

nature

LIFE SPEED

Cell mutation rates
scale with lifespan
in mammals



Commercial prospect
Can brain-computer
interfaces move from
clinic to market?

Keeping track
How organizations
can quantify their
biodiversity footprint

Entangled states
Quantum computing
using arrays of neutral
atoms

Vol. 594 No. 7398
www.nature.com

Nature.2022.04.23

[Sat, 23 Apr 2022]

- [This Week](#)
- [News in Focus](#)
- [Opinion](#)
- [Work](#)
- [Research](#)
- [Amendments & Corrections](#)

This Week

- **Le Pen election win would be disastrous for research, France and Europe** [19 April 2022]
Editorial • Marine Le Pen is promising to repeal unpopular changes to research institutions. But the wider impact of her presidency would be catastrophic.
- **To advance equality for women, use the evidence** [19 April 2022]
World View • These are three mistakes universities make when they attempt to improve gender equity.

- EDITORIAL
- 19 April 2022

Le Pen election win would be disastrous for research, France and Europe

Marine Le Pen is promising to repeal unpopular changes to research institutions. But the wider impact of her presidency would be catastrophic.



Emmanuel Macron, pictured in 2016 during a visit to Sciences Po in Paris.Credit: Alain Jocard/AFP/Getty

In less than a week, French President Emmanuel Macron will face far-right leader Marine Le Pen in the second round of a presidential election. Last time this happened, in 2017, Macron won comfortably, by a margin of 66% to Le Pen's 34%. But this time, opinion polls are predicting a much tighter race.

Macron has struggled to fulfil the promise of a new kind of presidency after his then newly formed party *La République En Marche!* (Republic Forward!) smashed France's two major right- and left-wing parties in the 2017 poll. Macron is now less popular with voters, probably including academics and students. And Le Pen, who leads *Rassemblement National* (National Rally) aims to persuade those who are still undecided that they should back her.

Science is not often a big factor in France's elections, and this one is no different. But Le Pen is appealing to scientists by pledging to repeal controversial reforms to research institutions enacted between 2007 and 2009 by centre-right president Nicolas Sarkozy — which Macron has continued. Both presidents sought to align France's universities, research and funding systems more closely with those of the United States and the United Kingdom by giving universities more autonomy; improving links between academics and businesses; and increasing financial support for research-intensive corporations.

Sarkozy changed the law so that funders and university administrations could have more independence in making decisions. His government also provided generous tax breaks to businesses that invest in research and development. These reforms, however, are not popular among many researchers, and Le Pen's pledge to repeal them might be attractive to academics who otherwise would not consider voting for her, researchers who study France's research system have told *Nature*.

Fears of precarity

France's research system is regarded by the state as a government department — academics are classed as civil servants, positions that receive

generous pensions and protection against employers setting too high a workload.

The reforms meant that universities would be able to amend researchers' contracts and many scientists did not welcome these changes, partly because of the risk that they could make employment more precarious, or change job terms and conditions. Thousands took to the streets to protest. Some lawmakers also questioned whether companies really need a government research subsidy — one of the highest in the European Union, worth some €5.7 billion (US\$6.2 billion) annually — to boost their own research and development.

Le Pen has said she will roll back these changes, and will use some of the money saved from the tax breaks to boost public spending on science. In a written interview this month with Patrick Lemaire, president of the Council of French Academic Learned Societies in Rennes, she declared: “The state will increase the public research budget, in particular by recovering part of the research tax credit.”

Although Le Pen's policy on the Sarkozy reforms might be welcomed by some researchers, National Rally's wider programme for government will be anything but. For one, the party's policy on restricting immigration is likely to hit collaborations with scientists in other countries. And minority communities would face severe discrimination under Le Pen. For example, she has said she wants to ban the wearing of headscarves in public by extending a law that prohibits them in schools.

Severe EU tensions

Furthermore, a Le Pen presidency would put France on a collision course with the EU. Her party is intending to violate European laws and regulations by restricting employment or state benefits for EU citizens from outside France; withholding payments into the EU budget; and ending free movement of people between France and its EU neighbours. Universities and research funders must also confront the possibility that a Le Pen government would seek to restrict academic freedom.

Unlike in 2017, Le Pen is not now advocating leaving the EU, but her policies would no doubt create severe tensions with the bloc, as is already happening between the EU and Hungary. Earlier this month, Hungary's far-right leader, Viktor Orbán, was re-elected. On the same day, his government announced a €500-million investment in university science parks at 20 universities. But two days later, the European Commission in Brussels began a process that could lead to Hungary losing potentially billions of euros in funding, under what's known as a conditionality mechanism.

This can be triggered if a member state, despite warnings, passes laws or enacts policies that are contrary to the rule of law. In September 2020, the European Commission confirmed in a 'rule of law report' that Orbán's government has interfered with the independence of Hungary's judiciary and with the functioning of its universities and media, since coming to power in 2010.

Researchers should consider that any short-term gains in terms of funding would be completely outweighed by the disaster of a Le Pen win. And those dissatisfied with both presidential candidates and considering not voting at all should realize that this, too, is likely to be of benefit to Le Pen. Everyone should look at Hungary for an EU case study of what happens when a far-right leader is elected.

Nature **604**, 401 (2022)

doi: <https://doi.org/10.1038/d41586-022-01043-0>

This article was downloaded by **calibre** from <https://www.nature.com/articles/d41586-022-01043-0>

- WORLD VIEW
- 19 April 2022

To advance equality for women, use the evidence



These are three mistakes universities make when they attempt to improve gender equity.

- [Michelle Ryan](#) ⁰

I'm an academic who studies women in leadership roles and also heads an institute devoted to this topic at the Australian National University in Canberra. So I spend a lot of time thinking about what it takes to achieve gender equity in workplaces. By now, the vast majority of universities, research institutions and funding bodies have some sort of initiative aimed at gender parity. In 2005, the Athena Swan accreditation programme to promote gender equity at universities was launched in the United Kingdom and later adopted in North America, Australia and elsewhere. In 2011, China's ministry of science and technology issued a policy to help advance

women in science and technology careers. In 2020, ambitious programmes began in the European Research Area and India.

Yet the most recent European Commission data show that women make up about half of doctoral graduates and only about one-quarter of senior academics and people in decision-making positions. In North America and Western Europe, only 33% of those employed in research and development are women; this drops to 24% in east Asia and the Pacific area, and to 18.5% in south and west Asia.

Why this inertia? In my view, progress could be faster if institutions that trumpet efforts to promote equity applied established research in their initiatives.

Here are three common missteps.

First, an overemphasis on quantity. Yes, metrics such as the proportion of female professors and grant winners are important. But simple tallies erase disparities in quality. Any tracking must capture the experiences and influence that awards and positions bestow. Do those given to women bring the same visibility, recognition and resources as those given to men? The proportion of women achieving authorships and professorships matters less if these are concentrated in sub-optimal, low-influence or temporary roles.

Consider the ‘glass-cliff’ phenomenon. Fifteen years ago, my colleagues and I reported that women are more likely than men to be given leadership roles that are risky, precarious and even doomed to failure ([M. K. Ryan and S. A. Haslam *Acad. Mgmt Rev.* 32, 549–572; 2007](#)). Archival, experimental and qualitative research provided evidence for this ‘poisoned chalice’, which holds from firms in the FTSE 100, the main UK share index, to chief librarians and heads of state. As institutions try to navigate a pandemic, as well as rising costs and increasing societal divisions, such considerations matter more than ever. The crucial question is, are women getting the same quality of promotions as men?

A second mistake is emphasizing training for individuals, instead of overhauling systems and cultures. Again and again, I see women offered extra coaching to encourage them to take career risks, overcome ‘impostor

'syndrome' and boost their skills in leadership and grant writing. But the evidence is clear: it is not women who need fixing, but entrenched systems of inequality.

Our research demonstrates that women's confidence and ambition are not inherently lower than men's, but are eroded by experiences in unequal workplace cultures — not having role models, and being treated differently from male counterparts. Similarly, women are not inherently risk-averse; they operate within systems that reward men for risk-taking, but punish women for the same behaviour.

Individually targeted interventions, at best, provide a short-term fix for a few already privileged women, and, at worst, reinforce the assumptions of success and leadership that underlie systemic gender inequality. Indeed, training programmes for women can have perverse effects by becoming yet another unrewarding demand on their time.

And the third mistake undermines all sorts of efforts: over-optimism. Yes, improvements are real and should be celebrated. Still, surveys of representation in boardrooms, films and various professions show that men and women consistently overestimate women's representation.

In [one of our studies](#), we found that veterinary surgeons who felt that sexism is no longer a problem in their profession were the most likely to pay a female member of staff less than a male member and to give her fewer career opportunities. In [another study](#), men who overestimated the proportion of women in the medical profession were the least likely to support gender-equality initiatives.

What's needed are interventions that genuinely address gender inequality in all its complexity and nuance.

There are good examples of concrete things that can be done: (1) systematic changes that improve the visibility and voice of women, such as prohibiting 'manels' (all-male panels), or requiring conference organizers to report proportions of women who are keynote speakers and panel members; (2) making senior leaders accountable for progress towards gender equality, as the Australian Champions for Change programme does, in which members

track factors such as pay, promotions and employment experiences; and (3) making research funding contingent on having a transparent and appropriately resourced gender-equality plan in place, as happens in the European Union's research and innovation strategy. Downgrading such requirements, as Britain announced in 2020 that it would do, exemplifies the sorts of backsliding on women's progress that is happening all too often now times are tough.

Good intentions are not enough to bring about change; nor are simple tallies, training programmes or unwarranted rosy views. Change requires sustained investment, appropriate incentives and evidence-backed interventions.

Nature **604**, 403 (2022)

doi: <https://doi.org/10.1038/d41586-022-01045-y>

This article was downloaded by **calibre** from <https://www.nature.com/articles/d41586-022-01045-y>.

| [Section menu](#) | [Main menu](#) |

News in Focus

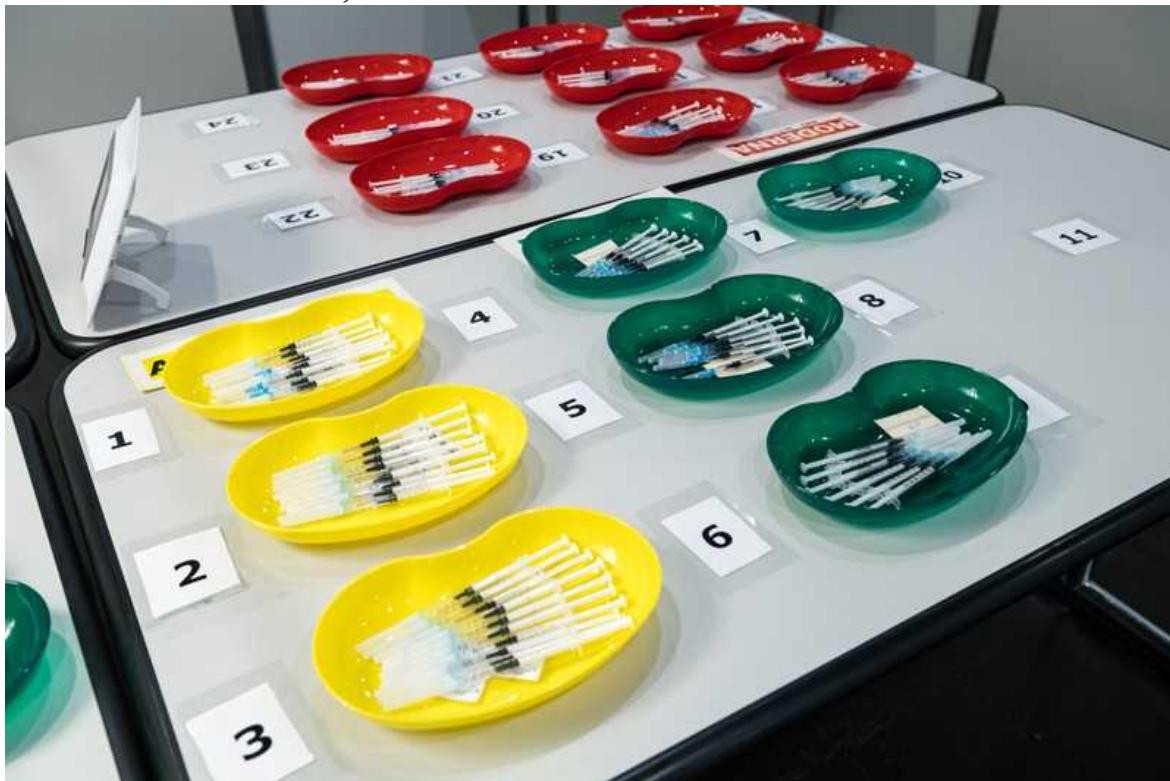
- **[COVID's diabetes risk, vaccine rankings and hidden industry ties](#)** [20 April 2022]
News Round-Up • The latest science news, in brief.
- **[How Shanghai's scientists are coping amid harsh COVID lockdown](#)** [15 April 2022]
News • From laps of the apartment to fears for students who desperately need data, four researchers speak about their lockdown experience.
- **[Jury finds University of Kansas chemical engineer guilty of hiding ties to China](#)** [08 April 2022]
News • After a two-week trial, Feng ‘Franklin’ Tao was convicted on four of eight charges — but a judge is reviewing the case.
- **[African clinical trial denied access to key COVID drug Paxlovid](#)** [04 April 2022]
News • Supply shortages and limits on research leave low- and middle-income countries struggling to access Pfizer’s COVID-19 antiviral.
- **[IPCC's starker message yet: extreme steps needed to avert climate disaster](#)** [05 April 2022]
News • Radical emissions cuts combined with some atmospheric carbon removal are the only hope to limit global warming to 1.5 °C, scientists warn.
- **[Ancient smells reveal secrets of Egyptian tomb](#)** [31 March 2022]
News • Jars contained fish, fruit and beeswax balm to sustain the tomb’s residents in the afterlife.
- **[China is hatching a plan to find Earth 2.0](#)** [12 April 2022]
News • A satellite will scour the Milky Way for exoplanets orbiting stars just like the Sun.
- **[The brain-reading devices helping paralysed people to move, talk and touch](#)** [20 April 2022]
News Feature • Implants are becoming more sophisticated — and are attracting commercial interest.

| [Next section](#) | [Main menu](#) | [Previous section](#) |

- NEWS ROUND-UP
- 20 April 2022

COVID's diabetes risk, vaccine rankings and hidden industry ties

The latest science news, in brief.



Pick one: a mass-vaccination centre in Leipzig, Germany, offers the COVID-19 vaccines made by Moderna (red dish), Pfizer–BioNTech (green) and AstraZeneca (yellow). Credit: Jens Schlueter/Getty

COVID vaccine comparison shows how jabs stack up

A [head-to-head comparison](#) shows that the COVID-19 vaccines made by Pfizer and Moderna outperform those from Johnson & Johnson (J&J) and Novavax. The data could aid vaccine design.

The study assessed the vaccines on 14 metrics, including the levels of immune cells and antibodies they produce.

The jabs fall into three classes. Those made by Moderna and by Pfizer–BioNTech are based on messenger RNA. J&J’s vaccine uses a virus to deliver SARS-CoV-2 genetic material into host cells. And Novavax’s jab contains pieces of a SARS-CoV-2 protein.

Antibody levels induced by two doses of the mRNA vaccines waned over six months, whereas antibody levels from J&J’s one-shot vaccine were stable or even increased. But antibody levels measured six months after a J&J vaccination were lower than those observed six months after an mRNA vaccination.

Novavax’s two-shot regimen induced antibody responses on a par with those to the mRNA vaccines, but levels of CD8⁺ T cells, which destroy infected cells, were low to undetectable.

The research was posted as a preprint on 21 March ([Z. Zhang *et al. Preprint at bioRxiv*](#)

Undisclosed payments rampant in drug-trial papers

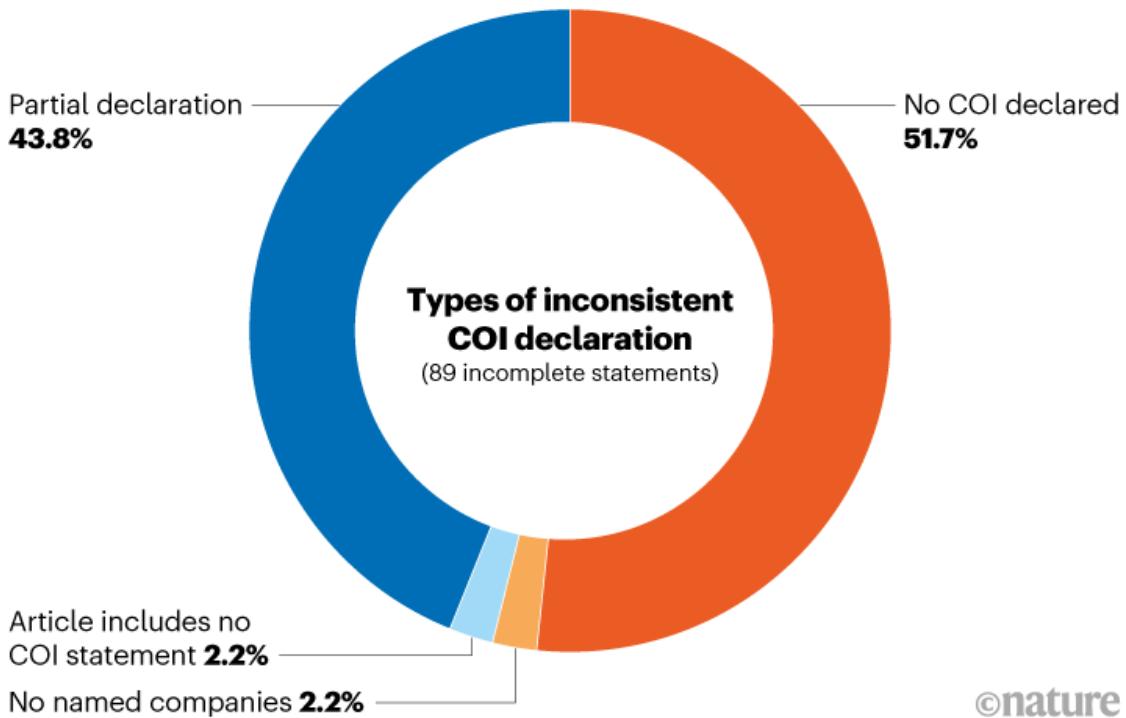
One in four Australian medical [researchers involved in drug trials failed to declare money](#) they had received from pharmaceutical companies when submitting journal manuscripts, a study reports.

The study’s authors cross-checked financial conflicts of interest listed by Australian authors of 120 drug trials published in the first 8 months of 2020 against a database of company-made payments reported to Medicines Australia, the country’s pharmaceutical-industry association. The research, published last month ([L. Flood *et al. J. Gen. Intern. Med.*](#)

<https://doi.org/hqbr; 2022>), found that missing or incomplete declarations were common.

MISSING INFORMATION

Most researchers with incomplete conflict of interest statements failed to declare any payments, and others included partial declarations.



©nature

Source: Flood, L. *et al. J. Gen. Intern. Med.* <https://doi.org/10.1007/s11606-022-07466-9> (2022).

Roughly half of the trial investigators with incomplete disclosures failed to declare any conflicts of interest when company reports showed they had been paid, and 43% of trials had partial declarations (see ‘Missing information’).

“We saw real inconsistencies where sometimes Australian researchers were not declaring funding from the same companies that their co-authors were declaring funding from,” says Barbara Mintzes, a pharmaceutical-policy researcher at the University of Sydney, Australia, who co-led the study. Undeclared financial ties are especially concerning in drug trials, in which transparency is paramount, she says.



People who were hospitalized with COVID-19 are at risk of a diabetes diagnosis. Credit: Bruna Prado/AP/Shutterstock

Diabetes risk rises after COVID

People who get COVID-19 have a [greater risk of developing diabetes](#) up to a year later, even after a mild infection, compared with those who never had the disease, a massive study shows.

The research, published this month, adds to evidence that COVID-19 can increase a person's risk of diabetes, months after infection ([Y. Xie and Z. Al-Aly *Lancet Diabetes Endocrinol.* \[Researcher Ziyad Al-Aly and epidemiologist Yan Xie at the Veterans Affairs \\(VA\\) St Louis Healthcare System in Missouri looked at the medical records of more than 180,000 people who had survived for longer than a month after catching COVID-19. They compared these with records from two groups, each of which comprised around four million people who had not had a SARS-CoV-2 infection. The analysis found that people who had had\]\(https://doi.org/hqb9; 2022\).</p></div><div data-bbox=\)](#)

COVID-19 were about 40% more likely to develop diabetes up to a year later than were those in the control groups. Almost all cases detected were of type 2 diabetes, and the chance of developing diabetes rose with increasing severity of COVID-19.

The pandemic will leave a legacy of chronic disease for which health-care systems are unprepared, says Al-Aly.

Nature **604**, 407 (2022)

doi: <https://doi.org/10.1038/d41586-022-01046-x>

This article was downloaded by **calibre** from <https://www.nature.com/articles/d41586-022-01046-x>

| [Section menu](#) | [Main menu](#) |

- NEWS
- 15 April 2022

How Shanghai's scientists are coping amid harsh COVID lockdown

From laps of the apartment to fears for students who desperately need data, four researchers speak about their lockdown experience.

- [Yvaine Ye](#)



Shanghai has been in lockdown since early April, and residents are barred from leaving their homes.Credit: Aly Song/Reuters/Alamy

As the highly infectious Omicron variant rages in parts of China, Shanghai, the country's financial hub and the home of many of its top research institutions, was plunged into a sweeping lockdown earlier this month. With restrictions seemingly set to continue, shortages of food and other supplies are making life tough for many residents — including many researchers and scientists.

China is one of the few remaining countries attempting to quash COVID-19 outbreaks. So when daily cases in Shanghai climbed steeply — more than 280,000 people in the city have tested positive since March — the authorities ordered its 25 million residents to stay at home from 1 April. The city also suspended all public transport, moved schools online and shut down businesses.

With cases continuing to rise, the lockdown has been extended, catching many residents unprepared. In some places, people are struggling to access health care. Public anger at the government's handling of the situation has also been rising.

Nature spoke with four researchers about what it's like living and working in the city.

Losing a year of data

As a conservation biologist at Fudan University, being stuck at home means Fang Wang can't conduct most of his research. Spring is the mating season for many animals that Wang studies. Before the lockdown, he had planned to visit the Qinling Mountains in northern China to study wild giant pandas (*Ailuropoda melanoleuca qinlingensis*), and the Liupan Mountains in northwestern China to observe endangered North China leopards (*Panthera pardus orientalis*). But he had to cancel the trips because of the Omicron outbreak. "For us, missing out on a season's data is like losing a year of data," he says. Wang hopes to do the field trips later this year.



Fang Wang on a research trip in Yunnan province in 2016.Credit: Zhiqiang Zhao

Without access to his university campus, Wang and his team can't analyse animal samples collected previously. Wang says the impact is manageable for him, but he worries about his students, whose dissertations rely on the data from such analyses.

"Over the past two years, our students have been under tremendous pressure and faced great uncertainty about their future. As a teacher, I worry such uncertainty will make my students prioritize stability over what they actually want to pursue when choosing a career," he says.

With Shanghai at a standstill, Wang also found himself overwhelmed by fresh responsibilities. Besides conducting research and mentoring his students, he needs to take care of his young child, who is also cooped up at home, and help with cooking. He must also look for ways to get groceries, which are difficult to find because residents are barred from leaving their

homes, so the city's entire population is relying on a limited number of delivery workers to bring necessities to their doors. "I feel really tired and stretched thin," he says.

3000 steps a day, at home

Before the lockdown, Jiahong Wen, a natural-disasters modeller at Shanghai Normal University, had planned to conduct field research in Linhai, a city nearby that was hit by a catastrophic typhoon in 2019. But Wen had to abandon his trip when the government announced that people in Shanghai could not leave home.

Wen can do most of his other work on his laptop at home. To keep fit, he's been exercising, including lifting weights. Sometimes he just walks around his apartment. "Every day I walk about 3,000 steps just by circling my home dozens of times. I also told my students to do so, to keep them healthy," he says.

In 2020, Wen worked on modelling the pandemic. Considering the size of Shanghai's current outbreak and the city's strict restrictions, he thinks COVID-19 cases will start to fall and that the city will return to normal as early as May. But he's concerned for his daughter, who is supposed to be sitting China's extremely competitive university entrance exam in June. "Omicron seems to be very stealthy and cases have been popping up unexpectedly. I don't want anything to happen to my daughter that would affect her exam."

Travel restrictions

"I consider myself extremely lucky, since our institute's administration and logistic staff on site have been trying their best to keep us safe and fed," says Xian Shi, an astronomer at the Shanghai Astronomical Observatory who also lives on her institute's campus in central Shanghai. The lockdown hasn't had a huge impact on her research — she studies small objects, such as asteroids, in the Solar System — because she's been working from home for the past two years of the pandemic. But Shi had to cancel a long-planned trip to give a talk and meet her collaborators in a nearby city.

“I expected that the lockdown probably would not end as planned, given the highly infectious variant and stringent criteria for opening up,” she says. As a result, she brought her laptops and hard drives, as well as a lot of coffee, back to her home ahead of time.

Shi worries that China’s strict travel restrictions will make it more difficult to attend international conferences and meet colleagues in other parts of the world.

A huge social experiment

To Jia Miao, a sociologist at New York University Shanghai, the current lockdown is like witnessing the focus of her past research unfolding in front of her. She studied community resilience during the 2020 citywide lockdown of Wuhan, where SARS-CoV-2 was first detected. Now, she’s experiencing a lockdown first-hand, which has given her ideas for future research.

“Although the lockdown is a misfortune for the city, for us sociologists, it’s a huge social experiment. It will give us lots of opportunities to study a variety of issues that arise.”

Her university gave staff and students a day’s notice before shutting down, so Miao had time to gather her laptop and assemble the data she needed to work from home. “Personally, I’m doing okay.” But the lockdown has forced her to postpone field trips, both for her own research and for the undergraduate classes she teaches. “I do hope there could be one last chance before the end of the term to take my students on a field trip.”

Miao teaches an urban sociology class which discusses topics such as the role of community in a pandemic. “After this outbreak, my students will have a fresh perspective when looking at communities, and they can bring their personal experience and insights to the class,” she says.

Nature **604**, 409-410 (2022)

doi: <https://doi.org/10.1038/d41586-022-01052-z>

| [Section menu](#) | [Main menu](#) |

- NEWS
- 08 April 2022
- Correction [11 April 2022](#)

Jury finds University of Kansas chemical engineer guilty of hiding ties to China

After a two-week trial, Feng ‘Franklin’ Tao was convicted on four of eight charges — but a judge is reviewing the case.

- [Sara Reardon](#)



Franklin Tao's trial was held at the Robert J. Dole Federal Courthouse in Kansas City. Credit: Michael Rosebrock/Alamy

University of Kansas (KU) chemical engineer Feng 'Franklin' Tao, [who was accused of hiding ties](#) to a Chinese university, has been found guilty of wire fraud and making false statements to the US government. On 7 April, a jury for the US district court of Kansas found that Tao, currently on unpaid leave from KU in Lawrence, had committed research-grant fraud by failing to tell his employer and federal funding agencies about an alleged faculty appointment in China.

The closely watched case is the latest prosecution of a scientist arrested under the China Initiative — a controversial US programme launched in 2018 by former president Donald Trump's administration to protect US institutions from economic espionage. On the basis of an internal investigation, the US Department of Justice (DoJ) [announced in February](#) that it was discontinuing the programme after accusations that the initiative was racially biased. The DoJ said it didn't find evidence of radical prejudice, but acknowledged that the initiative could have been perceived as fuelling a narrative of intolerance.

In a statement sent to *Nature*, Tao's attorney, Peter Zeidenberg, says that he hopes the verdict will be overturned. He notes that the judge has ordered a briefing on the government's evidence and whether anyone was actually defrauded, and that she did not set a sentencing date. "While we are deeply disappointed with the jury's verdict, we believe it was so clearly against the weight of the evidence we are convinced that it will not stand," Zeidenberg writes.

A [press release from the DoJ](#) says that Tao could face decades in federal prison and a fine of up to US\$250,000 for each offence.

A federal case

Tao, who was first arrested in 2019, is thought to be the first scientist charged under the China Initiative. DoJ prosecutors say he planned to defraud KU and US federal agencies by accepting a position at Fuzhou

University in China. They say Tao failed to report the alleged employment to the US Department of Energy and the US National Science Foundation, which funded his research at KU.

The prosecution's primary evidence included unsigned contracts from Fuzhou University that the FBI had found in Tao's e-mail account, along with e-mails suggesting that he had attempted to recruit student employees at the Chinese university, and requests for Fuzhou to buy equipment for his laboratory there. [Tao has told *The New Yorker* magazine](#) that he had considered moving to Fuzhou University, but decided against it.

Tao's defence team argued that the evidence came from an unreliable source: one of Tao's former research associates, Humin Liu. E-mails [seemed to show](#) Liu trying to extort Tao for \$300,000 and accusing him of having harmed her career prospects, before reporting him to KU and the FBI for espionage. Liu submitted the accusations under false names and apparently hacked into Tao's e-mail account to acquire the contracts.

Zeidenberg challenged the evidence suggesting that Tao was ever employed at Fuzhou University. The charges, he said, were minor administrative errors that had been blown out of proportion by federal investigators. He said there was no evidence that KU or anyone else was harmed. "Because it involved China, the government has turned it into a federal case," he told the court, [according to AP News](#).

An unclear future

The defence team called only two witnesses: Tao's wife and a pastor. Tao himself never took the stand. The trial lasted for two weeks, and the jury deliberated for more than a day before returning a verdict. It found Tao guilty of three out of six counts of wire fraud and one out of two counts of making false statements.

"When I heard the news, I was surprised, I was disappointed and I was confused about what this means for the future," says Jenny Lee, a social scientist at the University of Arizona in Tucson [who published a 2021 survey](#) showing that US scientists of Chinese descent feared being under

surveillance after the China Initiative launched. “I really anticipated the case would have been dropped a while ago.”

Tao is one of dozens of scientists of Chinese heritage to have been indicted for alleged ties to China. A [December 2021 analysis](#) by *MIT Technology Review* found that of the 77 known China Initiative cases, only 19 involved economic espionage; many of the rest involved corruption or research-integrity issues, such as the proper disclosure of foreign funding on grant applications. Most of the scientists were acquitted or had the charges against them dropped.

Among them is Gang Chen, a mechanical engineer at the Massachusetts Institute of Technology (MIT) in Cambridge who was arrested in 2021 on similar charges of hiding ties to China. The DoJ dropped Chen’s charges in January of this year. To Chen, Tao’s case suggests that the US government holds scientists of Chinese heritage to a different standard of research conduct from other researchers. “Most people would not disclose to their employer that they’re looking for a new job — that’s not a crime,” he says.

Chen adds that the government’s actions are creating a “chilling effect” on international scientists’ and students’ willingness to come to the United States. “The US is built on talent. If talent is leaving and not coming, how does that help with national security?” he says. And although MIT welcomed him back after his charges were dropped, Chen says he and his family still live in fear. “You can’t get rid of this fear, you fear you’re watched every day.”

[In a 23 February speech](#), US assistant attorney-general for national security Matthew Olsen said that although the China Initiative would be ending, the US government would continue to investigate espionage and fraud cases under a new plan called the Strategy for Countering Nation-State Threats. This initiative covers China, as well as several other “hostile” nations, including Iran and Russia. Olsen added that the DoJ will continue to stand behind the cases that it is currently prosecuting.

Pending cases

Critics of the China Initiative say that the decision by President Joe Biden's administration to end the policy should include a reexamination of pending cases, rather than continued prosecution. The decision to pursue Tao's case — and his conviction — is discouraging, Lee says. "It certainly does not reassure the Chinese American science community that the China Initiative in its current form will end, even though they're dropping the label."

Lee points out that US District Judge Julie Robinson, who is adjudicating Tao's case, took steps to distance the case from the China Initiative, banning testimony from either party that would have suggested that Tao was involved with espionage. However, Lee says, by focusing only on Tao's alleged crimes, the DoJ also distanced itself from the accusations of racial bias linked to the initiative, and the actual reasons that Tao was investigated in the first place. A spokesperson for the DoJ declined to comment to *Nature*.

Several other cases are set to go to trial later this year, and the Chinese American academic community will be watching them closely, says Haipei Shue, president of the non-profit organization United Chinese Americans in Washington DC. "We feel very strongly this is country-of-origin profiling," he says.

Shue also worries that the government's goal in treating cases such as Tao's as criminal rather than administrative is to dissuade Chinese espionage rather than punish a crime. He cites a Chinese proverb: "Kill a chicken to scare the monkey. If this is the case, I don't have the words to describe that. It sounds to me very medieval."

Nature **604**, 410-411 (2022)

doi: <https://doi.org/10.1038/d41586-022-01022-5>

Updates & Corrections

- **Correction 11 April 2022:** An earlier version of this story labelled the opening image as being of the Robert J. Cole Federal Courthouse. It is the Robert J. Dole Federal Courthouse. An earlier version of this story also said that of the 77 known China Initiative cases reported by *MIT*

Technology Review in December 2021, 19 involved economic espionage and the rest involved research-integrity issues. The story has been updated to clarify that the remainder also involved corruption and other issues.

This article was downloaded by **calibre** from <https://www.nature.com/articles/d41586-022-01022-5>

| [Section menu](#) | [Main menu](#) |

- NEWS
- 04 April 2022

African clinical trial denied access to key COVID drug Paxlovid

Supply shortages and limits on research leave low- and middle-income countries struggling to access Pfizer's COVID-19 antiviral.

- [Heidi Ledford](#) &
- [Amy Maxmen](#)



The ANTICOV trial aims to find COVID-19 treatments that are effective in resource-poor settings. Credit: Kenny Mbala/DNDi

Researchers in a pivotal African clinical trial have been denied access to a key COVID-19 antiviral drug, frustrating efforts to test the treatment in African populations and in combination with therapies that could potentially expand its utility on the continent.

The Drugs for Neglected Diseases Initiative (DNDi) has asked Pfizer for supplies of its antiviral drug Paxlovid to use in the ANTICOV clinical trial, a large study in ten African countries that aims to find treatments for mild to moderate COVID-19 — particularly those that will work in resource-poor settings.

Earlier this year, DNDi requested enough Paxlovid to treat 1,000–2,000 ANTICOV participants, but says that Pfizer denied the request and stated that it already had plans to conduct similar trials. The details of those trials have not yet been made public, and Pfizer did not elaborate on such plans or the request from DNDi when asked about them by *Nature*. “We are committed to well-controlled, hypothesis-driven clinical studies that can provide data that will be accepted by global regulatory agencies,” the company said in a statement. “We’re focusing our efforts and resources in a way that maximizes availability of our overall supply.”

But Pfizer’s decision has dismayed public-health advocates. DNDi’s inability to obtain medicine for clinical trials compounds ongoing inequities in access to vaccines and treatments, including Paxlovid, in low- and middle-income countries, says Brook Baker, a law professor at Northeastern University in Boston, Massachusetts, who specializes in global access to medicines and vaccines. “This is a trial in a population that needs to be studied,” he says. “Why don’t we get on with it?”

Two drugs in one

Paxlovid is a combination of two drugs: nirmatrelvir and ritonavir. Nirmatrelvir blocks a SARS-CoV-2 protein that is important for viral replication; ritonavir acts in the liver to prevent the rapid breakdown of nirmatrelvir, and is a component of some therapies for treating HIV and hepatitis C infections.

In December, Pfizer, which is headquartered in New York City, announced that Paxlovid reduced the risk of hospitalization or death from COVID-19 by 88% when adults at high risk of developing severe disease were given the drug within five days of symptom onset¹. The US Food and Drug Administration (FDA) followed soon after with an emergency-use authorization for the drug, and researchers hailed the development of an oral antiviral that didn't have to be administered in hospital.

But nirmatrelvir is in very short supply, and that put low- and middle-income countries at a disadvantage in the ensuing bidding war for Paxlovid. Pfizer has said it can make up to 120 million courses of the treatment by the end of the year. Of those, the company has promised up to four million courses to the United Nations children's charity UNICEF for distribution in low- and middle-income countries. "We are in continued conversations with a number of private partners and international organizations to provide Paxlovid to lower income countries," Pfizer said in a statement.

On 17 March, the Medicines Patent Pool, a non-profit organization with headquarters in Geneva, Switzerland, that has licensed Pfizer's Paxlovid patents, announced that 35 companies had signed agreements to make generic Paxlovid, which could then be eligible for sale in 95 low- and middle-income countries. But they are not expected to market significant supplies of the drug until 2023, and the list does not include many middle-income countries including Thailand, Malaysia and some in Latin America.

While the world waits for Paxlovid, some countries are leaning on another antiviral called molnupiravir (Lagevrio), authorized by the FDA at the end of last year. But the ANTICOV trial has decided against including molnupiravir for now, says Nathalie Strub-Wourgaft, director of the COVID-19 response at DNDi in Geneva. The drug, supplies of which are also limited, reduced hospitalizations and deaths from COVID-19 by only about 30% when given to adults at high risk for severe disease. And because of potential side effects, the FDA recommends that pregnant people not take the drug, and that people use contraception during treatment and for at least three months after. "Contraception is not so easy to control for men and women," says Strub-Wourgaft. "And our criteria are really to include the largest population that is possible — and that includes pregnant patients."

Combination treatments

There is a pressing need to test Paxlovid in a range of populations, says Melissa Barber, who studies health economics at the Harvard T.H. Chan School of Public Health in Boston, Massachusetts. “Clinically, we might expect population with different comorbidities — for example HIV or diabetes — might have a range of side effects or possibly effectiveness,” she says.

One potential downside of Paxlovid is that the drug is meant to be taken within five days of when COVID-19 symptoms first appear. “Getting a diagnosis that quickly anywhere is hard,” says Rachel Cohen, the regional executive director at DNDi in New York City, “but it’s pretty much impossible in low- and middle-income countries, where there are few tests and almost no rapid tests.”

An analysis of ANTOCOV participants showed that only about half of them were diagnosed within five days of the onset of symptoms. In hopes of prolonging the window during which Paxlovid is useful, DNDi wants to test the drug in combination with an inhalable steroid called budesonide, to see whether this can delay the onset of harmful immune responses that often drive later stages of COVID-19.

The initiative also had ambitions to test Paxlovid in combination with the antidepressant fluoxetine (Prozac), if results from its other fluoxetine trials show that the drug has antiviral properties, says Strub-Wourgaft.

Cohen is concerned about a clause in Pfizer’s agreement with the Medicines Patent Pool that requires the companies to obtain permission from Pfizer before combining their generic Paxlovid with other products, or from giving their drugs to researchers who intend to study combinations.

Pfizer did not respond to queries from *Nature* about why it would not allow companies to use the drug without its permission for combination studies.

It is not unusual for drug companies to be hesitant to allow external researchers to conduct studies with medicines such as Paxlovid that are not yet fully approved by regulators, says Cohen. But she adds that delays and a

lack of transparency on what the company is studying are surprising, given the emergency of the pandemic. “We have always said that we are happy to collaborate with them,” she says.

John Amuasi, a principal investigator on the ANTICOV study at the Kumasi Center for Collaborative Research in Tropical Medicine in Ghana, agrees. With just 15% of the country fully vaccinated, he’s anxious to obtain the drug for both treatment and research purposes. “We think this could be a game changer in an emergency situation.”

Nature **604**, 412-413 (2022)

doi: <https://doi.org/10.1038/d41586-022-00919-5>

References

1. Hammon, J. *et al.* *N. Engl. J. Med.* <https://doi.org/10.1056/NEJMoa2118542> (2022).

This article was downloaded by **calibre** from <https://www.nature.com/articles/d41586-022-00919-5>

| [Section menu](#) | [Main menu](#) |

- NEWS
- 05 April 2022
- Correction [06 April 2022](#)

IPCC's starker message yet: extreme steps needed to avert climate disaster

Radical emissions cuts combined with some atmospheric carbon removal are the only hope to limit global warming to 1.5 °C, scientists warn.

- [Jeff Tollefson](#)



One success story in the battle against climate change is that renewable energy sources, such as wind turbines, have dropped significantly in cost over the past decade. Credit: Vincenzo Izzo/LightRocket via Getty

Humanity probably isn't going to prevent Earth from at least temporarily warming 1.5 °C above pre-industrial levels — but aggressive action to curb greenhouse-gas emissions and extract carbon from the atmosphere could limit the increase and bring temperatures back down, [according to the latest report](#) from the United Nations Intergovernmental Panel on Climate Change (IPCC). The report makes it clear, however, that the window is rapidly closing, and with it the opportunity to prevent the worst impacts of global warming. Above the 1.5 °C limit — set by the Paris climate agreement in 2015 — the chances of extreme weather and collapsing ecosystems grow.

"The IPCC tells us that we have the knowledge and technology to get this done," Inger Andersen, executive director of the UN Environment Programme, said at the press conference for the report's release. "But increased action must begin this year, not next year; this month, not next month; and indeed today, not tomorrow."

Approved by 195 governments after a marathon negotiating session that ran over schedule by two days, [the roughly 2,900-page report](#) focuses on options for curbing emissions and mitigating the impacts of global warming. The document, compiled by hundreds of scientists across 65 countries, is the last of a trilogy comprising the IPCC's sixth climate assessment. The first two reports cover [the underlying science](#) and [impacts of climate on humans and ecosystems](#).

Multiple sources involved in the virtual session told *Nature* that the negotiations to finalize the report were bogged down as government delegates hashed out perennial arguments over climate mitigation. In particular, negotiators for India raised questions about emissions scenarios in the report, arguing that they assume too much action on the part of developing countries and do not adequately reflect questions of equity and responsibility. Negotiators for Saudi Arabia scrutinized language related to carbon-capture technologies and the future of fossil fuels. Although these debates pushed the negotiations into overtime, sources say they did not impact the findings or distort the underlying science in the report.

Coming more than three decades after the panel's first climate assessment, the sixth instalment delivers the most forceful warning yet about the consequences of inaction (see 'Key points from the report'). The question now, scientists say, is whether governments will at last step up to the challenge with actions rather than unfulfilled pledges.

"Despite more mitigation efforts by more governments at all scales, emissions continue to increase," says Karen Seto, a geographer at Yale University in New Haven, Connecticut, and a coordinating lead author on the report. "We need to do a lot more, and we need to do it quickly."

Key points from the report

- This report is one of the most stringent warnings yet from the IPCC. The message? Time has almost run out. Models suggest that global emissions need to peak, at the latest, by 2025 and then decline rapidly for the world to have a 50% chance of limiting warming to 1.5 °C. Carbon emissions would need to nearly halve by 2030 and hit 'net zero' in the early 2050s to meet the goal. Given current policies, some scientists estimate that the world is on track for a rise of nearly 3 °C rise above pre-industrial levels.
- But the report is not entirely doom and gloom. Although emissions continue to rise, there are signs that some mitigation efforts have had an impact. The price of renewable-energy technologies, such as wind turbines, solar panels and batteries, is plummeting, and the global economy is getting cleaner. Global energy intensity — a measure of the amount of energy required to drive the economy — decreased by 2% annually between 2010 and 2019, reversing the trend from the prior decade.
- To prevent temperatures from significantly overshooting the 1.5 °C threshold, some fossil fuels will need to remain in the ground. According to models that hold global warming to only slightly above this limit, emissions from existing and planned fossil-fuel projects already exceed the allowable carbon budget.
- For countries to achieve the net-zero emissions goals that they have set, dialling back emissions won't be enough — they will also need to extract carbon dioxide from the atmosphere. This will offset residual greenhouse-

gas emissions from sectors that are harder to clean up, such as industry and aviation. Nations could bolster carbon uptake by expanding forests and improving agricultural practices, or through a variety of nascent technologies that can capture carbon emissions either from industrial sources or directly from the atmosphere.

- Despite concerns about the costs of mitigation, meeting climate goals won't break the global bank: models suggest that global economic growth will continue over the coming decades, even with aggressive action to curb emissions. Although the global gross domestic product at mid-century is projected to dip slightly in scenarios in which climate policies have been enacted, most research suggests that the economic benefits of limiting warming — including improved health and reduced climate damages — exceed the cost of mitigation.
- Still, wealthy nations will need to contribute financial aid to low-income countries, to address inequities in vulnerability to climate change and to accelerate the clean-energy transition in a way that benefits all. The nations that have emitted the lowest amounts of greenhouse gases are often the ones most affected by climate change: the 88 countries that comprise the Least Developed Countries and Small Island Developing States groups within the UN climate framework are collectively responsible for less than 1% of historical carbon emissions.

The good news and the bad news

The report makes clear that current energy, economic and political trends put the world on course to shoot well past 1.5 °C of warming. Scientists have long been warning of this, but some say it's time to start thinking about what that means in terms of climate strategy.

"I think we are getting closer politically to a situation where we seriously have to ask how we are going to deal with that overshoot," says Oliver Geden, a social scientist with the German Institute for International and Security Affairs in Berlin and a lead author on the report. Although it still might be technically possible to limit warming to 1.5 °C, he says, the actions required would be unprecedented.

But the report provides reason for optimism by highlighting climate technologies and policies that are already driving emissions down in many countries. The immediate goal is to accelerate those efforts and ramp up climate finance to ensure that it's a truly global effort, says Nathaniel Keohane, president of the Center For Climate and Energy Solutions, an environmental think tank in Arlington, Virginia, and a White House adviser under former US president Barack Obama. Longer term, governments need to invest in research and development activities to explore the feasibility of carbon-removal technologies that could help to bend the curve in decades to come.

"It's a Herculean effort, and so we better get started," Keohane says.

Nature **604**, 413-414 (2022)

doi: <https://doi.org/10.1038/d41586-022-00951-5>

Updates & Corrections

- **Correction 06 April 2022:** An earlier version of this story said that social scientist Oliver Geden is a coordinating lead author on the IPCC report. He is a lead author.

This article was downloaded by **calibre** from <https://www.nature.com/articles/d41586-022-00951-5>

| [Section menu](#) | [Main menu](#) |

- NEWS
- 31 March 2022

Ancient smells reveal secrets of Egyptian tomb

Jars contained fish, fruit and beeswax balm to sustain the tomb's residents in the afterlife.

- [Colin Barras](#)



This papyrus from the tomb shows Kha and his wife Merit worshipping the lord of the afterlife, Osiris. Credit: Leemage/Corbis/Getty

More than 3,400 years after two ancient Egyptians were laid to rest, the jars of food left to nourish their eternal souls still smell sweet. A team of analytical chemists and archaeologists has analysed these scents to help identify the jars' contents¹. The study shows how the archaeology of smell can enrich our understanding of the past — and perhaps make museum visits more immersive.

The 1906 discovery of the intact tomb of Kha and Merit in the Deir el-Medina necropolis near Luxor was a landmark moment in Egyptology. The tomb of Kha — a ‘chief of works’, or an architect — and Merit, his wife, remains the most complete non-royal ancient burial ever found in Egypt, revealing important information about how high-ranking individuals were treated after death.

“It’s an amazing collection,” says Ilaria Degano, an analytical chemist at the University of Pisa, Italy. “Among the objects, there are even examples of Kha’s ancient Egyptian linen underwear, embroidered with his name.”

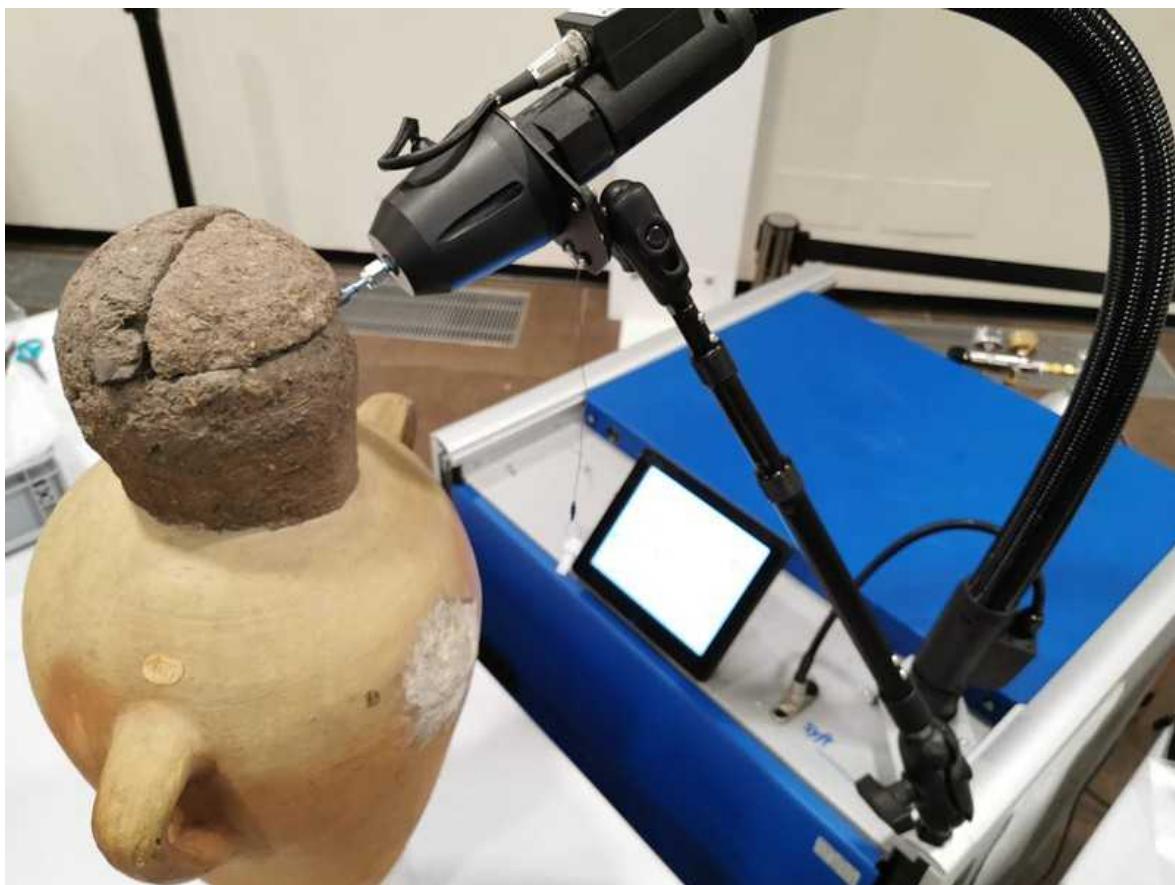
Unusually for the time, the archaeologist who discovered the tomb resisted the temptation to unwrap the mummies or peer inside the sealed amphorae, jars and jugs there, even after they were transferred to the Egyptian Museum in Turin, Italy. The contents of many of these vessels are still a mystery, although there are some clues, says Degano. “From talking with the curators, we knew there were some fruity aromas in the display cases,” she says.

Odour analysis

Degano and her colleagues placed various artefacts — including sealed jars and open cups laden with the rotten remains of ancient food — inside plastic bags for several days to collect some of the volatile molecules they still release. Then the team used a mass spectrometer to identify the components of the aromas from each sample. They found aldehydes and long-chain hydrocarbons, indicative of beeswax; trimethylamine, associated with dried fish; and other aldehydes common in fruits. “Two-thirds of the objects gave some results,” Degano says. “It was a very nice surprise.”

The findings will feed into a larger project to re-analyse the tomb's contents and produce a more comprehensive picture of burial customs for non-royals that existed when Kha and Merit died, about 70 years before Tutankhamun came to the throne.

This isn't the first time that scent compounds have revealed important information about ancient Egypt. In 2014, researchers extracted volatile molecules from linen bandages that are between 6,300 and 5,000 years old that were used to wrap bodies in some of the earliest known Egyptian cemeteries². The molecules confirmed the presence of embalming agents with antibacterial properties, showing that Egyptians were experimenting with mummification some 1,500 years earlier than had been thought.



One of the jars whose contents were analyzed using a mass spectrometer. Credit: J. La Nasa *et al.*/J. Archaeol. Sci

Odour analysis is still an underexplored area of archaeology, says Stephen Buckley, an archaeologist and analytical chemist at the University of York,

UK, who was involved in the 2014 study. “Volatiles have been ignored by archaeologists because of an assumption they would have disappeared from artefacts,” he says. But “if you want to understand the ancient Egyptians, you really want to go into that world of smell”.

For example, sweet-smelling incense derived from aromatic resins was essential for the ancient Egyptians. “Incense was necessary for temple ceremonies and for some mortuary rituals,” says Kathryn Bard, an archaeologist at Boston University in Massachusetts. Because resin-producing trees didn’t grow in Egypt, this necessitated [ambitious long-distance expeditions to obtain supplies](#).

Enriched exhibits

Aside from revealing more about past civilizations, ancient smells could add a dimension to the visitor experience at museums. “Smell is a relatively unexplored gateway to the collective past,” says Cecilia Bembibre at University College London. “It has the potential [to allow] us to experience the past in a more emotional, personal way.”

But [reconstructing ancient smells is not easy](#), says Bembibre. Degradation and decomposition can be a smelly business, so the scents from an artefact today do not necessarily match what Bembibre calls the original “smellscape” of a tomb.

With the right knowledge and understanding, it should be possible to pull the original and the decomposition scents apart, says Buckley. Whether visitors would actually want to experience the full and potentially unpleasant smellscape of an ancient tomb is still up for debate. “Curators might want to give people a choice over how far they want to push the smell experience,” says Buckley.

Nature **604**, 414 (2022)

doi: <https://doi.org/10.1038/d41586-022-00903-z>

References

1. La Nasa, J. *et al.* *J. Archaeol. Sci.* **141**, 105577 (2022).
 2. Jones, J. *et al.* *PLoS ONE* **9**, e103608 (2014).
-

This article was downloaded by **calibre** from <https://www.nature.com/articles/d41586-022-00903-z>

| [Section menu](#) | [Main menu](#) |

- NEWS
- 12 April 2022

China is hatching a plan to find Earth 2.0

A satellite will scour the Milky Way for exoplanets orbiting stars just like the Sun.

- [Yvaine Ye](#)



China is planning its first space mission to survey the skies for exoplanets similar to Kepler-186f, an Earth-size planet orbiting a distant star (artist's impression). Credit: NASA Ames/JPL-Caltech/T. Pyle

After sending robots to the Moon, landing them on Mars and building its own space station, China is now eyeing distant solar systems. This month,

scientists will release detailed plans for the country's first mission to discover exoplanets.

The mission will aim to survey planets outside the Solar System in other parts of the Milky Way, with the goal of finding the first Earth-like planet orbiting in the habitable zone of a star just like the Sun. Astronomers think such a planet, called an Earth 2.0, would have the right conditions for liquid water — and possibly life — to exist.

More than 5,000 exoplanets have already been discovered in the Milky Way, mostly with NASA's Kepler telescope, which was in use for 9 years before it ran out of fuel in 2018. Some of the planets were rocky Earth-like bodies orbiting small red-dwarf stars, but none fit the definition of an Earth 2.0.

With current technology and telescopes, it is extremely hard to find the signal of small, Earth-like planets when their host stars are one million times heavier and one billion times brighter, says Jessie Christiansen, an astrophysicist at the NASA Exoplanet Science Institute at California Institute of Technology in Pasadena.

The Chinese mission, called Earth 2.0, hopes to change that. It will be funded by the Chinese Academy of Sciences and is wrapping up its early design phase. If the designs pass a review by a panel of experts in June, the mission team will receive funding to start building the satellite. The team plans to launch the spacecraft on a Long March rocket before the end of 2026.

Seven eyes

The Earth 2.0 satellite is designed to carry seven telescopes that will observe the sky for four years. Six of the telescopes will work together to survey the Cygnus–Lyra constellations, the same patch of sky that the Kepler telescope scoured. “The Kepler field is a low-hanging fruit, because we have very good data from there,” says Jian Ge, the astronomer leading the Earth 2.0 mission at the Shanghai Astronomical Observatory of the Chinese Academy of Sciences.

The telescopes will look for exoplanets by detecting small changes in a star's brightness that indicate that a planet has passed in front of it. Using multiple small telescopes together gives scientists a wider field of view than a single, large telescope such as Kepler. Earth 2.0's 6 telescopes will together stare at about 1.2 million stars across a 500-square-degree patch of sky, which is about 5 times wider than Kepler's view was. At the same time, Earth 2.0 will be able to observe dimmer and more distant stars than does NASA's Transiting Exoplanet Survey Satellite (TESS), which surveys bright stars near Earth.

"Our satellite can be 10–15 times more powerful than NASA's Kepler telescope in its sky-surveying capacity," says Ge.

The satellite's seventh instrument will be a gravitational microlensing telescope for surveying rogue planets — free-roaming celestial objects that don't orbit any star — and exoplanets that are far from their star similar to Neptune. It will detect changes in starlight when the gravity of a planet or star warps the light of a background star that it is passing in front of. The telescope will target the centre of the Milky Way where massive numbers of stars are located. If successfully launched, this would be the first gravitational microlensing telescope that operates from space, says Ge.

"Our satellite can essentially conduct a census that identifies exoplanets of different sizes, masses and ages. The mission will provide a good collection of exoplanet samples for future research," he says.

Doubling the data

NASA launched Kepler in 2009, aiming to find out how common Earth-like planets are in the Galaxy. To confirm that an exoplanet is Earth-like, astronomers need to measure the time it takes to orbit its sun. Such planets should have an orbital period similar to Earth's and transit their suns about once a year. Chelsea Huang, an astrophysicist at the University of Southern Queensland in Toowoomba, says that scientists need at least three transits to work out a precise orbital period, which takes about three years of data, and sometimes more, if there are data gaps.

But four years into the Kepler mission, parts of the instrument failed, rendering the telescope unable to stare at one patch of the sky over an extended period of time. Kepler was on the cusp of finding some truly Earth-like planets, says Huang, who has worked with the Earth 2.0 team as a data-simulation consultant.

With Earth 2.0, astronomers could have another four years of data that, when combined with Kepler's observations, could help to confirm which exoplanets are truly Earth-like. "I am very excited about the prospect of returning to the Kepler field," says Christiansen, who hopes to study Earth 2.0's data if they are made available.

Ge hopes to find a dozen Earth 2.0 planets. He says he plans to publish the data within one or two years of their collection. "There will be a lot of data, so we need all the hands we can get," he says. The team already has about 300 scientists and engineers, mostly from China, but Ge hopes more astronomers worldwide will join. "Earth 2.0 is an opportunity for better international collaboration."

The European Space Agency is also planning an exoplanet mission — called Planetary Transits and Oscillations of Stars (PLATO) — that is scheduled to launch in 2026. PLATO's design has 26 telescopes, meaning that it will have a much larger field of view than Earth 2.0. But the satellite will shift its gaze every two years to observe different regions of the sky.

Nature **604**, 415 (2022)

doi: <https://doi.org/10.1038/d41586-022-01025-2>

This article was downloaded by **calibre** from <https://www.nature.com/articles/d41586-022-01025-2>

- NEWS FEATURE
- 20 April 2022

The brain-reading devices helping paralysed people to move, talk and touch

Implants are becoming more sophisticated — and are attracting commercial interest.

- [Liam Drew](#) ⁰



A person with paralysis controls a prosthetic arm using their brain activity.
Credit: Pitt/UPMC

James Johnson hopes to drive a car again one day. If he does, he will do it using only his thoughts.

In March 2017, Johnson broke his neck in a go-karting accident, leaving him almost completely paralysed below the shoulders. He understood his new reality better than most. For decades, he had been a carer for people with paralysis. “There was a deep depression,” he says. “I thought that when this happened to me there was nothing — nothing that I could do or give.”

But then Johnson’s rehabilitation team introduced him to researchers from the nearby California Institute of Technology (Caltech) in Pasadena, who invited him to join a clinical trial of a brain–computer interface (BCI). This would first entail neurosurgery to implant two grids of electrodes into his cortex. These electrodes would record neurons in his brain as they fire, and the researchers would use algorithms to decode his thoughts and intentions. The system would then use Johnson’s brain activity to operate computer applications or to move a prosthetic device. All told, it would take years and require hundreds of intensive training sessions. “I really didn’t hesitate,” says Johnson.

The first time he used his BCI, implanted in November 2018, Johnson moved a cursor around a computer screen. “It felt like *The Matrix*,” he says. “We hooked up to the computer, and lo and behold I was able to move the cursor just by thinking.”

Johnson has since used the BCI to control a robotic arm, use Photoshop software, play ‘shoot-’em-up’ video games, and now to drive a simulated car through a virtual environment, changing speed, steering and reacting to hazards. “I am always stunned at what we are able to do,” he says, “and it’s frigging awesome.”

Johnson is one of an estimated 35 people who have had a BCI implanted long-term in their brain. Only around a dozen laboratories conduct such research, but that number is growing. And in the past five years, the range of skills these devices can restore has expanded enormously. Last year alone, scientists described a study participant using a robotic arm that could send sensory feedback directly to his brain¹; a prosthetic speech device for

someone left unable to speak by a stroke²; and a person able to communicate at record speeds by imagining himself handwriting³.



James Johnson uses his neural interface to create art by blending images. Credit: Tyson Aflalo

So far, the vast majority of implants for recording long-term from individual neurons have been made by a single company: Blackrock Neurotech, a medical-device developer based in Salt Lake City, Utah. But in the past seven years, commercial interest in BCIs has surged. Most notably, in 2016, entrepreneur Elon Musk launched Neuralink in San Francisco, California, with the goal of connecting humans and computers. The company has raised US\$363 million. Last year, Blackrock Neurotech and several other newer BCI companies also attracted major financial backing.

Bringing a BCI to market will, however, entail transforming a bespoke technology, road-tested in only a small number of people, into a product that can be manufactured, implanted and used at scale. Large trials will need to show that BCIs can work in non-research settings and demonstrably improve the everyday lives of users — at prices that the market can support. The

timeline for achieving all this is uncertain, but the field is bullish. “For thousands of years, we have been looking for some way to heal people who have paralysis,” says Matt Angle, founding chief executive of Paradromics, a neurotechnology company in Austin, Texas. “Now we’re actually on the cusp of having technologies that we can leverage for those things.”

Interface evolution

In June 2004, researchers pressed a grid of electrodes into the motor cortex of a man who had been paralysed by a stabbing. He was the first person to receive a long-term BCI implant. Like most people who have received BCIs since, his cognition was intact. He could imagine moving, but he had lost the neural pathways between his motor cortex and his muscles. After decades of work in many labs in monkeys, researchers had learnt to decode the animals’ movements from real-time recordings of activity in the motor cortex. They now hoped to infer a person’s imagined movements from brain activity in the same region.

In 2006, a landmark paper⁴ described how the man had learnt to move a cursor around a computer screen, control a television and use robotic arms and hands just by thinking. The study was co-led by Leigh Hochberg, a neuroscientist and critical-care neurologist at Brown University in Providence, Rhode Island, and at Massachusetts General Hospital in Boston. It was the first of a multicentre suite of trials called BrainGate, which continues today.

“It was a very simple, rudimentary demonstration,” Hochberg says. “The movements were slow or imprecise — or both. But it demonstrated that it might be possible to record from the cortex of somebody who was unable to move and to allow that person to control an external device.”

Today’s BCI users have much finer control and access to a wider range of skills. In part, this is because researchers began to implant multiple BCIs in different brain areas of the user and devised new ways to identify useful signals. But Hochberg says the biggest boost has come from machine learning, which has improved the ability to decode neural activity. Rather

than trying to understand what activity patterns mean, machine learning simply identifies and links patterns to a user's intention.

"We have neural information; we know what that person who is generating the neural data is attempting to do; and we're asking the algorithms to create a map between the two," says Hochberg. "That turns out to be a remarkably powerful technique."

Motor independence

Asked what they want from assistive neurotechnology, people with paralysis most often answer "independence". For people who are unable to move their limbs, this typically means restoring movement.

One approach is to implant electrodes that directly stimulate the muscles of a person's own limbs and have the BCI directly control these. "If you can capture the native cortical signals related to controlling hand movements, you can essentially bypass the spinal-cord injury to go directly from brain to periphery," says Bolu Ajiboye, a neuroscientist at Case Western Reserve University in Cleveland, Ohio.

In 2017, Ajiboye and his colleagues described a participant who used this system to perform complex arm movements, including drinking a cup of coffee and feeding himself⁵. "When he first started the study," Ajiboye says, "he had to think very hard about his arm moving from point A to point B. But as he gained more training, he could just think about moving his arm and it would move." The participant also regained a sense of ownership of the arm.

Ajiboye is now expanding the repertoire of command signals his system can decode, such as those for grip force. He also wants to give BCI users a sense of touch, a goal being pursued by several labs.

In 2015, a team led by neuroscientist Robert Gaunt at the University of Pittsburgh in Pennsylvania, reported implanting an electrode array in the hand region of a person's somatosensory cortex, where touch information is

processed⁶. When they used the electrodes to stimulate neurons, the person felt something akin to being touched.

Gaunt then joined forces with Pittsburgh colleague Jennifer Collinger, a neuroscientist advancing the control of robotic arms by BCIs. Together, they fashioned a robotic arm with pressure sensors embedded in its fingertips, which fed into electrodes implanted in the somatosensory cortex to evoke a synthetic sense of touch¹. It was not an entirely natural feeling — sometimes it felt like pressure or being prodded, other times it was more like a buzzing, Gaunt explains. Nevertheless, tactile feedback made the prosthetic feel much more natural to use, and the time it took to pick up an object was halved, from roughly 20 seconds to 10.

Implanting arrays into brain regions that have different roles can add nuance to movement in other ways. Neuroscientist Richard Andersen — who is leading the trial at Caltech in which Johnson is participating — is trying to decode users' more-abstract goals by tapping into the posterior parietal cortex (PPC), which forms the intention or plan to move⁷. That is, it might encode the thought 'I want a drink', whereas the motor cortex directs the hand to the coffee, then brings the coffee to the mouth.

Andersen's group is exploring how this dual input aids BCI performance, contrasting use of the two cortical regions alone or together. Unpublished results show that Johnson's intentions can be decoded more quickly in the PPC, "consistent with encoding the goal of the movement", says Tyson Aflalo, a senior researcher in Andersen's laboratory. Motor-cortex activity, by contrast, lasts throughout the whole movement, he says, "making the trajectory less jittery".

This new type of neural input is helping Johnson and others to expand what they can do. Johnson uses the driving simulator, and another participant can play a virtual piano using her BCI.

Movement into meaning

"One of the most devastating outcomes related to brain injuries is the loss of ability to communicate," says Edward Chang, a neurosurgeon and

neuroscientist at the University of California, San Francisco. In early BCI work, participants could move a cursor around a computer screen by imagining their hand moving, and then imagining grasping to ‘click’ letters — offering a way to achieve communication. But more recently, Chang and others have made rapid progress by targeting movements that people naturally use to express themselves.

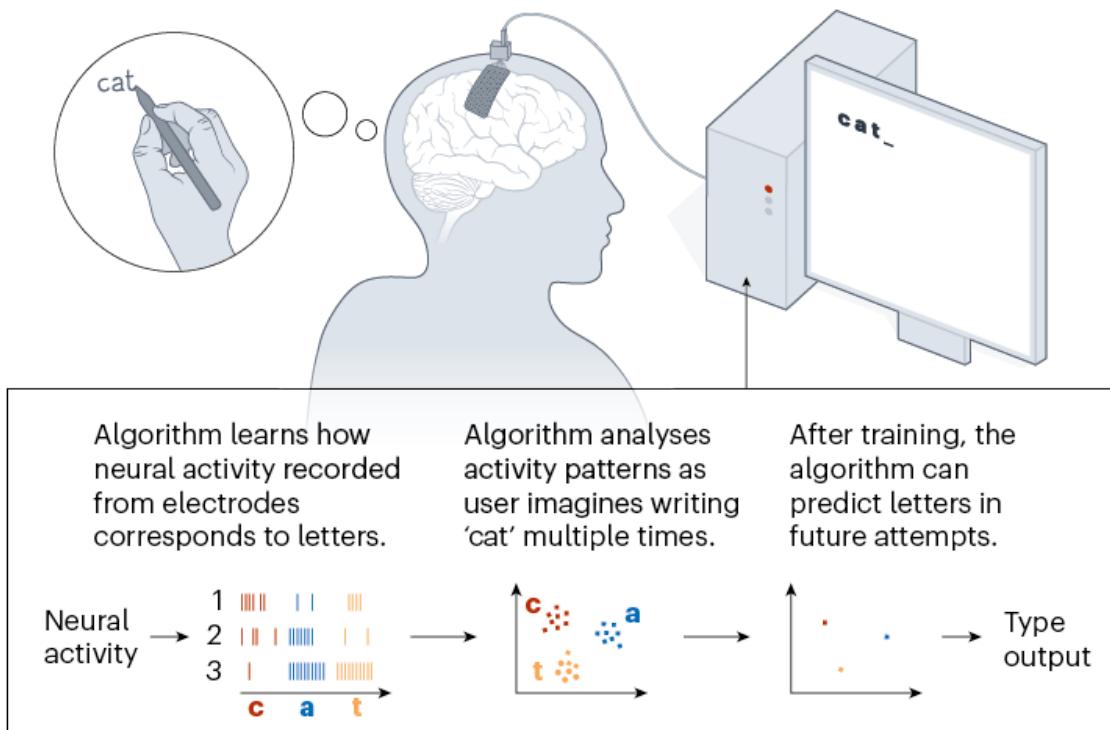
The benchmark for communication by cursor control — roughly 40 characters per minute⁸ — was set in 2017 by a team led by Krishna Shenoy, a neuroscientist at Stanford University in California.

Then, last year, this group reported³ an approach that enabled study participant Dennis Degrav, who can speak but is paralysed from the neck down, to double the pace.

Shenoy’s colleague Frank Willett suggested to Degrav that he imagine handwriting while they recorded from his motor cortex (see ‘Turning thoughts into type’). The system sometimes struggled to parse signals relating to letters that are handwritten in a similar way, such as r, n and h, but generally it could easily distinguish the letters. The decoding algorithms were 95% accurate at baseline, but when autocorrected using statistical language models that are similar to predictive text in smartphones, this jumped to 99%.

TURNING THOUGHTS INTO TYPE

To enable a man with paralysis to type, scientists implanted electrodes into the premotor area of his motor cortex and recorded neural activity while he imagined writing letters. They then used the activity to train a machine-learning algorithm. When the man imagines writing a word, the algorithm deciphers the neural activity to predict the letters, then corrects them using software that is similar to a smartphone's predictive text.



©nature

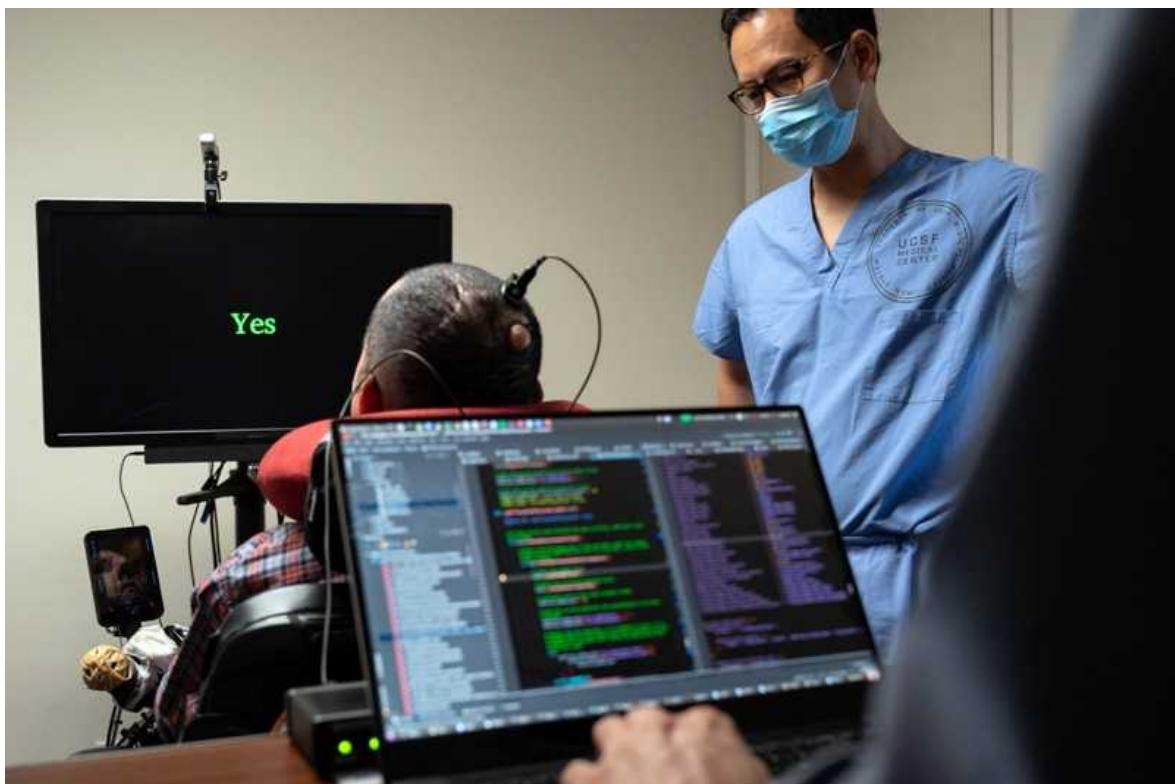
“You can decode really rapid, very fine movements,” says Shenoy, “and you’re able to do that at 90 characters per minute.”

Degray has had a functional BCI in his brain for nearly 6 years, and is a veteran of 18 studies by Shenoy’s group. He says it’s remarkable how effortless tasks become. He likens the process to learning to swim, saying, “You thrash around a lot at first, but all of a sudden, everything becomes understandable.”

Chang’s approach to restoring communication focuses on speaking rather than writing, albeit using a similar principle. Just as writing is formed of distinct letters, speech is formed of discrete units called phonemes, or

individual sounds. There are around 50 phonemes in English, and each is created by a stereotyped movement of the vocal tract, tongue and lips.

Chang's group first worked on characterizing the part of the brain that generates phonemes and, thereby, speech — an ill-defined region called the dorsal laryngeal cortex. Then, the researchers applied these insights to create a speech-decoding system that displayed the user's intended speech as text on a screen. Last year, they reported² that this device enabled a person left unable to talk by a brainstem stroke to communicate, using a preselected vocabulary of 50 words and at a rate of 15 words per minute. "The most important thing that we've learnt," Chang says, "is that it's no longer a theoretical; it's truly possible to decode full words."



Neuroscientist Edward Chang (right) at the University of California, San Francisco, helps a man with paralysis to speak through a brain implant that connects to a computer. Credit: Mike Kai Chen/The New York Times/Redux/eyevine

Unlike other high-profile BCI breakthroughs, Chang didn't record from single neurons. Instead, he used electrodes placed on the cortical surface that

detect the averaged activity of neuronal populations. The signals are not as fine-grained as those from electrodes implanted in the cortex, but the approach is less invasive.

The most profound loss of communication occurs in people in a completely locked-in state, who remain conscious but are unable to speak or move. In March, a team including neuroscientist Ujwal Chaudhary and others at the University of Tübingen, Germany, reported⁹ restarting communication with a man who has amyotrophic lateral sclerosis (ALS, or motor neuron disease). The man had previously relied on eye movements to communicate, but he gradually lost the ability to move his eyes.

The team of researchers gained consent from the man's family to implant a BCI and tried asking him to imagine movements to use his brain activity to choose letters on a screen. When this failed, they tried playing a sound that mimicked the man's brain activity — a higher tone for more activity, lower for less — and taught him to modulate his neural activity to heighten the pitch of a tone to signal 'yes' and to lower it for 'no'. That arrangement allowed him to pick out a letter every minute or so.

The method differs from that in a paper¹⁰ published in 2017, in which Chaudhary and others used a non-invasive technique to read brain activity. Questions were raised about the work and the paper was retracted, but Chaudhary stands by it.

These case studies suggest that the field is maturing rapidly, says Amy Orsborn, who researches BCIs in non-human primates at the University of Washington in Seattle. "There's been a noticeable uptick in both the number of clinical studies and of the leaps that they're making in the clinical space," she says. "What comes along with that is the industrial interest".

Lab to market

Although such achievements have attracted a flurry of attention from the media and investors, the field remains a long way from improving day-to-day life for people who've lost the ability to move or speak. Currently, study participants operate BCIs in brief, intensive sessions; nearly all must be

physically wired to a bank of computers and supervised by a team of scientists working constantly to hone and recalibrate the decoders and associated software. “What I want,” says Hochberg, speaking as a critical-care neurologist, “is a device that is available, that can be prescribed, that is ‘off the shelf’ and can be used quickly.” In addition, such devices would ideally last users a lifetime.

Many leading academics are now collaborating with companies to develop marketable devices. Chaudhary, by contrast, has co-founded a not-for-profit company, ALS Voice, in Tübingen, to develop neurotechnologies for people in a completely locked-in state.

Blackrock Neurotech’s existing devices have been a mainstay of clinical research for 18 years, and it wants to market a BCI system within a year, according to chairman Florian Solzbacher. The company came a step closer last November, when the US Food and Drug Administration (FDA), which regulates medical devices, put the company’s products onto a fast-track review process to facilitate developing them commercially.

This possible first product would use four implanted arrays and connect through wires to a miniaturized device, which Solzbacher hopes will show how people’s lives can be improved. “We’re not talking about a 5, 10 or 30% improvement in efficacy,” he says. “People can do something they just couldn’t before.”

Blackrock Neurotech is also developing a fully implantable wireless BCI intended to be easier to use and to remove the need to have a port in the user’s cranium. Neuralink and Paradromics have aimed to have these features from the outset in the devices they are developing.

These two companies are also aiming to boost signal bandwidth, which should improve device performance, by increasing the number of recorded neurons. Paradromics’s interface — currently being tested in sheep — has 1,600 channels, divided between 4 modules.

Neuralink’s system uses very fine, flexible electrodes, called threads, that are designed to both bend with the brain and to reduce immune reactions, says Shenoy, who is a consultant and adviser to the company. The aim is to

make the device more durable and recordings more stable. Neuralink has not published any peer-reviewed papers, but a 2021 blogpost reported the successful implantation of threads in a monkey's brain to record at 1,024 sites (see go.nature.com/3jt71yq). Academics would like to see the technology published for full scrutiny, and Neuralink has so far trialled its system only in animals. But, Ajiboye says, "if what they're claiming is true, it's a game-changer".

Just one other company besides Blackrock Neurotech has implanted a BCI long-term in humans — and it might prove an easier sell than other arrays. Synchron in New York City has developed a 'stentrode' — a set of 16 electrodes fashioned around a blood-vessel stent¹¹. Fitted in a day in an outpatient setting, this device is threaded through the jugular vein to a vein on top of the motor cortex. First implanted in a person with ALS in August 2019, the technology was put on a fast-track review path by the FDA a year later.



The 'stentrode' interface can translate brain signals from the inside of a blood vessel without the need for open-brain surgery. Credit: Synchron, Inc.

Akin to the electrodes Chang uses, the stentrode lacks the resolution of other implants, so can't be used to control complex prosthetics. But it allows

people who cannot move or speak to control a cursor on a computer tablet, and so to text, surf the Internet and control connected technologies.

Synchron's co-founder, neurologist Thomas Oxley, says the company is now submitting the results of a four-person feasibility trial for publication, in which participants used the wireless device at home whenever they chose. "There's nothing sticking out of the body. And it's always working," says Oxley. The next step before applying for FDA approval, he says, is a larger-scale trial to assess whether the device meaningfully improves functionality and quality of life.

Challenges ahead

Most researchers working on BCIs are realistic about the challenges before them. "If you take a step back, it is really more complicated than any other neurological device ever built," says Shenoy. "There's probably going to be some hard growing years to mature the technology even more."

Orsborn stresses that commercial devices will have to work without expert oversight for months or years — and that they need to function equally well in every user. She anticipates that advances in machine learning will address the first issue by providing recalibration steps for users to implement. But achieving consistent performance across users might present a greater challenge.

"Variability from person to person is the one where I don't think we know what the scope of the problem is," Orsborn says. In non-human primates, even small variations in electrode positioning can affect which circuits are tapped. She suspects there are also important idiosyncrasies in exactly how different individuals think and learn — and the ways in which users' brains have been affected by their various conditions.

Finally, there is widespread acknowledgement that ethical oversight must keep pace with this rapidly evolving technology. BCIs present multiple concerns, from privacy to personal autonomy. Ethicists stress that users must retain full control of the devices' outputs. And although current technologies cannot decode people's private thoughts, developers will have records of

users' every communication, and crucial data about their brain health. Moreover, BCIs present a new type of cybersecurity risk.

There is also a risk to participants that their devices might not be supported forever, or that the companies that manufacture them fold. There are already instances in which users were let down when their implanted devices were left unsupported.

Degray, however, is eager to see BCIs reach more people. What he would like most from assistive technology is to be able to scratch his eyebrow, he says. "Everybody looks at me in the chair and they always say, 'Oh, that poor guy, he can't play golf any more.' That's bad. But the real terror is in the middle of the night when a spider walks across your face. That's the bad stuff."

For Johnson, it's about human connection and tactile feedback; a hug from a loved one. "If we can map the neurons that are responsible for that and somehow filter it into a prosthetic device some day in the future, then I will feel well satisfied with my efforts in these studies."

Nature **604**, 416–419 (2022)

doi: <https://doi.org/10.1038/d41586-022-01047-w>

References

1. Flesher, S. *et al.* *Science* **372**, 831–836 (2021).
2. Moses, D. A. *et al.* *N. Engl. J. Med.* **385**, 217–227 (2021).
3. Willett, F. R. *et al.* *Nature* **593**, 249–254 (2021).
4. Hochberg, L. R. *et al.* *Nature* **442**, 164–171 (2006).
5. Ajiboye, A. B. *et al.* *Lancet* **389**, 1821–1830 (2017).
6. Flesher, S. *et al.* *Sci. Transl. Med.* **8**, 361ra141 (2016).

7. Aflalo, T. *et al.* *Science* **348**, 906–910 (2015).
8. Pandarinath, C. *et al.* *eLife* **6**, e18554 (2017).
9. Chaudhary, U. *et al.* *Nature Commun.* **13**, 1236 (2022).
10. Chaudhary, U., Xia, B., Silvoni, S., Cohen, L. G. & Birbaumer, N. *PLoS Biol.* **15**, e1002593 (2017); retraction **17**, e3000607 (2019).
11. Oxley, T. J. *et al.* *Nature Biotechnol.* **34**, 320–327 (2016).

This article was downloaded by **calibre** from <https://www.nature.com/articles/d41586-022-01047-w>

| [Section menu](#) | [Main menu](#) |

Opinion

- **[Analysis: the biodiversity footprint of the University of Oxford](#)** [20 April 2022]

Comment • Every large organization should plot a path to net gain in biodiversity — here's how.

- COMMENT
- 20 April 2022

Analysis: the biodiversity footprint of the University of Oxford

Every large organization should plot a path to net gain in biodiversity — here's how.

- [Joseph William Bull](#)⁰,
- [Isobel Taylor](#)¹,
- [Elizabeth Biggs](#)²,
- [Henry M. J. Grub](#)³,
- [Tom Yearley](#)⁴,
- [Harriet Waters](#)⁵ &
- [E. J. Milner-Gulland](#)⁶



The Botanic Garden at the University of Oxford, UK. Credit: Greg Smolonski/Oxford University Images

To help to achieve ecological recovery worldwide, more multinational corporations are making commitments to biodiversity conservation^{1–3}. According to the most recent assessment in 2018, 31 of the 100 largest companies by revenue worldwide (the global Fortune 100) have done so, from the retail corporation Walmart to the insurance company AXA⁴.

To deliver real gains — in the population sizes of endangered species, say, or in the number of hectares of restored forests, grasslands or wetlands — large organizations need to determine which of their activities have the greatest impacts on biodiversity⁵. And they need to disclose and mitigate those impacts. Currently, methods for doing this are lacking (see ‘Promises are hard to keep’). (By large organizations, we mean formal entities composed of hundreds of people or more that act towards a certain purpose, whether in the public, private or non-profit sectors.)

Promises are hard to keep

A lack of consensus on methods and metrics means companies are struggling to clearly define — and deliver on — commitments relating to biodiversity.

So far, most studies of the environmental impacts of organizations, such as multinational corporations and universities, have focused on greenhouse-gas emissions.

The G7 group of the world's largest economies endorsed the new Taskforce on Nature-related Financial Disclosures (TNFD) only last year. This builds on a similar approach used for climate change — the Taskforce on Climate-related Financial Disclosures. The TNFD aims to guide organizations on how to disclose environmental harms tied to their activities, but is still being developed.

The number of corporations making commitments to achieve ‘net gain’ or ‘no net loss’ outcomes in relation to biodiversity has risen steadily in the past two decades³. But some of these promises have subsequently been retracted. In 2016, for example, the mining corporation Rio Tinto moved away from its 2006 agenda-setting ‘net positive impact’ biodiversity commitment, reportedly to focus on minimizing impacts³ (see also go.nature.com/3xtjggo).

Many other commitments are not quantitative. As of 2018, only 5 of the 31 global Fortune 100 companies making biodiversity-related commitments had provided ones that were SMART — specific, measurable, ambitious, realistic and time-bound⁴ (the global Fortune 100 is an annual list of the 100 largest firms worldwide by revenue, as ranked by *Fortune* magazine).

When quantitative analyses have been done, they tend to be of limited use, mainly because of inconsistencies in the biodiversity metrics used, and limitations in the scope of the assessment made. Disclosure of results is also limited.

When quantitative analyses have been done, a variety of metrics have been used to quantify impacts. These range from the proportion of local species that would be lost as a result of an activity, to factors such as hectares of habitat affected, or the amount of sustainably sourced paper, fish or palm oil

that is used⁴. But the choice of metric can radically alter the results of an impact assessment, so it is difficult to compare organizations. Likewise, few analyses consider the impacts of activities that are not under the direct control of the organization, such as those associated with supply chains⁶.

As a proof of principle, we conducted a comprehensive assessment of biodiversity losses associated with activities at the University of Oxford, UK. We used data on purchasing, travel bookings, utility bills and other information from the 2018–19 and 2019–20 academic years. The 60 activities we assessed included the day-to-day running of buildings and transport services; travel (including flights) for students and researchers; construction of laboratories and other buildings; consumption of food and beverages at restaurants and cafeterias; and use of medical supplies and other materials in research labs.

Our aim was to demonstrate what it would take for a large organization such as the University of Oxford to bring about a net gain in biodiversity — meaning that, thanks to its actions, the world's biodiversity is left in a better state than it was before. As part of our analysis, we assessed how the university's various activities and operations also affect greenhouse-gas emissions, and how those, in turn, affect biodiversity by driving climate change.

We are confident that the approach we've developed for Oxford could be applied more broadly. Indeed, we hope that such a well-known institution disclosing a full assessment of its biodiversity footprint will offer powerful inspiration for others. (All seven of us have a current or previous affiliation with the university.)

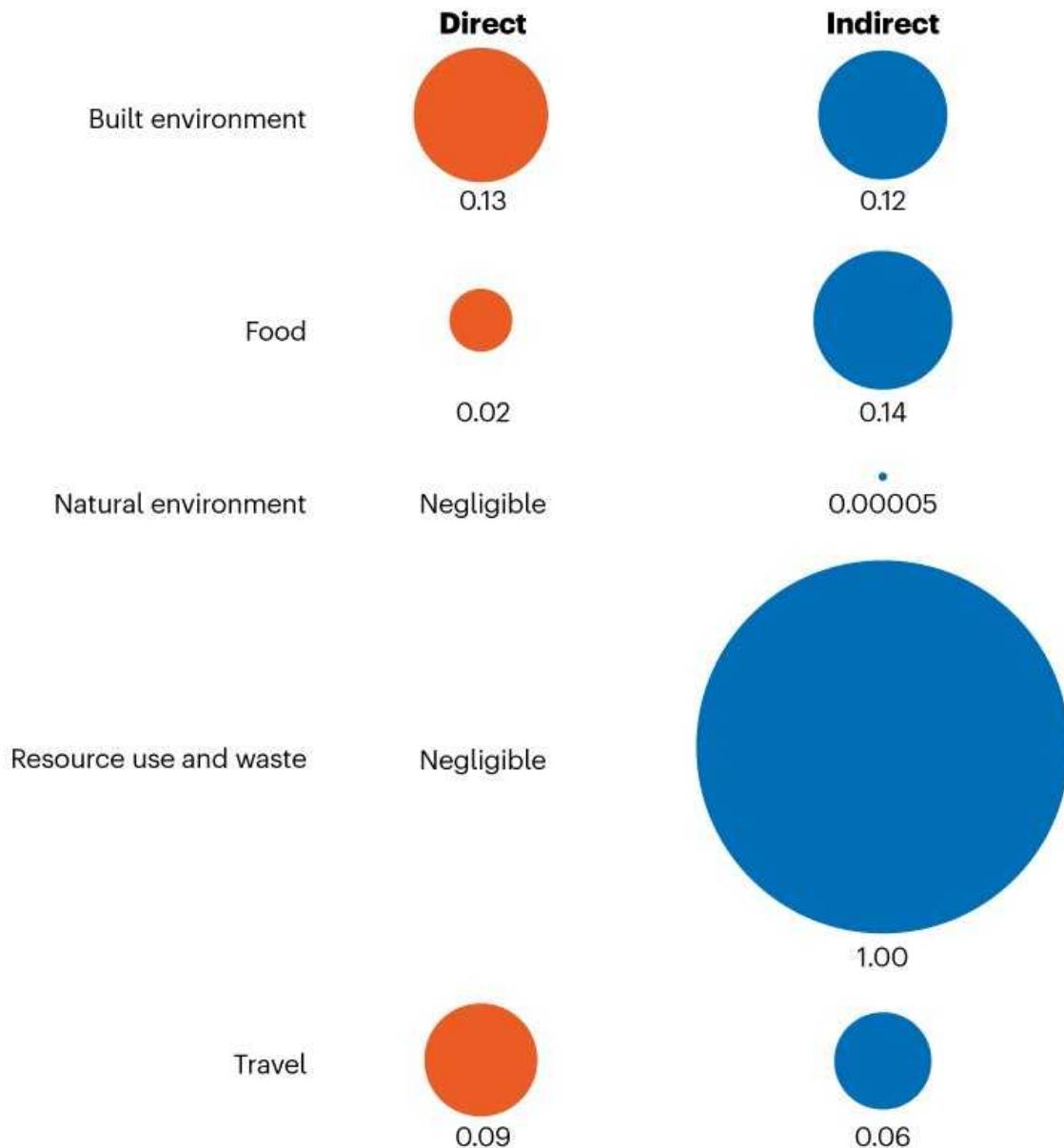
What we did

The University of Oxford launched an ambitious environmental sustainability strategy in March 2021. Its two main goals are to achieve biodiversity net gain and net-zero carbon, both by 2035. (The latter means that the university will remove as much carbon from the atmosphere as it adds.)

To understand how challenging these goals might be to fulfil, we assessed the environmental impacts of the university's various activities. This covered all those to do with research, education and operations during an academic year for staff and students (see 'Upstream effects'). For our purposes, operations includes the university transport fleet, consumption of departmental food and utilities, waste disposal and the operational supply chain, including for paper.

UPSTREAM EFFECTS

The University of Oxford's biggest impact on biodiversity* is from the indirect effects of resource use and waste in external supply chains it does not control.



*As measured by local relative species loss for each impact category (see M. A. J. Huijbregts *et al.* *Int. J. Life Cycle Assess.* **22**, 138–147 (2017) for method).

©nature

Source: J. W. Bull *et al.*

As a first step, we defined a conceptual framework to systematically categorize the environmental impacts. We grouped activities in research, education and operations according to whether they involved any of five

features: travel; food; the built environment (university buildings); the natural environment (any green space or land owned by the university, including managed parks and gardens); and resource use and waste (see ‘What we left out’). Each of these is associated with five general environmental impacts: greenhouse-gas emissions, the use of land and water, and pollution of water and air.

What we left out

Other organizations could assess different types of impact on biodiversity.

Our biodiversity analysis of the University of Oxford, UK, included most upstream impacts — those resulting from consumption of goods and services created outside the university, such as food or medical supplies. We excluded the downstream impacts of research and education, such as those of a discovery in gene editing or chemistry, because it would be impossible to comprehensively account for all of the environmental impacts of knowledge generation. Also not included in our analysis were the university’s 39 colleges, 6 permanent private halls and more than 260 commercial buildings. These are independent legal entities that manage sustainability issues separately.

Other analyses in different sectors might well be able to include downstream impacts. The effects of discarded plastic bottles or clothes could be included for a soft-drinks company or clothing manufacturer, for example.

To further categorize the environmental impacts, we assigned each activity to one of two groups: those under direct university control or influence (through staff and key contractors), and those that the university can influence only indirectly (through students and supply chains). We deemed students buying tuna sandwiches from a university-owned cafe as direct control, for instance, because the university could decide to serve only vegetarian food. However, it can influence only indirectly what happens up the supply chain, before materials are used in a research lab, for example.

Using this organizational framework, we worked with administrators to obtain the relevant information, such as travel bookings for staff and

students, electricity and water bills, and purchasing records for goods, services and materials used in construction projects.

Next, we used various tools to convert all the activities data into estimates of ‘mid-point environmental impacts’ (amount of carbon dioxide emitted, land or water used, and air or water pollutants produced). The database Exiobase 3 is one of the most extensive sources of international supply-chain impacts worldwide⁷. It shows, for instance, that the roughly US\$3.5 million the university spent on paper and paper products in 2019–20 contributed to atmospheric acidification by releasing 2,448 kilograms of sulfur dioxide equivalent. Similarly, the UK Higher Education Supply Chain Emissions Tool uses spending data on goods and services to estimate greenhouse-gas emissions. The roughly \$23 million Oxford spent on personal computers, printers and calculators in 2019–20, for example, produced an estimated 20,105 tonnes of CO₂ equivalent.

We then needed to estimate the extent of biodiversity loss associated with this wide range of broad environmental impacts. So we converted the mid-point environmental impacts into ‘end-point impacts’ that are specifically concerned with biodiversity. To do this, we used an established conversion methodology called ReCiPe⁸. The output metric ultimately linked to each activity is based on the proportion of local species that would be lost as a result of that activity, relative to the number that exists currently (see Supplementary information for all results and conversion factors).

Caveats

We refined our methods slightly when analysing data from the 2019–20 academic year. This, combined with the disruption caused by the COVID-19 pandemic, makes it difficult to compare years. So for simplicity, we report our results only from the 2019–20 academic year.

The biodiversity metric we obtain using ReCiPe is based on strong evidence: the conversion tool is derived from the results of hundreds of studies of the impacts of human pressures on biodiversity⁸. But, in general, we weren’t able to factor in fine-level variables, such as whether the beef steaks in a university-owned restaurant are sourced from a UK or Brazilian farm. As

such, our approach is best seen as a way to evaluate relative impacts, rather than as an indicator of precise absolute impacts.

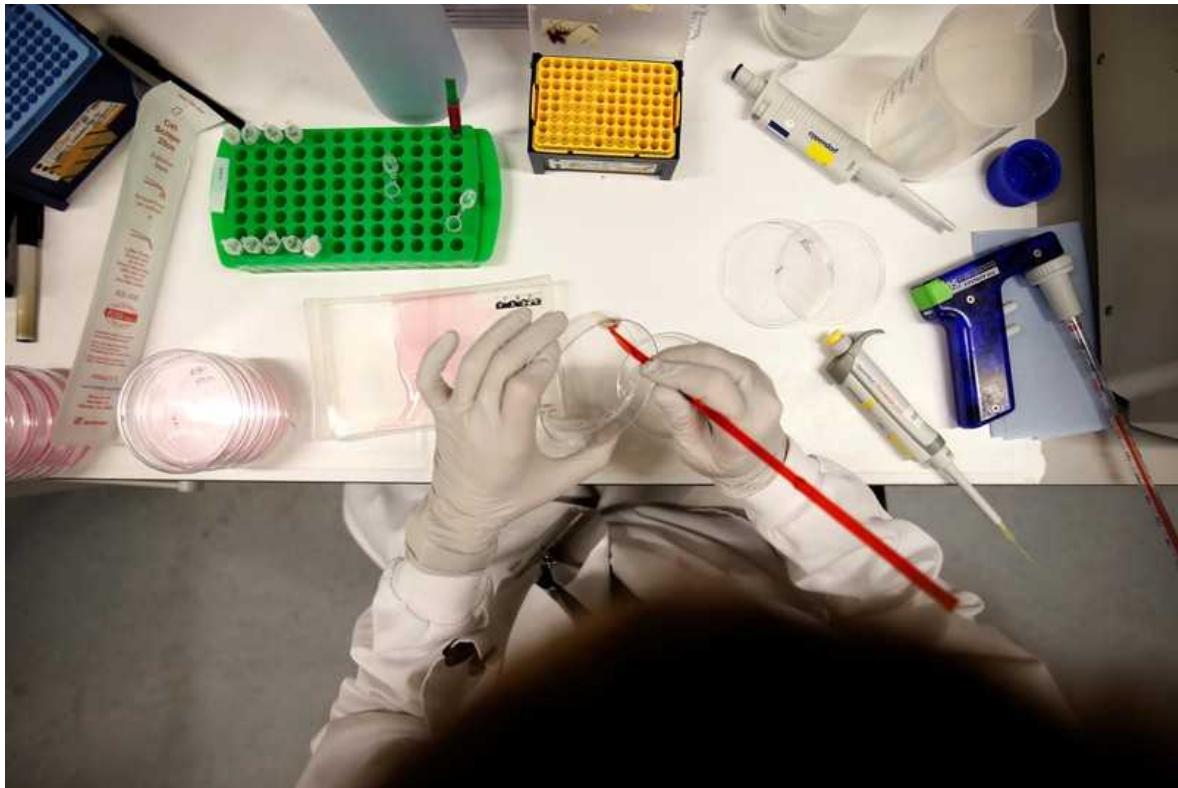
This difficulty aside, it is hard to compare the impact of the University of Oxford on biodiversity with that of similarly sized organizations. As yet, and as far as we know, no other organization has comprehensively evaluated and disclosed its impact on biodiversity, and then had its assessment independently validated. (Ecologists and other stakeholders at the University of Jyväskylä in Finland have begun to explore the impacts of that university's activities on biodiversity [using a similar approach to ours](#).)

Using the greenhouse-gas metric, however, we can compare the impacts of the University of Oxford on emissions (which are related to its impacts on biodiversity) with those of comparably sized organizations.

What we found

The absolute size of the university's greenhouse-gas footprint is astonishingly large — comparable to that of the eastern Caribbean island nation of Saint Lucia. It is two orders of magnitude smaller than Microsoft's greenhouse-gas footprint, but one order of magnitude larger than that of the London Stock Exchange, according to estimates publicly disclosed by those organizations.

Perhaps the most striking finding in our assessment of impacts specifically on biodiversity is that most of the harms are tied to university activities that are not under its direct control. In fact, the activities with the five biggest impacts on biodiversity are (from biggest to smallest): the supply chain for research activities (such as for chemicals, medical products, organic tissue and plastics); the supply chain for the day-to-day running of buildings (for paper, information technology and so on); food consumption; electricity consumption; and the supply chain for construction. All of these activities are associated with resource use and waste, food and the built environment.



The University of Oxford's use of laboratory materials has a large impact on biodiversity because of the upstream supply chain. Credit: Peter Nicholls/Reuters

In short, supplies of lab equipment have much greater impacts on biodiversity overall than do international flights, the university's consumption of electricity or its use of construction materials. (Personal protective equipment used in the lab, for example, requires the extraction and industrial processing of hydrocarbons, often from areas that are rich in biodiversity.)

This observation is in line with the results of a handful of studies that suggest that supply chains, not transport or the day-to-day running of buildings, are the main contributors to greenhouse-gas emissions for universities (see, for example, ref. 9). It also aligns with the results of assessments by the fashion giant Kering since 2012, using its Environmental Profit & Loss account — a tool designed to quantify the environmental impacts of the company's activities. These have revealed that Kering's procurements of commodities, such as leather, wool and metals, have much

more impact on greenhouse-gas emissions, particularly on those from land use, than does the day-to-day running of its factories and offices¹⁰.

Yet the sustainability strategies of large organizations typically focus not on supply chains, but on recycling, reducing the number of flights people take or the amount of electricity used^{11–13} (see also *Nature* 546, 565–567; 2017).

Another important finding is the scale of intervention needed. Restoring the university's owned land (around 1,000 hectares) to native woodland or some other natural habitat would make little difference when it comes to compensating for the impacts on biodiversity that result from just one year of activity. The university colleges own much more land than the university itself — some 50,000 hectares — but we excluded them from our analysis because they are independent legal entities that manage sustainability issues separately.

Biodiversity boost

How could the university reverse the biodiversity losses stemming from its activities and operations?

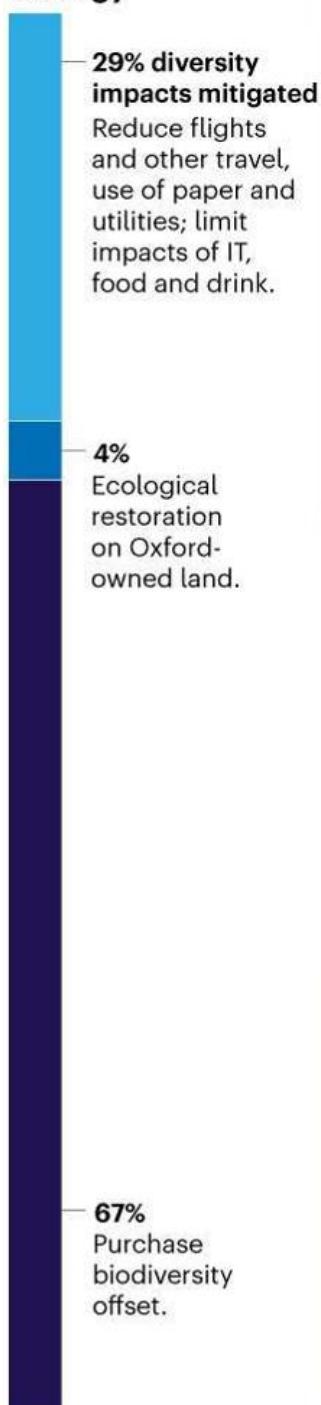
Here we consider three options. It could pursue its current environmental sustainability strategy. This entails (among other steps) setting quantitative targets to reduce flights, limiting consumption of all single-use products, making university-catered food vegetarian by default, and achieving 20% net gain for biodiversity in new construction projects. Alternatively, it could focus more heavily on preventing harms to biodiversity. We model a scenario in which all staff flights are prevented, all use of paper and any further construction is stopped, and the purchasing of lab materials is halved. Or the university could focus on compensating for the impacts that its activities and operations have on the planet, by taking steps to increase biodiversity in other places (see ‘Oxford’s options’).

OXFORD'S OPTIONS

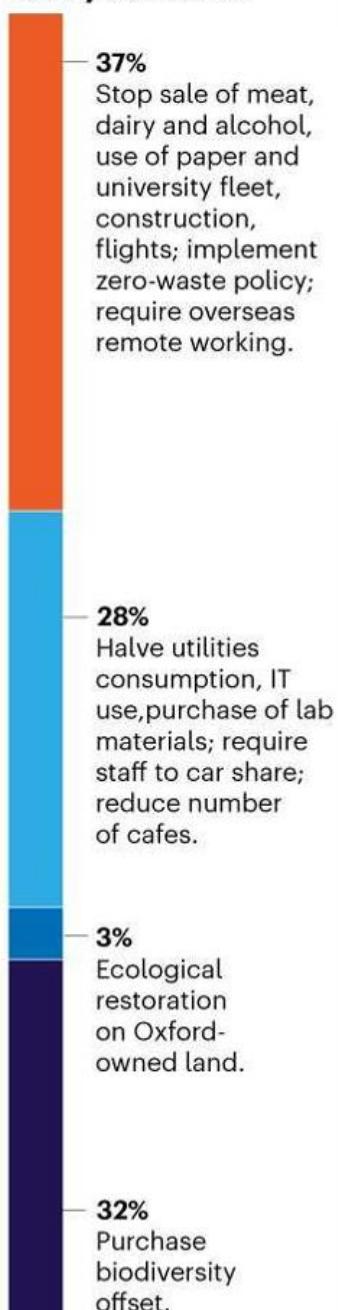
To achieve no net loss of biodiversity, the University of Oxford could focus more heavily on preventing harms to biodiversity (option 1). Or it could try to compensate for the impacts that its activities and operations have on the planet (option 2).

■ Avoid ■ Minimize ■ Remediate ■ Offset

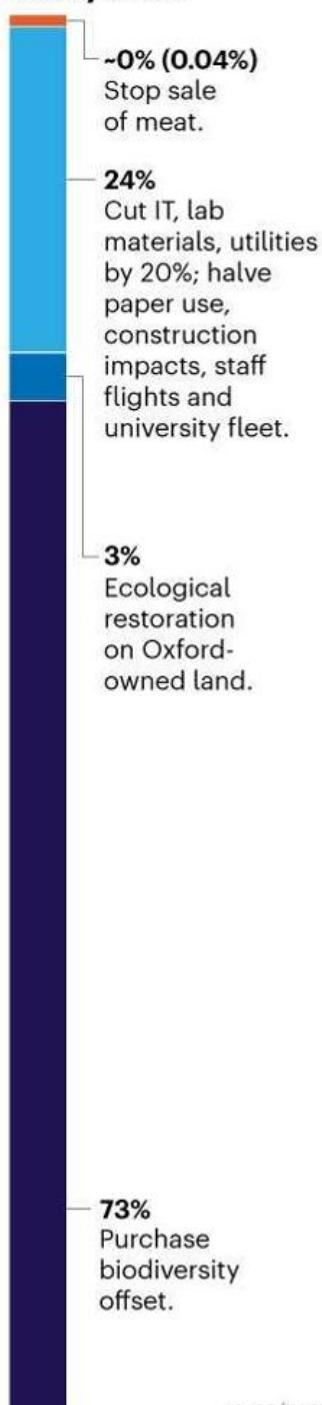
Current strategy



Option 1: Heavy avoidance



Option 2: Heavy offset



Source: J. W. Bull *et al.*

Using the 2018–19 academic year results (selected because the COVID-19 pandemic made 2019–20 so unusual), we estimated how far these mitigation strategies might take the university towards biodiversity net gain.

Our analysis indicates that the set of preventive measures proposed under the university’s environmental sustainability strategy get it about one-third of the way towards net gain. The findings also indicate that focusing mainly on the prevention of impacts is operationally unfeasible. Activities that have most effect on biodiversity, such as purchasing lab consumables, are central to the university’s existence and cannot simply stop.

To achieve net gain, preventive measures, such as reducing flights and paper use, will have to be accompanied by additional and extensive actions to compensate for the remaining impacts on biodiversity.

Such actions could include investing in reforestation, wetland restoration, sustainable land-management programmes and prevention of habitat loss caused by independent parties. For example, those directing the Oyu Tolgoi mining project in Mongolia are seeking to achieve biodiversity net gain by spending around 0.6% of the total project cost on actions that benefit biodiversity, such as sustainable grazing practices (see go.nature.com/3tkkbjh). Similarly, the Ambatovy metals mine in Madagascar is on course to offset its impacts on biodiverse eastern rainforests by preventing deforestation of those same habitats through small-scale agriculture¹⁴.

Achieving true biodiversity net gain will require substantial offsetting that does not necessarily contribute to the university’s reductions in greenhouse-gas emissions. But whatever mix of approaches the institution pursues, it should strive for win–wins on both biodiversity and climate.

Many types of action can simultaneously increase biodiversity and reduce greenhouse-gas emissions. For example, restoring mangroves in Bangladesh increased populations of wintering water birds 20-fold in just three years from 2004. And these restored mangroves can absorb carbon four times faster than land-based forests can¹⁵. But in other cases, there are trade-offs.

Constructing wind turbines and solar photovoltaics to produce renewable energy, for instance, requires extensive mining of metals in places that can be rich in biodiversity^{[16](#)}.

Net gain for other organizations

Our calculations are likely to be comparable to results for other universities. In our analysis, we do not include the impacts of individual colleges. But because similar kinds of activity occur in colleges as in the rest of the university, their inclusion — or of halls of residence at other universities — is unlikely to qualitatively change our main findings. In fact, because of the colleges' unusually large land holdings, including them would arguably result in an assessment that doesn't so easily compare with those of other universities.

Crucially, however, the analytical framework we have developed can be applied to a wide range of large organizations — whether they be universities, multinational corporations or government institutions.



Restoring mangroves in western Bangladesh increased populations of wintering water birds, such as this oriental darter (*Anhinga melanogaster*). Credit: Muhammad Mostafiqur Rahman/Alamy

Governments, intergovernmental organizations and multinational corporations are increasingly recognizing that it will not be enough to simply slow the loss of the world's biodiversity. Damaged habitats and depleted natural resources must be restored to prevent the collapse of ecosystems.

Last year, the United Nations called for the urgent revival of nature in farmlands, forests and other ecosystems, declaring 2021–30 to be the Decade on Ecosystem Restoration. Later this year, at a meeting in Kunming, China, it is hoped that 196 nations will agree to the Post-2020 Global Biodiversity Framework of the Convention on Biological Diversity. Among the goals listed in the draft document are a “net gain in the area, connectivity and integrity of natural systems of at least 5 per cent”¹⁷.

We urge all large organizations, academic or otherwise, to commit to strategies for a net gain in biodiversity — and to adopt formalized approaches that quantify current impacts and allow transparent tracking of progress. Otherwise, the degree of worldwide recovery of natural resources increasingly recognized as crucial for resilient societies to function will not happen.

A key challenge is the lack of traceability for commodities. Both our assessment of the University of Oxford and those of others have revealed that large organizations often don't know which country their commodities (such as cotton, flour or cement) come from — let alone which supplier or what kinds of biodiversity are being affected as a result.

[According to its 2022 report](#), for example, even a sector leader such as Kering could trace the source of only about three-quarters of its cotton. Supply chains for other commodities, such as sand, are even more opaque¹⁸.

Encouragingly, various initiatives are being developed to provide more transparency about environmental impacts across supply chains. These

include the supply-chain mapping tool [TRASE](#), which aims to address deforestation.

A related challenge, covered extensively elsewhere^{[19,20](#)}, is how to ensure that biodiversity offsets are effectively and appropriately implemented such that they lead to conservation outcomes that are truly additional.

Currently, there are uncertainties around how long it takes for a restored forest to start delivering biodiversity gains, whether promises to offset harms to biodiversity are actually met, what level of biodiversity gain is delivered by the restoration of a particular habitat, and so on. Take the Ambatovy mine in Madagascar. Its directors began protecting areas of eastern rainforest in 2009 to offset the impacts of deforestation directly caused by the mine. Yet forest gains are not estimated to balance losses until sometime between 2018 and 2033^{[14](#)}.

Despite such challenges, however, we think that a commitment to full transparency, and to improving data collection over time, will enable organizations to compare performance and drive change — both in their own operations and throughout supply chains.

Time is too short to let the perfect be the enemy of the good, or to claim that biodiversity net gain is too hard to achieve because there is no universal biodiversity metric. Individual metrics are imperfect but improving, and their limitations should not be a reason to delay measuring, disclosing and tackling impacts on biodiversity.

Nature **604**, 420–424 (2022)

doi: <https://doi.org/10.1038/d41586-022-01034-1>

References

1. Folke, C. *et al. Nature Ecol. Evol.* **3**, 1396–1403 (2019).
2. Barbier, E. B., Burgess, J. C. & Dean, T. J. *Science* **360**, 486–488 (2018).

3. de Silva, G. C., Regan, E. C., Pollard, E. H. B. & Addison, P. F. E. *Bus. Strategy Environ.* **28**, 1481–1495 (2019).
4. Addison, P. F. E., Bull, J. W. & Milner-Gulland, E. J. *Conserv. Biol.* **33**, 307–318 (2019).
5. Bull, J. W. *et al.* *Nature Ecol. Evol.* **4**, 4–7 (2020).
6. Moran, D., Giljum, S., Kanemoto, K. & Godar, J. *One Earth* **3**, 5–8 (2020).
7. Ciroth, A. & Bunsen, J. *Exiobase 3.4 in openLCA* (GreenDelta, 2019).
8. Huijbregts, M. A. J. *et al.* *Int. J. Life Cycle Assess.* **22**, 138–147 (2017).
9. Thurston, M. & Eckelman, M. J. *Int. J. Sustain. Higher Educ.* **12**, 225–235 (2011).
10. Kering. *Environmental Profit & Loss (EP&L) 2020 Group Results* (Kering, 2021); available at go.nature.com/3tnw5s2.
11. Findler, F., Schönherr, N., Lozano, R., Reider, D. & Martinuzzi, A. *Int. J. Sustain. Higher Educ.* **20**, 23–38 (2019).
12. Li, X., Tan, H. & Rackes, A. *J. Cleaner Prod.* **106**, 97–108 (2015).
13. Klöwer, M., Hopkins, D., Allen, M. & Higham, J. *Nature* **583**, 356–359 (2020).
14. Devenish, K., Desbureaux, S., Willcock, S. & Jones, J. P. G. *Nature Sustain.* <https://doi.org/10.1038/s41893-022-00850-7> (2022).
15. Smith, A. C. *et al.* *Front. Environ. Sci.* <https://doi.org/10.3389/fenvs.2021.737659> (2021).
16. Sonter, L. J., Dade, M. C., Watson, J. E. M. & Valenta, R. K. *Nature Commun.* **11**, 4174 (2020).

17. Convention on Biological Diversity. *First Draft of the Post-2020 Global Biodiversity Framework* CBD/WG2020/3/3 (CBD, 2021).
18. Torres, A., Brandt, J., Lear, K. & Liu, J. *Science* **357**, 970–971 (2017).
19. Maron, M. *et al.* *BioScience* **66**, 489–498 (2016).
20. zu Ermgassen, S. O. S. E. *et al.* *Conserv. Lett.* **12**, e12664 (2019).

This article was downloaded by **calibre** from <https://www.nature.com/articles/d41586-022-01034-1>

| [Section menu](#) | [Main menu](#) |

Work

- **First-in-family scholars bust generational barriers** [19 April 2022]
Career Feature • First-generation students face challenges navigating graduate school. How can institutions better support them?
- **Ultracold storage ensures a future for endangered plants**
[19 April 2022]
Where I Work • Plant scientist Liang Lin develops cryopreservation technology to store seeds and plant tissues at liquid-nitrogen temperature to conserve biodiversity.

- CAREER FEATURE
- 19 April 2022

First-in-family scholars bust generational barriers

First-generation students face challenges navigating graduate school. How can institutions better support them?

- [Nikki Forrester](#) 0

[Find a new job](#)



Qaphelani Ngulube, a first-generation student from Zimbabwe, faces unique challenges as a graduate student.Credit: Donald Tapfuma

“As a first-generation student, the challenges start before grad school,” says Jisub Hwang, a PhD candidate at the Korea Polar Research Institute in Incheon, South Korea. Hwang is the first person in his family to attend university. “It’s tough to help my family understand the long journey of getting a PhD. They want me to earn money at a company.”

Hwang says South Korean undergraduates typically secure industry jobs after earning their degrees, but he realized this path wasn’t for him, given his passion for polar research. In a conversation with his parents about his career plans, Hwang explained why going to graduate school makes him happy and is necessary to achieve his long-term career goal of becoming an independent researcher. “The more they understand about graduate-school life, the more they support me and respect my decision,” he says.

Having difficult conversations with family members is just one of many challenges first-generation graduate students can face. For example, parents are often proud of their children’s undergraduate achievements, but they might worry about the financial instability that comes with more years of study. Students who are the first in their families to attend university might struggle with imposter syndrome and feel guilty for leaving their families or not financially contributing enough to them. Although many first-generation students are interested in pursuing graduate degrees, they might not have the financial resources or guidance to navigate applying to graduate school and to thrive in the academic environment¹.

Despite the challenges, first-generation students make up a large portion of undergraduate and graduate students. In 2016, [56% of undergraduate students at US institutions](#) were first-generation, and approximately 30% of US PhD recipients are first-generation². In England, two-thirds of graduates are the first in their families to attend univerrsity³.

“That’s a tremendous amount of students,” says Sarah Whitley, assistant vice-president of the Center for First-generation Student Success, an initiative based in Washington DC that provides information about advancing the success of first-generation students and practices for doing so.

“But institutions are such complex bureaucratic and jargon-filled entities that we are making it difficult for first-generation students to access the support and resources imperative to their success.”

Opening doors

First-generation students, who are more likely to come from lower income families and to belong to under-represented groups in science², often lack the financial resources, research opportunities and mentorship needed to make graduate school a viable option. “If you’re the first in your family, you may not have the cultural capital to understand how to navigate graduate education,” says Whitley.

“The biggest challenge is the lack of awareness,” says Arnaldo Díaz Vázquez, assistant dean for diversity inclusion at the University of Texas Southwestern Medical Center in Dallas. It can be difficult for students to determine how to pick the right graduate programmes, apply and advocate for themselves if they don’t know someone who’s gone through the process, he says.

Some first-generation students are even discouraged from pursuing university degrees after they graduate from high school. Natalia Jagielska, who moved with her parents from Poland to Bolton, UK, during the 2007 economic crisis, recalls career advisers and teachers telling her that she wouldn’t be competitive for a place at a research university after attending a state, rather than private, secondary school. “My grades were lower than my colleagues’,” says Jagielska. But that didn’t prevent her from pursuing an integrated bachelor’s and master’s programme in geochemistry and palaeobiology at the University of Manchester, UK.

“My priority was finding an integrated master’s, otherwise I wasn’t able to afford it,” says Jagielska, now a PhD candidate studying Jurassic pterosaur evolution at the University of Edinburgh, UK. Because her master’s programme was integrated, it took only eight months, and she could use a government student loan to cover both degrees. To save money at university, she lived at home with her parents and worked part-time. She also received

support through the [Manchester Access Programme](#), which helps to prepare local secondary-school students for university.

Along with funding, the access programme helped Jagielska to develop the knowledge and skills she needed to be successful as an undergraduate student. “They showed us how to read research papers, do referencing, write essays and get to know what university life looks like,” she says.

Although she’s grateful for the programme’s support, she thinks universities could do more to support first-generation students. For instance, university administrators could account for discrepancies in the quality of secondary education that students receive: state and private schools can vary greatly in their resources and staff. “We are competing on the same exam board but coming from very different starting points,” she says. “Universities have resources to help you, but the issue is that people don’t access them.”

The cost of graduate education is another significant barrier. “I waited 15 years to get my master’s degree,” says Qaphelani Ngulube, who graduated from university in 2005. Despite years of effort, he didn’t receive a scholarship until 2020 to pursue a master’s degree in chemistry at Uppsala University in Sweden. Ngulube says many students from his home country, Zimbabwe, aren’t fully aware of the available opportunities, or they don’t apply for scholarships because they don’t feel that their achievements are competitive enough for them to receive funding for studying abroad. “If they don’t get a scholarship, they can’t pay the graduate-school fees because many people don’t have that much saved,” he says.

Since starting his master’s programme, Ngulube has supported students from Zimbabwe and other countries who are interested in pursuing graduate degrees. He points them to a variety of scholarship, university and study-abroad websites to find funding opportunities, such as [Studyportals’ global scholarships portal](#). Talking to people who have secured scholarships can help first-generation students to understand the application process and build confidence in their ability to earn financial support, says Ngulube.



Natalia Jagielska, a first-generation PhD student, studies a pterosaur fossil. Credit: Callum Bennetts, Maverick Photo Agency

Graduate application fees can also be expensive. Although many institutions offer waivers, fellowships, scholarships and other forms of financial aid, some students might not be aware that they're available. "It's the issue of exposure," he adds.

At an institutional level, undergraduate programmes can do more to help first-generation students determine whether graduate studies are right for them, says Whitley. "More institutions are growing their academic-aid funds, particularly for first-generation students, where students can apply for support to take graduate school standardized tests, visit an institution that they are interested in and get study materials or books," she says. "Some even help them get clothing for a graduate school interview."

Staying on track

Once first-generation students start graduate studies, they must navigate the academic responsibilities of their programmes. Although the overall adjustments to graduate school aren't necessarily unique for first-generation students, they can be compounded by external factors, such as being far from home and not having a support system of people who understand academia. "Non-first-generation students can get expert guidance both on campus and at home, but I can only get that from campus and outsourcing," says Ngulube.

Undergraduate training doesn't necessarily groom first-generation students to excel in graduate programmes. "I wasn't prepared for graduate school at all," says Silindile Maphosa, who is from Zimbabwe and is currently pursuing a PhD in biotechnology at the University of Pretoria in South Africa. "I couldn't go back to my family to ask, 'Can you please tell me how this works?'" she describes. Instead, she advises new students to quickly find their own network of people to help them navigate university systems.

After months of struggling to manage a new e-mail system, coursework expectations, meetings and research, Maphosa reached out to a more senior graduate student in her laboratory, who was also a first-generation student, for help. "I decided to open up to her about how stressed I was without anyone to help me. Then I realized that she'd gone through the same thing."

Maphosa wishes she had turned to her peers earlier for help and to share her anxieties about starting graduate school in a new environment. "Ask questions of those who have gone through the process. They can be our guides."

"We also need faculty and staff to be proactive," says Díaz Vázquez. He says faculty members should not assume that a quiet first-generation student is not engaged or not struggling. Graduate advisers should check in with students and ask what hurdles they're facing, so they can provide support or recommend other mentors. "I don't know how to navigate the health-care system, but I can put students in contact with someone who does. That's why, as a mentor, I have to keep building my network," says Díaz Vázquez.

Whitley encourages institutions to be more transparent about graduate-school expectations and requirements. "If we do everything with first-gen

students in mind, everyone benefits. If you make institutional navigation easier, if you make policies and procedures more clear, if you make jargon less complicated, every graduate student is going to benefit,” she says.

Coping with pressure

Perhaps some of the biggest struggles that first-generation graduate students face are emotional and mental: dealing with fear of failure or guilt for moving away from their families or not contributing enough money to them, and an intense pressure to succeed. For example, first-generation students and those with family members who’ve completed university experience similar degrees of imposter syndrome, but it is more strongly associated with stress among those who are first generation⁴.

“The anxiety of being first-gen is just a lot,” says Nicole Schroeder, a postdoctoral researcher in history at Kean University in Union, New Jersey. Throughout university, Schroeder felt hesitant to speak up in class if she didn’t have specific questions or comments. “I felt like I was underprepared, even if that wasn’t necessarily the case.”

To build her confidence, Schroeder wrote out prepared answers to discussion questions before class. “Even when I made comments that were off the mark, my educators would often gently correct me or use it as an opportunity to clarify,” she says. In time, she learnt that nothing bad was going to happen, and that bringing up topics that she was confused about probably helped other students, too.

Now, in her teaching-intensive postdoc position, Schroeder supports first-generation undergraduates in the classroom. “I can tell they struggle with some of the same things, such as whether their perspective matters,” she says. To ease the anxieties of these students, Schroeder provides multiple avenues for students to demonstrate their participation: she evaluates their notes on course readings, and breaks the class into smaller discussion groups so it’s easier for everyone to speak up.

Many students experience another burden — the guilt of pursuing a different path. Díaz Vázquez struggled to support his parents financially during his

graduate studies. “I felt pressure to help because I was making more money than my parents — but, at the same time, I was not making enough to survive myself,” he says.

Maphosa left her three-year-old son in Zimbabwe to obtain her graduate degree in South Africa. Although her parents support her decision, she says they struggle to understand her graduate-school obligations. “It was as though I had abandoned my duties. I was anxious, ashamed, and felt like an imposter. It felt like I had to choose between being a mother or a student,” she says.

In time, Maphosa realized that this perception wasn’t accurate and that she was serving as a role model for her son. “Put yourself first sometimes. Your children will be proud of you,” she says. “If you fail, that’s also OK, pick yourself up and try again.”

Guilt is often coupled with an intense drive to succeed. “My family has sacrificed a lot for me to study at the graduate level — I’m away for long periods of time and they have to make do with my limited income,” says Ngulube. “I feel a drive to perform because not performing will be like wasting precious time.”

Ngulube’s family is supportive because they see his graduate experience as an example for the next generation. “I call it positive pressure,” he says. If he pushes the boundaries in his career, then his younger family members will broaden their own career aspirations.

Although the feelings of guilt and pressure to succeed might not go away, first-generation students can cope by building a community and support network that understands them. “Finding your people is the most important thing,” says Schroeder, noting that this group doesn’t have to be limited to other first-generation students.

When Schroeder began her postdoctoral position, she was assigned a faculty mentor. But it wasn’t a perfect match — Schroeder has Ehlers–Danlos syndrome, an incurable connective-tissue disorder, and had brain surgery before starting her PhD programme. “My mentor is not disabled, is older and is not a woman. I am dealing with problems in the academy that my mentor

cannot help with,” she says. Instead, she has relied on her peers for support. “When I didn’t find a community around me, I built it,” says Schroeder, who founded the Disabled Academic Collective, a blog and support network.

“Networking can be incredibly intimidating,” she says. At conferences, Schroeder seeks out other early-career researchers who are in PhD, postdoc and tenure-track positions. “They are going to be your best bet because they are the most invested in bringing new light into the academy.”

She openly identifies as first-generation and disabled but notes that these topics are rarely, if ever, addressed in formal conference and networking events. Instead, she focuses on networking with peers online through her website, Twitter and Zoom. She also meets with two to four early-career scholars from her PhD programme over Zoom every other day or so, to work together, share resources and give presentations as practise for job interviews. “That has been my anchor,” says Schroeder.

Some departments and institutions have first-generation and identity-based organizations, as well as mental-health, physical-health and spiritual resources, that students can tap into as soon as they start graduate programmes. Connecting with other first-generation students early in graduate studies and openly sharing experiences is immensely important, says Díaz Vázquez. “We have to share the challenges because the challenges that you have as a first-gen undergraduate just escalate in intensity as grad students and then as a faculty. But by sharing and being vulnerable, we can empower each other.”

Nature **604**, 585–587 (2022)

doi: <https://doi.org/10.1038/d41586-022-01048-9>

References

1. Christophers, B. *et al. Nature Med.* **27**, 752–755 (2021).
2. Roksa, J., Feldon, D. F. & Maher, M. *J. Higher Educ.* **89**, 728–752 (2018).

3. Henderson, M., Shure, N. & Adamecz-Völgyi, A. *Oxford Rev. Educ.* **46**, 734–751 (2020).
 4. Holden, C. L., Wright, L. E., Herring, A. M. & Sims, P. L. *J. Coll. Stud. Retent. Res. Theory Pract.* <https://doi.org/10.1177/15210251211019379> (2021).
-

This article was downloaded by **calibre** from <https://www.nature.com/articles/d41586-022-01048-9>

| [Section menu](#) | [Main menu](#) |

- WHERE I WORK
- 19 April 2022

Ultracold storage ensures a future for endangered plants

Plant scientist Liang Lin develops cryopreservation technology to store seeds and plant tissues at liquid-nitrogen temperature to conserve biodiversity.

- [Yvaine Ye](#) 0



Liang Lin is a plant scientist at the Kunming Institute of Botany, Yunnan, China. Credit: Li Lianyi

Here at the Germplasm Bank of Wild Species of China at the Kunming Institute of Botany, we want to preserve the seeds of as many wild plants as possible from across China's vast land area. I work on developing the best techniques to freeze plant seeds and tissues at ultracold temperatures, to maintain their viability for years. The idea is that if we plant these seeds again in hundreds of years, a plant will grow.

The picture shows me taking a sample of embryos from the seeds of a magnolia tree out of a liquid-nitrogen cryopreservation tank to test whether they'll regrow when thawed. I dress in protective equipment from head to toe to protect me from the nitrogen, which has a temperature of -196°C .

When I came to the institute in 2009 as a PhD student, it had just purchased its first liquid-nitrogen cryopreservation system, but no one knew how to operate it. I was the one to work it out.

Over the years, human activities and climate change have had a negative impact on plant biodiversity. The ultimate goal of the plant seed bank is to collect and preserve all wild plant species in China that are endangered, rare or valuable. We want to save these species before they go extinct. We've collected seeds from nearly 11,000 plant species, but that's only one-third of what grows in China.

Many wild plants have genes that help them to survive in harsh environments and make them disease- or drought-resistant. In the future, we might need these genetic materials to breed new crops that can better adapt to the changing climate.

I am constantly amazed by how diverse and beautiful seeds are. Some of them are brightly coloured and others are star-shaped. I feel proud when I see the unfrozen seeds germinate in test tubes and gradually grow into large plants. We have three plants in the seed-bank lobby that we cultivated from preserved tissues, and they are all now taller than me.

Nature **604**, 590 (2022)

doi: <https://doi.org/10.1038/d41586-022-01049-8>

This article was downloaded by **calibre** from <https://www.nature.com/articles/d41586-022-01049-8>

| [Section menu](#) | [Main menu](#) |

Research

- **[The Human Pan-genome Project: a global resource to map genomic diversity](#)** [20 April 2022]

Perspective • The Human Pan-genome Reference Consortium aims to offer the highest quality and most complete human pan-genome reference that provides diverse genomic representation across human populations.

- **[A quantum processor based on coherent transport of entangled atom arrays](#)** [20 April 2022]

Article • A quantum processor is realized using arrays of neutral atoms that are transported in a parallel manner by optical tweezers during computations, and used for quantum error correction and simulations.

- **[Global seasonal forecasts of marine heatwaves](#)** [20 April 2022]

Article • Climate forecast systems are used to develop and evaluate global predictions of marine heatwaves (MHWs), highlighting the feasibility of predicting MHWs and providing a foundation for operational MHW forecasts to support climate adaptation and resilience.

- **[Climate effects on archaic human habitats and species successions](#)** [13 April 2022]

Article • A new model simulation of climate change during the past 2 million years indicates that the appearances and disappearances of hominin species correlate with long-term climatic anomalies.

- **[Somatic mutation rates scale with lifespan across mammals](#)** [13 April 2022]

Article • Whole-genome sequencing is used to analyse the landscape of somatic mutation in intestinal crypts from 16 mammalian species, revealing that rates of somatic mutation inversely scale with the lifespan of the animal across species.

- **[Brain charts for the human lifespan](#)** [06 April 2022]

Article • MRI data from more than 100 studies have been aggregated to yield new insights about brain development and ageing, and create an interactive open resource for comparison of brain structures throughout the human lifespan, including those associated with neurological and psychiatric disorders.

- **[Molecular basis of receptor binding and antibody neutralization of Omicron](#)** [28 February 2022]

Article • The structures of the open and closed states of the Omicron spike protein and its complex with the ACE2 receptor or a broadly neutralizing antibody are resolved and shed light on the receptor engagement and antibody neutralization of Omicron.

- **Antibody evasion properties of SARS-CoV-2 Omicron sublineages** [03 March 2022]

Article • A study reports on the antigenic characterization of SARS-CoV-2 BA.1, BA.1.1 and BA.2 and the neutralizing activity of different monoclonal antibodies and sera against them.

- **Nonlinear control of transcription through enhancer–promoter interactions** [13 April 2022]

Article • The transcriptional effect of an enhancer depends on its contact probabilities with the promoter through a nonlinear relationship, and enhancer strength determines absolute transcription levels as well as the sensitivity of a promoter to CTCF-mediated transcriptional insulation.

[Download PDF](#)

- Perspective
- [Published: 20 April 2022](#)

The Human Pangenome Project: a global resource to map genomic diversity

- [Ting Wang](#) ORCID: [orcid.org/0000-0002-6800-242X^{1,2,3}](https://orcid.org/0000-0002-6800-242X),
- [Lucinda Antonacci-Fulton³](#),
- [Kerstin Howe](#) ORCID: [orcid.org/0000-0003-2237-513X⁴](https://orcid.org/0000-0003-2237-513X),
- [Heather A. Lawson¹](#),
- [Julian K. Lucas⁵](#),
- [Adam M. Phillippy](#) ORCID: [orcid.org/0000-0003-2983-8934⁶](https://orcid.org/0000-0003-2983-8934),
- [Alice B. Popejoy⁷](#),
- [Mobin Asri⁵](#),
- [Caryn Carson](#) ORCID: [orcid.org/0000-0003-2091-3109^{1,2,3}](https://orcid.org/0000-0003-2091-3109),
- [Mark J. P. Chaisson](#) ORCID: [orcid.org/0000-0001-5395-1457⁸](https://orcid.org/0000-0001-5395-1457),
- [Xian Chang⁵](#),
- [Robert Cook-Deegan⁹](#),
- [Adam L. Felsenfeld¹⁰](#),
- [Robert S. Fulton³](#),
- [Erik P. Garrison](#) ORCID: [orcid.org/0000-0003-3821-631X¹¹](https://orcid.org/0000-0003-3821-631X),
- [Nanibaa' A. Garrison](#) ORCID: [orcid.org/0000-0002-6228-3216^{12,13,14}](https://orcid.org/0000-0002-6228-3216),
- [Tina A. Graves-Lindsay³](#),
- [Hanlee Ji](#) ORCID: [orcid.org/0000-0003-3772-3424¹⁵](https://orcid.org/0000-0003-3772-3424),
- [Eimear E. Kenny^{16,17,18}](#),
- [Barbara A. Koenig¹⁹](#),
- [Daofeng Li](#) ORCID: [orcid.org/0000-0001-7492-3703^{1,2,3}](https://orcid.org/0000-0001-7492-3703),
- [Tobias Marschall](#) ORCID: [orcid.org/0000-0002-9376-1030²⁰](https://orcid.org/0000-0002-9376-1030),

- [Joshua F. McMichael](#) [ORCID: orcid.org/0000-0002-0197-2245³](#),
- [Adam M. Novak⁵](#),
- [Deepak Purushotham^{1,2,3}](#),
- [Valerie A. Schneider²¹](#),
- [Baergen I. Schultz¹⁰](#),
- [Michael W. Smith¹⁰](#),
- [Heidi J. Sofia¹⁰](#),
- [Tsachy Weissman²²](#),
- [Paul Flicek](#) [ORCID: orcid.org/0000-0002-3897-7955²³](#),
- [Heng Li](#) [ORCID: orcid.org/0000-0003-4874-2874^{24,25}](#),
- [Karen H. Miga](#) [ORCID: orcid.org/0000-0002-3670-4507⁵](#),
- [Benedict Paten](#) [ORCID: orcid.org/0000-0001-8863-3539⁵](#),
- [Erich D. Jarvis](#) [ORCID: orcid.org/0000-0001-8931-5049^{26,27}](#),
- [Ira M. Hall](#) [ORCID: orcid.org/0000-0003-4442-6655²⁸](#),
- [Evan E. Eichler](#) [ORCID: orcid.org/0000-0002-8246-4014^{29,30}](#),
- [David Haussler](#) [ORCID: orcid.org/0000-0003-1533-4575^{5,31}](#) &
- [the Human PanGenome Reference Consortium](#)

Nature volume **604**, pages 437–446 (2022)

- 13k Accesses
- 268 Altmetric
- [Metrics details](#)

Subjects

- [Ethics](#)
- [Genetics research](#)
- [Genome](#)
- [Genome informatics](#)
- [Genomic analysis](#)

Abstract

The human reference genome is the most widely used resource in human genetics and is due for a major update. Its current structure is a linear composite of merged haplotypes from more than 20 people, with a single individual comprising most of the sequence. It contains biases and errors within a framework that does not represent global human genomic variation. A high-quality reference with global representation of common variants, including single-nucleotide variants, structural variants and functional elements, is needed. The Human Pangenome Reference Consortium aims to create a more sophisticated and complete human reference genome with a graph-based, telomere-to-telomere representation of global genomic diversity. Here we leverage innovations in technology, study design and global partnerships with the goal of constructing the highest-possible quality human pangenome reference. Our goal is to improve data representation and streamline analyses to enable routine assembly of complete diploid genomes. With attention to ethical frameworks, the human pangenome reference will contain a more accurate and diverse representation of global genomic variation, improve gene–disease association studies across populations, expand the scope of genomics research to the most repetitive and polymorphic regions of the genome, and serve as the ultimate genetic resource for future biomedical research and precision medicine.

Main

The human reference genome is the foundational open-access resource of modern human genetics and genomics, providing a centralized coordinate system for reporting and comparing results across studies^{1,2,3,4}. Its release set the bar for genomic data sharing, essential for nearly all human genomics applications, including alignments, variant detection and interpretation, functional annotations, population genetics and epigenomic analyses. The current human reference (GRCh38.p13) is a mosaic of genomic data assembled from more than 20 individuals, with approximately 70% of the sequence contributed by a single individual^{5,6,7}. Dependence on a single mosaic assembly (which does not represent the sequence of any one person) creates reference biases, adversely affecting variant discovery, gene–disease association studies and the accuracy of genetic analyses^{8,9}. More than two decades after the first human genome reference sequences were released, the current reference genome still contains errors, rare structural configurations

that do not exist in most human genomes, and gaps in regions that have been difficult to assemble^{7,10} because of their repetitive and highly polymorphic nature. The human reference genome, like most technology-driven resources, is overdue for an upgrade¹¹.

For years, the Genome Reference Consortium has updated the linear reference by fixing errors, filling in gaps and adding newly discovered variants^{1,4,7,12}. When enough changes accumulate, new builds are generated and released. Although this process has served the community well, shortcomings have been identified along the way. Segments of genome sequences sampled from individuals may differ considerably from the reference genome, leading to errors in read mapping to the reference and reducing the accuracy of variant calls^{13,14}. Identification of structural variants (more than 50-bp deletions, insertions, tandem duplications, inversions and translocations) relies on detecting patterns of discordant read pairs or split read alignments, which in turn depend on the accuracy of read mapping^{15,16}. Assembling and detecting these structural variants are challenging when the reads are too short to cover long, repetitive regions of the genome⁸. This is because short reads (50–300 bp) from different repeats may be identical and/or overlapping with one another such that it is impossible to determine where they should map. Both the limitations of short reads and reference biases mean that we may have missed more than 70% of structural variants in traditional whole-genome sequencing studies^{17,18}.

Advances in sequencing technologies and a greater appreciation for the importance of genetic diversity make improving the human reference sequence both timely and practical. First, the development of long-read (more than 10 kb) sequencing technologies has enabled the assembly of large, repeat-rich regions, facilitated phasing and assembly of maternal and paternal haplotypes, and improved representation of GC-rich regions of the genome that are often missing in short-read assemblies^{8,19,20,21,22}. Second, growing recognition of the importance of diversity and inclusion in human genomics²³ has led to widespread calls to improve representation and methods for detecting and presenting global variation.

In this Perspective article, we outline the goals, strategies, challenges and opportunities for the Human Pangenome Reference Consortium (HPRC). We will engage scientists and bioethicists in creating a human pangenome reference and resource that represents genomic diversity across human populations, as well as improving technology for assembly and developing an ecosystem of tools for analyses of graph-based genome sequences. This new reference will maintain essential ties to the original reference for continuity, even as we strive to develop complete and error-free telomere-to-telomere (T2T) assemblies of all chromosomes of individual human genomes, referred to here as ‘haplotypes’.

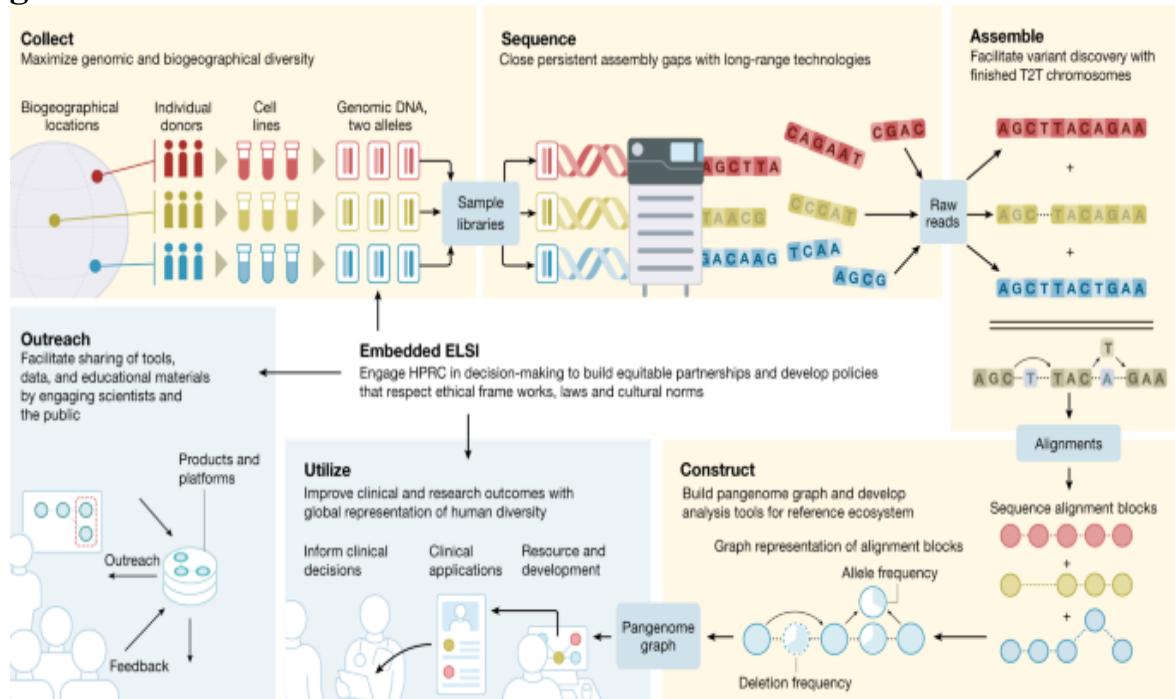
Goals and strategies of the HPRC

A ‘pangenome’ is the collective whole-genome sequences of multiple individuals representing the genetic diversity of the species. Originally popularized in the context of highly dynamic bacterial genomes²⁴, the concept has been adapted to the field of human genomics, in which the full extent of human genomic variation is expected to be much broader than has thus far been revealed. The pangenome data infrastructure depends on the high-throughput production of high-quality, phased haplotypes (segments of a chromosome identified as being maternally or paternally inherited) that improve upon the current human reference genome. Highly accurate and complete haplotype-phased genome assemblies will be organized into a graph-based data structure for the pangenome reference that compresses and indexes information^{25,26,27}. This data structure will contain a coordinate system with a simple, intuitive framework for referring to genomic variants, as well as preserving backward compatibility with GRCh38 and previous linear reference builds. Managing and interpreting these data require transdisciplinary collaboration and innovation, focused on the development of novel conceptual frameworks and analytic methods to construct the pangenome infrastructure and tools for downstream analyses and visualization. The goals of the HPRC are laid out in Box 1.

The HPRC functions through multidisciplinary collaborations, convening cross-institutional and multinational working groups dedicated to sample collection and consent, population genetic diversity, technology and production, phasing and assembly, approaches to construction of a

pangenome reference, resource improvement and maintenance, and resource sharing and outreach (Fig. 1). The HPRC has begun the process of engaging international partnerships with the Australian National Centre for Indigenous Genomics (NCIG; <https://ncig.anu.edu.au>), the US Food and Drug Administration (FDA)-recognized Clinical Genome Resource (ClinGen)²⁸, the National Institutes of Health (NIH)-funded Human Heredity and Health in Africa (H3Africa; <https://h3africa.org>) Consortium, the Personal Genome Project (PGP; <https://www.personalgenomes.org>), the Vertebrate Genomes Project⁸ and the Global Alliance for Genomics and Health (GA4GH; <https://www.ga4gh.org>). The HPRC will integrate perspectives from the international scientific community through these collaborators and others yet to be identified to inform the development of HPRC references, methods and standards.

Fig. 1: The HPRC.



An overview of several components of the HPRC. Collect: 1,000 Genomes samples start the project and will be followed by additional samples collected through community engagement and recruitment. Sample selection efforts will ensure that the graph-based reference captures global human genomic diversity. Sequence: long-read and long-range technologies are used to generate genome graphs and bridge gaps in difficult-to-assemble

genomic regions. Assemble: T2T finished diploid genomes will foster variant discovery, especially in complex, difficult-to-assemble genomic regions. Construct: scalable bioinformatics approaches assemble, quality control, call variants and benchmark graph assembly accuracy. The graph is annotated with gene descriptions and transcriptome data, making it more accessible and interpretable. Utilize: collaboration across scientific and stakeholder communities will create a new ecosystem of analysis tools. Clinical applications and research use will involve analysis, validation, interpretation and publication of results. Outreach: members of the HPRC outreach community engage and educate the user community and broadly share all genomic products and informatics platforms. ELSI: ELSI scholars will develop selection processes and policy frameworks that meet investigator needs as well as respect research partner autonomy and cultural norms.

Box 1 Goals of the HPRC

- Identify individuals from diverse genomic and biogeographical backgrounds to include in the pangenome reference, with at least 350 reference quality haplotype-phased human diploid genomes (700 haplotypes in total).
- Integrate ethical, legal and social implications (ELSI) scholarship in the development of recommended policies and protocols for inclusion, data acquisition and stewardship from study recruitment to publication of findings.
- Prioritize the use of long-read and long-range technologies for assemblies, with haplotype-aware algorithms to generate the highest quality phased genomes possible.
- Create methods to finish diploid genomes from T2T across complex regions, closing gaps and ensuring hard-to-measure variants are identified.
- Foster an ecosystem of pangenome reference tools to facilitate the annotation of genes and other genomic features.

- Implement an iterative design–development–engagement process to understand and respond to user community needs.
- Develop communication strategies that will assure understanding of the pangenome reference resource, including the ability of the community to fix and report errors.
- Enable appropriately controlled access to data through genomics platforms such as the INSDC^{59,71}, the NCBI, the UCSC Genome Browser^{72,73}, Ensembl^{74,75}, the WashU Epigenome Browser^{76,77} and NHGRI’s cloud-based analysis platform AnVIL^{46,78}.
- Foster an international human pangenome reference alliance that actively engages the diverse populations it seeks to represent.

Inclusion criteria

For initial inclusion, the HPRC selected individual genomes for high-quality sequencing among existing cell lines established by the 1000 Genomes Project (1KGP), which offers a deep catalogue of human variation from 26 populations²⁹. These cells were originally collected from volunteer donors using consent procedures designed for unrestricted data use, and the cell lines are available in the National Human Genome Research Institute (NHGRI) Biorepository at Coriell (<https://www.coriell.org>). The selected cell lines were prioritized on the basis of a combination of criteria, ranging from genetic and geographical diversity of the donors, to the availability of relevant parental data (for haplotype phasing), and limited time in cell culture (to minimize the accumulation of de novo mutations).

Differences between individuals were initially identified using clustering and visualization techniques (uniform manifold approximation and projection clusters generated from 1KGP data) and observed allelic diversity (heterozygosity), and then selected for inclusion in the first phase of the HPRC ($n = 100$). Our inclusion criteria and recruitment strategies are evolving with the project, and we recognize that there are inherent limitations to clustering algorithms and using only the 1KGP dataset.

Although useful for the first phase of the HPRC, genomes selected from the 1KGP data represent a limited scope of geographical and genomic diversity. One reason is that the resource was developed by sampling in 26 geographical locations across the globe, and the discrete number of individuals included from each location limits the amount of genomic variation representing those regions, especially regarding rare variants that are less likely to be observed in small sample sizes. The genomes of individuals sampled from each 1KGP location cannot be assumed to have sufficient variation to be comprehensive of the genomic diversity in the natural population of the region, let alone to represent an entire continent. Furthermore, 1KGP populations were often selected by asking potential study participants questions about their racial, ethnic or ancestral identities, assigning ancestry on the basis of geographical location, or some combination, which would not necessarily produce a representative sampling of any natural population. As population descriptors can be inconsistent on clinical forms³⁰ and are fluid across cultural contexts^{31,32,33}, there are many unknown layers of diversity within each geographical sampling cohort of the 1KGP data.

Because the 1KGP data are insufficient to support the ambitious sampling and genetic diversity goals of the HPRC, the consortium will include additional genomes, including from participants identified through the BioMe Biobank at Mount Sinai and a cohort of African American individuals recruited by Washington University. All participants will give informed consent, and their sequence data will be deidentified in open access. Some will have cell lines generated at Coriell. In later phases, the HPRC will foster additional domestic and international partnerships to explore additional avenues to broaden diversity and enhance inclusion (Box 2).

Box 2 Commitment to diversity and inclusion

There are many aspects of diversity to consider for broad inclusion, and the first step is to assess current gaps in diversity. Researchers have demonstrated a lack of diversity in genomics research using biogeographical ancestry groupings at the continental level, as well as sociocultural categories such as racial and ethnic identities⁷⁹. It is important to distinguish

biological and sociocultural diversity, as sociocultural labels are not derived from genotype data, and vice versa. Owing to gaps in genomic sampling worldwide, the distribution of allelic variants appears to be correlated with continental-level biogeographical ‘ancestry’. However, variants are rarely unique to a single biogeographical ‘population’, and factors such as effective population size, founder events and genetic drift are responsible for differences in allele frequencies between such groups.

Capturing the full range of human genomic diversity is a daunting task: some gaps are understood and predictable, but we also face ‘unknown’ unknowns. The initial HPRC dataset cannot be comprehensive of global genomic variation, but it can set a foundation to build on. The HPRC will initially produce high-quality genome data for 350 individuals (700 haploid genomes) selected to maximize global representation within the logistical constraints of the initial HPRC efforts. Strategic partnerships with organizations such as the GA4GH and H3Africa are underway, which we anticipate will help to facilitate international engagement and broaden our understanding of the cultural, ethical, legal, social and political considerations of the HPRC. However, further partnerships will be needed to include populations that are underrepresented or entirely missing from current data resources. The HPRC actively welcomes additional partners and collaborators to join us in rising to this challenge.

Embedded ELSI scholarship

Most human genomics has been based on individuals of European ancestry, and the datasets available for analyses are thus biased. As a result, current precision medicine is based on genomic variation found in populations with primarily European ancestry. Much of the global genetic diversity that contributes to clinical phenotypes is missing from clinical genetic tests. Many ethical, legal and social challenges arise in efforts to include previously excluded populations, communities or groups.

The HPRC has formed an ‘embedded’ team of scholars to address ethical, legal and social implications (ELSI) of its work, with expertise at the intersections of genomics with biomedical ethics, law, social sciences, demography, community engagement and population genetics. The main

objective of the HPRC-ELSI team is to identify, investigate and ultimately offer consortium investigators advice about the issues they face, which must be addressed if the HPRC is to meet its goals. In the embedded model, with ELSI scholars participating in key meetings during which decisions are made, investigators can engage these colleagues in discussions that deepen their understanding and appreciation of what is at stake as we seek to improve the human reference genome.

Large-scale human population genetics projects aimed at broadening the diversity in genomic datasets and analyses have often missed the mark in demonstrating respect for individuals and communities. The Human Genome Diversity Project encountered strong opposition three decades ago^{[34](#)}, facing objections that its approach was extractive and its goals benefitted scientists and institutions in rich countries, but did not match the priorities of Indigenous peoples or people in resource-poor regions who were asked to donate their samples and data. The Havasupai Tribe of northern Arizona sued the Arizona Board of Regents in 2002 when they learned that samples donated for diabetes research were shared with other researchers and re-used for studies of schizophrenia and population origins, to which tribal members did not agree. That case was settled in 2010 (refs. ^{[35](#)},^{[36](#)}), but the effect it had on relations between tribal communities and genomics research has persisted.

More recently, the Wellcome Sanger Institute was criticized for licensing access to data arising from southern African samples, despite institutions in Africa asserting terms of informed consent that did not permit commercial uses. The NIH was also criticized for inadequate tribal engagement and consultation in the All of Us programme^{[37](#)},^{[38](#)}. With a keen awareness of this history, the HPRC has initiated a process to consult with, engage and genuinely include groups who are currently not well represented in the genomic database. Indigenous scholars have spearheaded a movement for Indigenous data sovereignty^{[39](#)}, for example, including the development of the CARE (collective benefit, authority to control, responsibility, and ethics) principles for Indigenous data governance^{[40](#)} to layer onto the FAIR (findable, accessible, interoperable, reusable) principles that support the open-science approach that the HPRC and similar projects take^{[41](#)}. The HPRC is reaching out to Indigenous geneticists, leaders and community

members to engage and collaboratively develop a truly global and inclusive reference resource, taking into account FAIR and CARE principles.

Furthermore, similar efforts will be made for other diverse populations that the HPRC will work with.

Some groups who we seek to engage with may develop sampling and sequencing efforts parallel to, rather than directly participating in, the HPRC. In developing a state-of-the-art pangenome reference sequence, the HPRC will continue to disseminate standards for accuracy and completeness of sequencing, as well as emphasizing the importance of ELSI considerations. It is a priority for the HPRC to actively communicate with parallel genome-wide sequencing efforts to ensure compatibility between efforts, thus enabling integration into a global pangenome reference resource. The HPRC is committed to assessing local policies and promoting broad sharing of resources developed through interdisciplinary engagement, scholarship and innovative technical solutions. The HPRC will establish procedures for navigating potential tensions among its technical, research and resource-generating objectives with local customs, laws and data sharing policies for the groups within the HPRC as well as for those in the parallel projects.

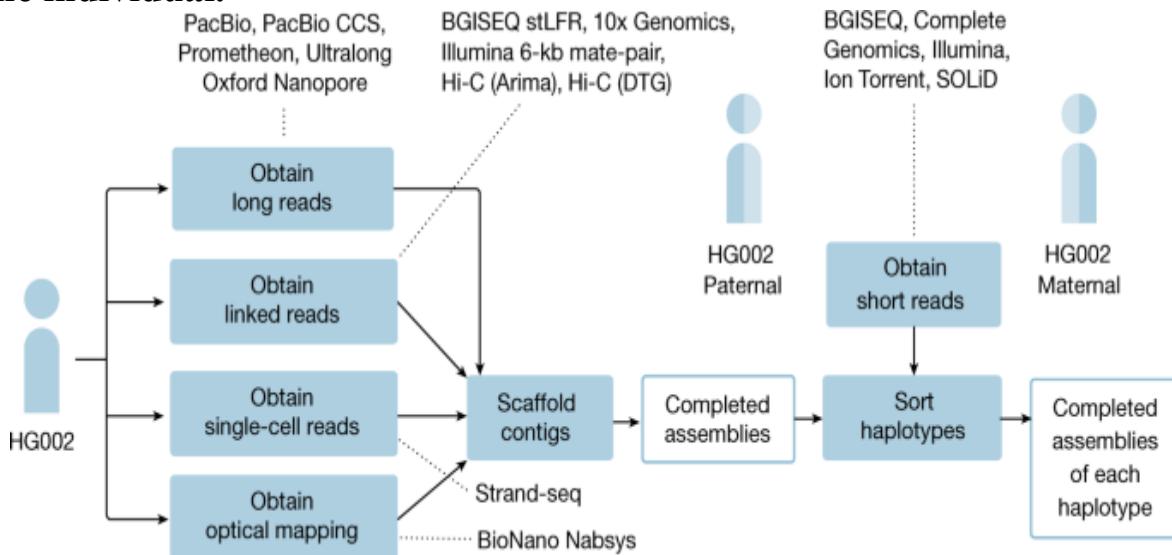
Initial data generation and release

Technological advances in genomics enable sequencing long repeats, physical mapping to chromosomes, and phasing maternally and paternally inherited haplotypes (Box 3).

For the initial phase of the project, we sequenced a single individual, HG002, whose genomic sequence has been thoroughly characterized by the Genome in a Bottle (GIAB) Consortium⁴². We evaluated multiple sequencing technologies and assembly algorithms to identify the optimal combination of platforms and develop an automated pipeline that generated the most complete and accurate genome representation⁴³ (Fig. 2). We began with the now well-established assumption that long reads (more than 10 kb) yield more complete genome assemblies than short reads alone⁸. The technologies tested included Pacific Biosciences (PacBio) and/or ONT long reads for generating contigs, 10x Genomics linked reads, Hi-C paired reads,

Strand-seq long reads, and/or BioNano optical maps for scaffolding contigs into chromosomes. This pilot benchmark study created the standards for sequencing technologies and computational methodologies that are critical to the success of the HPRC.

Fig. 2: Standards were developed through a pilot benchmark study of one individual.



Multiple long-read and long-range technologies and computational methods were evaluated to develop the combination of platforms and an automated pipeline that provides the most complete and accurate genome graph. CCS, circular consensus sequencing; stLFR, single-tube long fragment read.

We found that the trio approaches using parental short-read sequence data to sort haplotypes of the long-read data of offspring gave the most complete assemblies of each haplotype with the fewest structural errors⁴³. Furthermore, all methods attempting to separate haplotype sequences performed much better in generating highly contiguous assemblies than those that merged the consensus between haplotypes into one assembly. The algorithm that gave the highest haplotype separation accuracy for contigs was HiFiasm⁴⁴, which incorporates separation of reads of each haplotype into the assembly graph⁴⁵. Generation of contigs were more structurally accurate than scaffolds, where the HPRC identified areas of improvements that were necessary to prevent contig miss-joins, missed-joins, collapsed repeats and other structural assembly errors. On the basis of these findings, an initial set of 47 1KGP genomes from parent–offspring trios was

assembled with HiFiasm^{43,44}, creating high-quality diploid contig-only genome assemblies. Going forward, we will further optimize sequencing, assembly and analysis methods with the goal of creating fully-phased T2T diploid genomes, including repetitive and structurally variable regions such as centromeres, telomeres and segmental duplications. We anticipate that the high-quality assemblies created in the project will drive tool creation and improvement for diploid genome assembly and quality control in which new and recently created existing tools (from the T2T assembly of CHM13 (ref. 22)) are applied to diploid genome assembly.

The first HPRC data release comprises the sequencing data from 47 participants, mostly from the 1KGP (listed and described in Supplementary Table 1). All sequencing data are publicly available and can be downloaded without egress fees from the Amazon Web Services (AWS) Public Datasets program and can be analysed with the AWS cloud. Data are also available for analysis within the AnVIL (Analysis, Visualization and Informatics Lab-space) cloud platform, organized as a public workspace (https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_HPRC). AnVIL is the genomic data science analysis, visualization, and informatics lab-space of the NHGRI that provides a cloud environment for analysis of large genomic datasets, and supports multiple globally used analysis tools including Terra, Bioconductor, Jupyter and Galaxy⁴⁶.

Box 3 Sequencing and assembly

Notable improvements in long-read technologies have resulted in complete chromosome assemblies^{20,21,22} and have demonstrated the ability to broaden variant analysis to span large, complex human structural variants⁵¹. Use of highly accurate consensus reads (99.9%, or Q30) of moderate length (for example, 10–20 kb), such as high fidelity (HiFi) reads from PacBio, routinely resolves long tandem repeats, or satellite arrays, and large segmental duplications^{20,21,63}. In parallel, the nanopore-based sequencing platform (Oxford Nanopore Technologies) offers long-read data that routinely generate substantial coverage of reads that are hundreds of kilobytes in length (or ‘ultra-long’ data) with an increasing number of reported reads greater than 1 million bases. Like the HiFi data, ultra-long data is used to close large and persistent assembly gaps, including in human

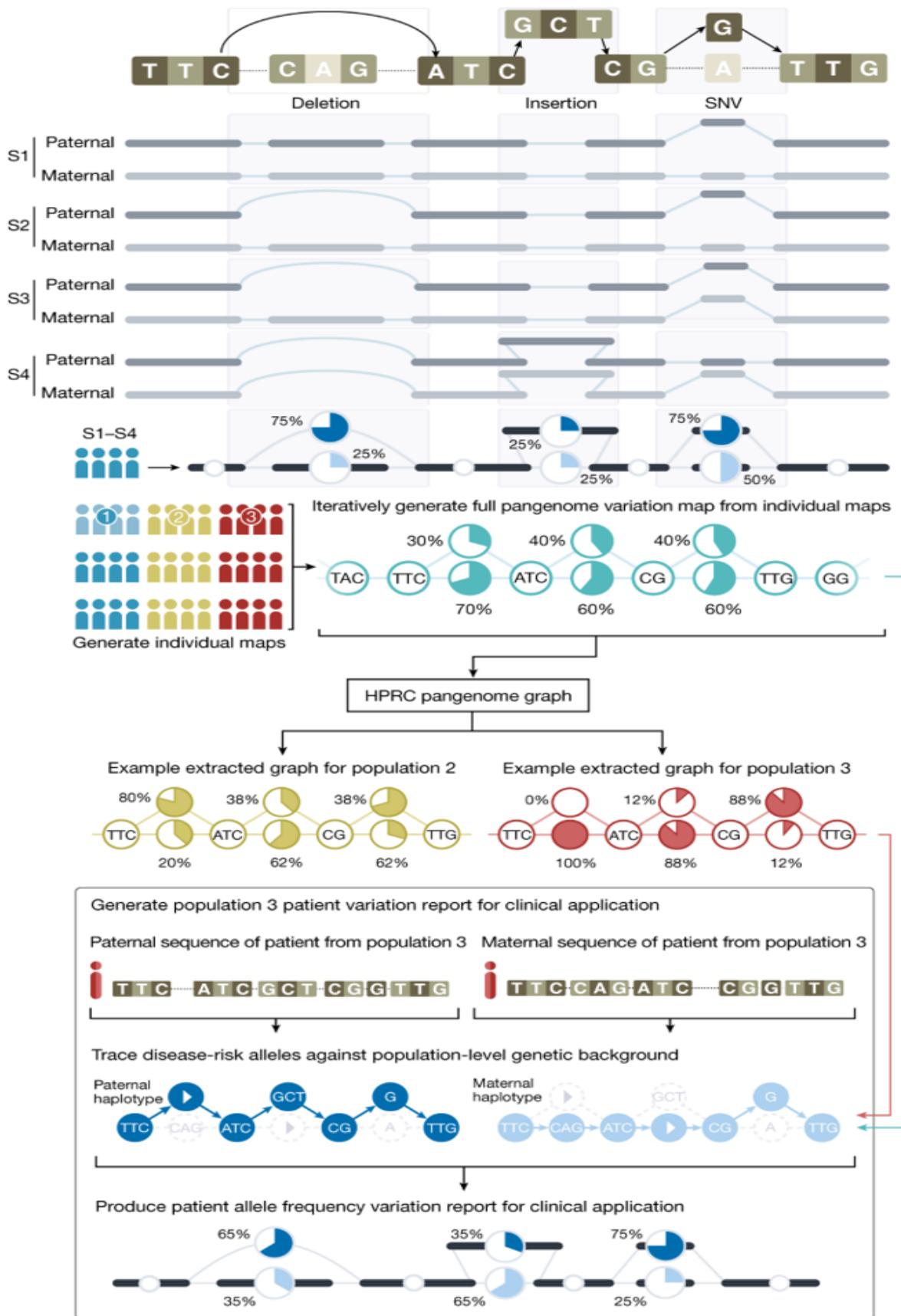
centromeres, subtelomeric regions and large segmental duplications^{22,63}. Furthermore, chromosome conformation capture methods produce long-range data for both short-read (Hi-C⁸⁰) and long-read (Pore-C⁸¹) sequencing. Such chromatin crosslinking protocols generate chimeric DNA fragments from interacting chromosomal regions that are covalently linked together. These ligated DNA molecules are sequenced to help determine phasing and spatial organization at the level of an entire chromosome. With continuing gains in both HiFi read length and nanopore single-read base-level quality, and improved methods for the use of chromosome conformation capture methods to guide phased haplotype assembly, we are entering into a new era of routine complete chromosome-level assemblies^{20,21,22,78}. In collaboration with the T2T Consortium, which aims to use long-read sequencing and cutting-edge algorithmic approaches to close the hundreds of gaps persisting in the human reference genome and that of other species²², the HPRC will generate accurate assemblies of entire chromosomes. These assemblies will empower us to characterize variations in large, repeat-rich regions that have historically been out of reach for standard genetic analysis and interpretation.

Pangenome reference

We are building a pangenome reference with three complementary parts: (1) the haplotypes, which are the sequences within the input assemblies; (2) the pangenome alignment, which is a sequence graph and an efficient embedding of each of the input haplotypes as paths within this graph; and (3) the coordinate system, which is a backward-compatible coordinate system and set of sequences that make it possible to refer to all variations encoded within the reference equally (Fig. 3). The haplotypes provide hundreds of individual representations of the genome, spanning global diversity. Each haplotype assembly will be useful individually as a reference for studying genomic sequences that are divergent from the current human reference assembly. The pangenome alignment represents the homology relationships among the individual assemblies. This canonical alignment will support coordinate translation (liftOver) between the haplotypes and defines the allelic relationships. It will form the substrate for many emerging pangenome tools and pipelines that will improve important genomic workflows, for example, by making genotyping accuracy less dependent on

ancestry. The coordinate system provides a global, unambiguous means to refer to all the variations within the pangenome. It makes all the variations within the haplotypes first-class objects that can be referred to equally. Ultimately, it will provide a more complete means to refer to variations not contained within the existing linear reference, proving useful for databases and tooling that will build on the pangenome reference.

Fig. 3: The human pangenome reference.



Graph-aware mappers can be used to genotype samples by directly mapping against the graph. This simplified example shows how to create a pangenome graph for four people and calculate the allele frequency of three variants. Iterating through each individual produces the structure of the graph, which improves as new genomes are added. Genomic data are arranged into a sequence variation map based on edges. Alternative haplotypes are depicted as alternate pathways across the graph, with the edges being the primary data-bearing elements. The pangenome reference catalogues genomic variation and allows for population-scale analysis because of its graph structure. Tracing a path through the network and connecting sequences at access edges yield haplotypes for individuals. For clinical interpretation, allele frequencies are reported. SNV, single-nucleotide variant.

Supporting these parts is a new proposed set of file standards⁴⁷, notably, the rGFA format for representing a pangenome and the GAF format for representing read mappings to a pangenome. We hope that these will have an effect on the field similar to how SAM/BAM⁴⁸ and VCF⁴⁹ formats have generated a broad range of interoperable tools that have become widely used and accessible. To start this process, we have developed the vg toolkit⁵⁰ and minigraph⁴⁷, which incorporate downstream tools for graph construction and long-read and short-read mapping and genotyping.

We anticipate releasing an alpha pangenome reference based on existing variant calls and assembled contig genomes. Using the proposed incremental coordinate system, we will subsequently release updated graphs that incorporate the growing numbers of assemblies.

Variant detection

A central aim of this research is to document the genetic similarities and differences among the human genomes included in the pangenome reference. Comprehensive variant detection, however, is still a challenge even when high-quality genome assemblies are available. No single data type or bioinformatic approach yet achieves high performance across all variant classes and genomic regions^{51,52}. Therefore, we are pursuing multiple complementary approaches to variant detection using a

combination of whole-genome multiple assembly alignment, pairwise assembly–assembly alignment and traditional reference-based read alignment.

Ideally, we will accomplish variant detection in a single step that is designed to build pangenome graphs directly from whole-genome, multiple assembly alignments. Genetic variants will be represented naturally as features in the resulting graph because any variant would be captured by the assembly process. This offers a substantial advantage, enabling optimal breakpoint reconstruction via joint analysis of all input genomes. Accurate multiple alignment and graph construction of entire human genomes is extremely challenging, but recent improvements to tools such as minimap2 (ref. [53](#)), minigraph⁴⁷, cactus⁵⁴ and pggb⁵⁵ make this feasible. However, errors in variant calling can still arise from errors in assembly and sequence alignment, especially in repetitive regions of the genome. Given this, and the fact that pangenome graph construction tools have not been thoroughly evaluated at scale with real-world data, we are also pursuing the complementary approaches described below.

An alternative approach to multiple alignments is to map variants from pairwise assembly–assembly alignments. Towards this end, we are using minimap2 and Winnowmap to align each draft assembly to the GRCh38 and T2T-CHM13 references to perform variant detection of single-nucleotide variants, indels and other structural variants. This approach is more straightforward than whole-genome multiple alignment; however, complications can arise from reference genome effects and the need to merge results across many pairwise comparisons. The exact coordinates of complex and repetitive variants may differ due to alignment ambiguity. To alleviate reference effects, we are mapping variants via pairwise alignment of the two haploid genomes of each individual, enabling detection of natural heterozygous variants within sequences that are missing or poorly represented in GRCh38 and CHM13. Methods of pairwise alignment assembly help to control for potential errors from the multiple alignment and graph construction process outlined above; however, they still fail to detect variants that are not captured in the underlying assemblies. We are also running a host of traditional variant callers that rely on the alignment of raw reads to the GRCh38 and CHM13 references to control for potential assembly errors. Although limited by reference genome quality and

alignment accuracy, these traditional tools are able to capture a subset of variants that are not accurately assembled, and they will serve as a cross-check on newer and less mature assembly-based tools.

In summary, we expect the above methods to capture most genetic variants in genomic regions that are accessible to current assembly and alignment methods. We will compare our variant calls to published call sets from the 1KGP (<https://www.internationalgenome.org>), HGSCV (Human Genome Structural Variation Consortium, <https://www.internationalgenome.org/human-genome-structural-variation-consortium>) and GIAB⁴² using samples from these projects that are also included in the HPRC references to assess quality. We will evaluate and validate variant calls using independent data types generated by the HPRC but that are not used for contig assembly—such as ONT, Hi-C, Strand-seq and BioNano data—and assess the read-level support for each variant call based on the alignment of raw data to assemblies and pangenome graphs.

Achieving comprehensive T2T variant detection across the entire genome will require improved methods for genome assembly, multiple alignments and graph construction. The development and application of these methods in subsequent years is a major goal of the HPRC, and will help to extend the impact of pangenomics to the full spectrum of variant classes.

Pangenome annotation

Annotation of the current GRCh38 reference includes genes and genomic features, such as repeats, CpG islands, regulatory regions and chromatin immunoprecipitation–seq peaks, among others. The pangenome reference will have these same utilities and more, including the following.

For genes, the two primarily used gene sets in genomic analysis are the National Center for Biotechnology Information’s (NCBI’s) RefSeq⁵⁶, which exists as independent mRNA definitions, and EMBL-EBI’s Ensembl/Gencode⁵⁷, which is built on the GRCh38 reference. The pangenome reference will support both the RefSeq and the Ensembl/Gencode gene sets. We will map both annotations to each

haplotype. Specifically, we will evaluate the mapping of the core reference set of human transcriptome data to each haplotype and incorporate putative new genes that are not represented in either RefSeq or Ensembl/GENCODE. Mapping these gene sets in conjunction with other transcriptomic datasets will annotate the pangenome graph. Other tools will support spliced alignments and transcript reconstruction on a mature graphical data structure. We will integrate the results of these approaches into the annotation released for each haplotype, accompanied by a description of whether transcripts are identical by both methods or whether changes were identified, including transcripts that are disabled, duplicated or missing on a given haplotype. We will also annotate all transcript haplotypes for their global frequency using Haplosaurus tools⁵⁸. We will initially annotate haplotype-by-haplotype, and will explore methods for direct annotation of the pangenome, such as those currently being developed in the GENCODE Consortium. Direct annotation methods simultaneously cover all relevant haplotypes and result in both an annotated genome graph and haplotype-specific annotations. One of the critical use cases of direct annotation of the pangenome will be large transcriptomic datasets aligned directly to the graphical structure that natively annotate it.

For functional elements and other genome features, a central goal in biology is to understand how sequence variants affect genome function to influence phenotypes. Genome function includes regulatory regions that influence gene expression, enhancers that modulate expression levels, and the three-dimensional interactions that control chromosome structural organization within a cell. We will use the pangenome reference to annotate such functional information using existing RNA sequencing, methylC sequencing and assay for transposase-accessible chromatin with high-throughput sequencing datasets from Roadmap Epigenomics, ENCODE, 4D Nucleome (4DN), Genotype-Tissue Expression (GTEx) and the Center for Common Disease Genomics (CCDG), among others. This will enhance the functional human genetic variation catalogue.

Integrating functional data with the pangenome reference will facilitate the development of toolkits and analysis pipelines that evaluate the effect of genetic variants on complex traits and variation in phenotypes. The HPRC will work with developers to define rules and mechanisms to engage with multimodal ‘big bio-data’ for both data providers and consumers. We will

co-create user-friendly informatics platforms to manage, integrate, visualize and compare highly heterogeneous datasets in the context of the genetic diversity represented in the pangenome. Box 4 lists available resources for working with pangenome graphs. We will also make all haplotype-by-haplotype annotation methods available in AnVIL so that others can run them to create custom annotation tracks on all or a selected subset of assemblies. These platforms will serve as a foundation for significant clinical datasets and global biobank initiatives that will ultimately improve precision medicine and medical breakthroughs. For example, the NHGRI is establishing the Impact of Genomic Variation on Function (IGVF) Consortium, which aims to develop a framework for systematically understanding the effects of genomic variation on genome function. Data generated by the IGVF will include high-resolution identification and annotation of functional elements and cell-type-specific perturbation studies to assess the effect of genomic variants on function. The pangenome will be an important foundation for predicting functional outcomes in these studies.

Box 4 Pangenome graph tools

Graph building

- minigraph⁴⁷
- PGGB⁵⁵

Graph aligners

- deBGA⁸²
- BGREAT⁸³
- BrownieAligner⁸⁴
- GenomeMapper⁸⁵
- HISAT2 (ref. ⁸⁶)
- VG⁸⁷

- GraphAligner⁸⁸
- GRAF (<https://www.internationalgenome.org/human-genome-structural-variation-consortium>)
- PaSGAL⁸⁹
- SPAAligner⁹⁰

Graph indexing

- CHOP⁹¹
- PSI⁹²

Graph visualization

- Bandage⁹³
- GfaViz⁹⁴
- SGTK⁹⁵
- AGB⁹⁶
- Sequence Tube Map⁹⁷
- MoMI-G⁹⁸
- VG view⁸⁷
- VG vis⁸⁷
- ODGI viz⁹⁹

Gene prediction

- Path Racer¹⁰⁰

Variant detection

- PanGenie¹⁰¹
- Cortex with Bubbleparse¹⁰²
- BayesTyper¹⁰³
- Paragraph¹⁰⁴
- GraphTyper2 (ref. [105](#))
- VG⁸⁷

Data sharing

To enhance community access and sharing, we will submit sequence data (PacBio HiFi, ONT and Hi-C, among others), assemblies and pangenomes produced by the consortium to AnVIL⁴⁶ and the International Nucleotide Sequence Database Collaboration (INSDC)⁵⁹. Data will also be stored and made publicly available on both S3 and Google Cloud Storage. This general model supports future efforts to use cloud-based strategies for biological data analysis that spans multiple centres. Users of various clouds worldwide will know that they are using the same datasets. Data coordination within the consortium will leverage the established methods in use and the constant development since the inception of the 1KGP more than a decade ago^{60,61}. These processes will ensure that we rapidly release data in an organized manner, with proper accessioning of archival datasets, and future traceability of analysis objects and primary data items. Data stored in the INSDC will use BioProjects and BioProject umbrella structures similar to the 1KGP and the Vertebrate Genome Project⁸ to ensure that data are appropriately organized and easily identifiable in the public archives. This approach ensures that sample identifiers are effectively managed via the BioSamples database⁶², including metadata provisions, and makes any data generated from the same samples readily tractable. The INSDC will archive all reads and assembly data, and other relevant archives will be used, as appropriate for a specific data type. Each haplotype assembly will receive a genome

collections accession number (GCA_ *), which we will version as we make assembly updates. We will address additional data sharing considerations as they arise through our expanded recruitment and sampling efforts to broaden the diverse representation of global variation.

Adoption and outreach

Achieving widespread international adoption of a pangenome reference will be a challenge¹¹. The HPRC will design a pragmatic model and transition plan that are simple and compelling enough to gain traction among researchers and clinical laboratories. Working across scientific and other stakeholder communities, we will foster a new ecosystem of analysis tools. We will maintain and improve the reference, establish scalable bioinformatics methods for resolving errors, improve resolution in difficult-to-resolve genomic regions and respond to user feedback. Importantly, we envision an integrated pangenome transition plan that involves broad community engagement via outreach and education, from tool developers to end-users. These efforts will create a software ecosystem and expert user base to support the next generation of human genetics. The pangenome reference will provide improved genomic research standards, data sharing and reproducible cloud-based workflows. Understanding the barriers to adoption will lead to effective outreach and training, ensuring that the pangenome reference resource is widely adopted.

Adoption will ultimately be driven by the creation of a data resource that sustains continued improvement in its accuracy and completeness, enables a range of uses and improves genomic analyses. We will actively publicize the benefits of using the pangenome. As a starting point for our outreach efforts, we have created a website (<https://humanpangenome.org>) to publicize the consortium. We have also created social media accounts for the human pangenome that directly connect our consortium with the end-user community (for example, @HumanPangenome on Twitter).

To facilitate adoption, we will explore who the user community will be, their needs and, most importantly, the technical and non-technical barriers that they may encounter. Addressing potential obstacles is essential, as we know that adopting an updated version of the linear reference can result in

significant bottlenecks for many laboratories. The cost of switching can be significant, and the HPRC is aware that many clinical laboratories worldwide still use the GRCh37 build from February 2009 for this reason. The HPRC will examine how to reduce switching costs and expedite transition. User data will be collected in self-reporting surveys, including user characteristics, location, specific applications and barriers to adopting a pangenome reference framework.

Creating a coordinate system that builds on GRCh38 and includes both GRCh37 and GRCh38 assemblies is central to user adoption. The HPRC will develop training materials that explain the additional sequences included in the pangenome reference coordinates and how these sequences relate to GRCh37 or GRCh38. Existing linear reference tools will continue to work with the expanded pangenome reference coordinate system, and pangenome-based results will be translatable to these existing coordinate systems with improved genotype accuracy.

We will develop liftOver tools that make it easy to go backward from the pangenome reference to GRCh37 or GRCh38 when necessary. We already have algorithms for this purpose and demonstration of functionality to predict read mappings from a prototype pangenome to GRCh37 or GRCh38. We will precompute all mappings between the previous assemblies and the pangenome and provide these coordinate translation functions with the pangenome reference release. This information should ease the transition of other databases and resources that rely on these coordinates and provide an annotation directly onto the GRCh37 or GRCh38 assemblies in areas where mappings and interpretation on the pangenome are more reliable than current linear sequence representations.

We will augment the displays of the human genome browser to transition to any haplotype assembly in the pangenome reference and display the haplotype alignments. Visualizations will include relevant genetic backgrounds for specific tracks, for example, picking the right HLA haplotype for a read mapping track. To ensure that we use these tools effectively, we will add detailed information that explains these novel views to our existing training materials and make this information part of our respective workshops.

We have adopted the GA4GH principles and will develop exchange formats analogous to SAM/BAM and utility libraries analogous to htslib/samtools, facilitating the development of transition tools and workflows for the pangenome reference. We will deposit these tools and their guides in the HPRC resource repository. We have also developed a prototype transcript archive that facilitates annotation discovery in GRCh37, GRCh38, CHM13 and the pangenome, and visualizes the differences between two transcripts (for example, on two different genomes).

We aim to engage pilot users to obtain feedback about these resources. The HPRC programme and related tool developers connected with the community of users will develop new tools that gain additional value from using the pangenome reference rather than linear reference genome assemblies. We will report on our discoveries in publications and talks, through the blog, webinars and on the HPRC website, and provide educational tools and forums on using and switching to a pangenome reference.

Relevance to disease research

We expect that the resources and methods that we are developing will profoundly impact studies of the genetic basis of human disease and precision medicine. Although we recognize that adoption by the clinical research community will take time, there are three important benefits to using a pangenome reference. First, a more complete reference that incorporates and displays human genetic diversity will produce fewer ambiguous mappings and more accurate analyses of copy number variation throughout the genome when patient samples are sequenced and analysed^{[63,64](#)}. This will improve genetic diagnosis and the functional annotation of variants. Second, the resource will enable the discovery of disease-risk alleles and previously unobserved rare variants, especially in regions that are inaccessible to standard, short-read sequencing technologies. Studies of unsolved Mendelian genetic disease, for example, have shown that approximately 25% of ‘missing’ disease variants can be recovered when longer reads are applied and more complex repetitive regions are characterized^{[65](#)}. Important genetic risk loci, such as *SMN1* and *SMN2* (spinal muscular atrophy), *LPA* (lipoprotein A and coronary heart disease), *CYP2D6*

(pharmacogenomics), as well as numerous triplet repeat expansion loci are now being sequenced and assembled in large human cohort studies. These studies are revealing the standing pattern of natural genetic variation for loci that are typically excluded from previous analyses^{51,63}. Resolution of these loci by long-read sequencing in even a limited number of human haplotypes improves our ability to genotype them in other patient-derived short-read datasets, allowing for the discovery of new genetic associations, through both genome-wide association study and expression quantitative trait locus methods⁵¹. Last, the pangenome approach represents a fundamental change in how human genetic variation is discovered. Instead of simply mapping sequence reads to a reference, we are constructing phased genome assemblies and aligning them to the graph, which in turn will pinpoint all genetic differences, both large and small, at the base-pair level^{26,66}. As long-read sequencing costs fall and pangenome methods evolve²⁶, we predict that patient samples will probably be sequenced using long-read technology to increase sensitivity and accuracy.

Outlook

As we write this Perspective article, the world is reeling from the COVID-19 pandemic and the spread of new SARS-CoV-2 variants. Scientists can trace the epidemiology of the virus, determine why humans are susceptible^{67,68} and determine why some individuals are more susceptible than others^{69,70}. The current GRCh38 human reference is one of many resources that have made this possible, but we know that it can be improved. Through years of strategic investments in the public and private sectors, we find ourselves with the technologies and methods to build additional references that better represent global human genomic diversity.

The human pangenome reference will collect accurate haplotype-phased genome assemblies generated by efficient algorithmic innovations, which we anticipate will be widely used by the scientific community. The collection of individual genomes, comprising sequence information, genomic coordinates and annotations, will be a critical resource with more accurate representation of human genomic diversity. The original Human Genome Project enabled major advances in human health and genomic medicine^{1,2,3,4}; it is time to

build a more inclusive resource with better representation of human genomic diversity to better serve humanity.

References

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Gibbs, R. A. The Human Genome Project changed everything. *Nat. Rev. Genet.* **21**, 575–576 (2020).
4. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
5. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
6. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* **21**, 243–254 (2020).
7. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
8. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* **25**, 489–494 (2009).
9. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
10. Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature* **475**, 163–165 (2011). **Emphasizes the importance of reference data from ancestral and diverse genomes, as well as stating that researchers should invest time and money into**

education and outreach to explain why studying global (and local) health is so important.

11. Miga, K. H. & Wang, T. The need for a human pangenome reference sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021).
12. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
13. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018). **A model for presenting genomes that aims to improve read mapping by representing genetic variation in the reference.**
14. Martiniano, R., Garrison, E., Jones, E. R., Manica, A. & Durbin, R. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biol.* **21**, 250 (2020).
15. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
16. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
17. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
18. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
19. Li, R. et al. Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
20. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020). **The sequence of the first**

complete human chromosome.

21. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
22. Nurk, S. et al. The complete sequence of a human genome. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.05.26.445798> (2021). **The first complete genome assembly issued from the T2T Consortium, which closed all remaining gaps in the GRCh38, including all acrocentric short arms, segmental duplications and human centromeric regions.**
23. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
24. Tettelin, H. et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc. Natl Acad. Sci. USA* **102**, 13950–13955 (2005).
25. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* **23**, 148–154 (2015).
26. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief Bioinform.* **19**, 118–135 (2018).
27. Eizenga, J. M. et al. Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).
28. Rehm, H. L. et al. ClinGen—the clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
29. Genomes Project Consortium. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
30. Popejoy, A. B. et al. The clinical imperative for inclusivity: race, ethnicity, and ancestry (REA) in genomics. *Hum. Mutat.* **39**, 1713–1720 (2018).

31. Popejoy, A. B. et al. Clinical genetics lacks standard definitions and protocols for the collection and use of diversity measures. *Am. J. Hum. Genet.* **107**, 72–82 (2020).
32. Bonham, V. L. et al. Physicians' attitudes toward race, genetics, and clinical medicine. *Genet. Med.* **11**, 279–286 (2009).
33. Race, Ethnicity & Genetics Working Group. The use of racial, ethnic, and ancestral categories in human genetics research. *Am. J. Hum. Genet.* **77**, 519–532 (2005).
34. Dodson, M. & Williamson, R. Indigenous peoples and the morality of the Human Genome Diversity Project. *J. Med. Ethics* **25**, 204–208 (1999).
35. Couzin-Frankel, J. Ethics. DNA returned to tribe, raising questions about consent. *Science* **328**, 558 (2010).
36. Dukepoo, F. C. The trouble with the Human Genome Diversity Project. *Mol. Med. Today* **4**, 242–243 (1998).
37. Fox, K. The illusion of inclusion—the “All of Us” research program and Indigenous peoples’ DNA. *N. Engl. J. Med.* **383**, 411–413 (2020).
38. Devaney, S. A., Malerba, L. & Manson, S. M. The “All of Us” program and Indigenous peoples. *N. Engl. J. Med.* **383**, 1892 (2020).
39. Hudson, M. et al. Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nat. Rev. Genet.* **21**, 377–384 (2020).
40. Carroll, S. R., Herczog, E., Hudson, M., Russell, K. & Stall, S. Operationalizing the CARE and FAIR principles for Indigenous data futures. *Sci. Data* **8**, 108 (2021).
41. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

42. Genome in a Bottle. *NIST* <https://www.nist.gov/programs-projects/genome-bottle> (updated 16 February 2022).
43. Jarvis, E. D. et al. Automated assembly of high-quality diploid human reference genomes. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.03.06.483034> (2021).
44. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with HiFiasm. *Nat. Methods* **18**, 170–175 (2021). **HiFiasm is a haplotype-resolved assembler specifically designed for PacBio HiFi reads that aims to represent haplotype information in a phased assembly graph.**
45. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
46. Schatz, M. C. et al. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genom.* **2**, 100085 (2022). **The AnVIL platform provides scalable solutions for genomic data access, analysis and education.**
47. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with Minigraph. *Genome Biol.* **21**, 265 (2020). **The Minigraph toolkit has been used to efficiently construct a pangenome graph, which is useful for mapping and constructing graphs that encode structural variation.**
48. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
49. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
50. Rosen, Y., Eizenga, J. & Paten, B. Modelling haplotypes with respect to reference cohort variation graphs. *Bioinformatics* **33**, i118–i123 (2017).

51. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021). **The use of long-read data from 64 human genomes to predict structural variants and the patterns of variation across diverse populations.**
52. Abel, H. J. et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).
53. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
54. Paten, B. et al. Cactus: algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011). **Cactus is a highly accurate, reference-free multiple genome alignment program that is useful for studying general rearrangement and copy number variation.**
55. Pangenome Graph Builder. *GitHub* <https://github.com/pangenome/pggb> (2022).
56. O’Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
57. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
58. Spooner, W. et al. Haplosaurus computes protein haplotypes for use in precision drug design. *Nat. Commun.* **9**, 4128 (2018).
59. Arita, M., Karsch-Mizrachi, I. & Cochrane, G. The international nucleotide sequence database collaboration. *Nucleic Acids Res.* **49**, D121–D124 (2021).
60. Clarke, L. et al. The 1000 Genomes Project: data management and community access. *Nat. Methods* **9**, 459–462 (2012).

61. Clarke, L. et al. The International Genome Sample Resource (IGSR): a worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Res.* **45**, D854–D859 (2017).
62. Courtot, M. et al. BioSamples database: an updated sample metadata hub. *Nucleic Acids Res.* **47**, D1172–D1178 (2019).
63. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.05.26.445678> (2021).
64. Aganezov, S. et al. A complete reference genome improves analysis of human genetic variation. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.07.12.452063> (2021). **The importance of complete T2T genomes in novel variant discovery and of offering major improvements of variant calls within clinically relevant genes are highlighted.**
65. Miller, D. E. et al. Targeted long-read sequencing identifies missing disease-causing variation. *Am. J. Hum. Genet.* **108**, 1436–1449 (2021).
66. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
67. Kim, D. et al. The architecture of SARS-CoV-2 transcriptome. *Cell* **181**, 914–921.e90 (2020).
68. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
69. Toh, C. & Brody, J. P. Evaluation of a genetic risk score for severity of COVID-19 using human chromosomal-scale length variation. *Hum. Genomics* **14**, 36 (2020).
70. Zeberg, H. & Paabo, S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* **587**, 610–612 (2020).

71. Okubo, K., Sugawara, H., Gojobori, T. & Tateno, Y. DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res.* **34**, D6–D9 (2006).
72. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
73. Navarro Gonzalez, J. et al. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.* **49**, D1046–D1057 (2021).
74. Stalker, J. et al. The Ensembl web site: mechanics of a genome browser. *Genome Res.* **14**, 951–955 (2004).
75. Howe, K. L. et al. Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
76. Zhou, X. et al. The Human Epigenome Browser at Washington University. *Nat. Methods* **8**, 989–990 (2011).
77. Li, D., Hsu, S., Purushotham, D., Sears, R. L. & Wang, T. WashU Epigenome Browser update 2019. *Nucleic Acids Res.* **47**, W158–W165 (2019).
78. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016). **Analysis of sample descriptions included in the genome-wide association study catalogue indicates that some populations are still under-represented and left behind in studies of genomic medicine.**
79. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun. Biol.* **2**, 9 (2019).
80. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
81. Ulahannan, N. et al. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. Preprint at *bioRxiv* <https://doi.org/10.1101/833590> (2019).

82. Liu, B., Guo, H., Brudno, M. & Wang, Y. deBGA: read alignment with de Bruijn graph-based seed and extension. *Bioinformatics* **32**, 3224–3232 (2016).
83. Limasset, A., Cazaux, B., Rivals, E. & Peterlongo, P. Read mapping on de Bruijn graphs. *BMC Bioinformatics*. **17**, 237 (2016).
84. Heydari, M., Miclotte, G., Van de Peer, Y. & Fostier, J. BrownieAligner: accurate alignment of Illumina sequencing data to de Bruijn graphs. *BMC Bioinformatics* **19**, 311 (2018).
85. 1001 Genomes. GenomeMapper. *1001 Genomes* https://www.1001genomes.org/software/genomemapper_graph.html (accessed 2021).
86. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
87. Hickey, G. et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
88. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
89. Jain, C., Misra, S., Zhang, H., Dilthey, A. & Aluru, S. Accelerating sequence alignment to graphs. *IEEE Int. Parallel and Distributed Processing Symp. (IPDPS)* 451–461 (2019).
90. Dvorkina, T., Antipov, D., Korobeynikov, A. & Nurk, S. SPAAligner: alignment of long diverged molecular sequences to assembly graphs. *BMC Bioinformatics* **21**, 306 (2020).
91. Mokveld, T., Linthorst, J., Al-Ars, Z., Holstege, H. & Reinders, M. CHOP: haplotype-aware path indexing in population graphs. *Genome Biol.* **21**, 65 (2020).
92. Ghaffaari, A. & Marschall, T. Fully-sensitive seed finding in sequence graphs using a hybrid index. *Bioinformatics* **35**, i81–i89 (2019).

93. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
94. Gonnella, G., Niehus, N. & Kurtz, S. GfaViz: flexible and interactive visualization of GFA sequence graphs. *Bioinformatics* **35**, 2853–2855 (2019).
95. Kunyavskaya, O. & Prjibelski, A. D. SGTK: a toolkit for visualization and assessment of scaffold graphs. *Bioinformatics* **35**, 2303–2305 (2019).
96. Mikheenko, A. & Kolmogorov, M. Assembly Graph Browser: interactive visualization of assembly graphs. *Bioinformatics* **35**, 3476–3478 (2019).
97. Beyer, W. et al. Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics* **35**, 5318–5320 (2019).
98. Yokoyama, T. T., Sakamoto, Y., Seki, M., Suzuki, Y. & Kasahara, M. MoMI-G: modular multi-scale integrated genome graph browser. *BMC Bioinformatics* **20**, 548 (2019).
99. ODGI. *GitHub* <https://github.com/pangenome/odgi> (2021).
100. Shlemov, A. & Korobeynikov, A. in *Algorithms for Computational Biology* (eds Holmes, I., Martín-Vide, C. & Vega-Rodríguez, M. A.) 80–94 (Springer, 2019).
101. Ebler, J. et al. Pangenome-based genome inference. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.11.11.378133> (2020).
102. Leggett, R. M. et al. Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de Bruijn graphs. *PLoS ONE* **8**, e60058 (2013).
103. Sibbesen, J. A. et al. Accurate genotyping across variant classes and lengths using variant graphs. *Nat. Genet.* **50**, 1054–1059 (2018).

104. Chen, S. et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
105. Eggertsson, H. P. et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).

Acknowledgements

We thank the NHGRI for funding multiple components to improve and update the Human Genome Reference Program, which has supported the work represented by this report (1U41HG010972, 1U01HG010971, 1U01HG010961, 1U01HG010973 and 1U01HG010963). This work was also supported, in part, by the Intramural Research Program of the NHGRI, NIH (A.M.P.).

Author information

Affiliations

1. Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

Ting Wang, Heather A. Lawson, Caryn Carson, Daofeng Li & Deepak Purushotham

2. Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO, USA

Ting Wang, Caryn Carson, Daofeng Li & Deepak Purushotham

3. McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA

Ting Wang, Lucinda Antonacci-Fulton, Caryn Carson, Robert S. Fulton, Tina A. Graves-Lindsay, Daofeng Li, Joshua F. McMichael & Deepak Purushotham

4. Wellcome Sanger Institute, Cambridge, UK
Kerstin Howe
5. UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA
Julian K. Lucas, Mobin Asri, Xian Chang, Adam M. Novak, Karen H. Miga, Benedict Paten & David Haussler
6. Genome Informatics Section, National Human Genome Research Institute, Bethesda, MD, USA
Adam M. Phillippy
7. Epidemiology Division, Department of Public Health Sciences, University of California, Davis, CA, USA
Alice B. Popejoy
8. Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA
Mark J. P. Chaisson
9. Arizona State University, Barrett & O'Connor Washington Center, Washington DC, USA
Robert Cook-Deegan
10. National Institutes of Health (NIH)–National Human Genome Research Institute, Bethesda, MD, USA
Adam L. Felsenfeld, Baergen I. Schultz, Michael W. Smith & Heidi J. Sofia
11. Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA

Erik P. Garrison

12. Institute for Society & Genetics, College of Letters and Science,
University of California, Los Angeles, Los Angeles, CA, USA

Nanibaa' A. Garrison

13. Institute for Precision Health, David Geffen School of Medicine,
University of California, Los Angeles, Los Angeles, CA, USA

Nanibaa' A. Garrison

14. Division of General Internal Medicine & Health Services Research,
David Geffen School of Medicine, University of California, Los
Angeles, Los Angeles, CA, USA

Nanibaa' A. Garrison

15. Department of Medicine, Stanford University, School of Medicine,
Stanford, CA, USA

Hanlee Ji

16. Department of Genetics and Genomic Science, Icahn School of
Medicine at Mount Sinai, New York, NY, USA

Eimear E. Kenny

17. Department of Medicine, Icahn School of Medicine at Mount Sinai,
New York, NY, USA

Eimear E. Kenny

18. Institute for Genomic Health, Icahn School of Medicine at Mount Sinai,
New York, NY, USA

Eimear E. Kenny

19. Program in Bioethics and Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA

Barbara A. Koenig

20. Heinrich Heine University, Medical Faculty, Institute for Medical Biometry and Bioinformatics, Düsseldorf, Germany

Tobias Marschall

21. National Center for Biotechnology Information (NCBI), National Library of Medicine, Bethesda, MD, USA

Valerie A. Schneider

22. Department of Electrical Engineering, Stanford University, Stanford, CA, USA

Tsachy Weissman

23. European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK

Paul Flicek

24. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Heng Li

25. Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA

Heng Li

26. Vertebrate Genome Lab and and Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, USA

Erich D. Jarvis

27. Howard Hughes Medical Institute, Chevy Chase, MD, USA

Erich D. Jarvis

28. Yale School of Medicine, New Haven, CT, USA

Ira M. Hall

29. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA

Evan E. Eichler

30. Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

Evan E. Eichler

31. Howard Hughes Medical Institute, University of California, Santa Cruz, CA, USA

David Haussler

Consortia

the Human Pangenome Reference Consortium

Contributions

All authors contributed to writing the manuscript.

Corresponding authors

Correspondence to [Ting Wang](#), [Paul Flicek](#), [Heng Li](#), [Karen H. Miga](#), [Benedict Paten](#), [Erich D. Jarvis](#), [Ira M. Hall](#), [Evan E. Eichler](#) or [David Haussler](#).

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Kazuto Kato and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information

Supplementary Information

This file contains Supplementary Table 1 (47 genomes comprise the initial pangenome data resources) and the current membership of the Human Pangenome Reference Consortium.

Rights and permissions

Reprints and Permissions

About this article

Cite this article

Wang, T., Antonacci-Fulton, L., Howe, K. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446

(2022). <https://doi.org/10.1038/s41586-022-04601-8>

- Received: 30 August 2021
- Accepted: 01 March 2022
- Published: 20 April 2022
- Issue Date: 21 April 2022
- DOI: <https://doi.org/10.1038/s41586-022-04601-8>

Share this article

Anyone you share the following link with will be able to read this content:

[Get shareable link](#)

Sorry, a shareable link is not currently available for this article.

[Copy to clipboard](#)

Provided by the Springer Nature SharedIt content-sharing initiative

This article was downloaded by **calibre** from <https://www.nature.com/articles/s41586-022-04601-8>

| [Section menu](#) | [Main menu](#) |

[Download PDF](#)

- Article
- Open Access
- [Published: 20 April 2022](#)

A quantum processor based on coherent transport of entangled atom arrays

- [Dolev Bluvstein](#)¹,
- [Harry Levine](#)¹ nAff6,
- [Giulia Semeghini](#)¹,
- [Tout T. Wang](#)¹,
- [Sepehr Ebadi](#)¹,
- [Marcin Kalinowski](#) [ORCID: orcid.org/0000-0003-0605-8791](#)¹,
- [Alexander Keesling](#)^{1,2},
- [Nishad Maskara](#)¹,
- [Hannes Pichler](#)^{3,4},
- [Markus Greiner](#)¹,
- [Vladan Vuletić](#) [ORCID: orcid.org/0000-0002-9786-0538](#)⁵ &
- [Mikhail D. Lukin](#) [ORCID: orcid.org/0000-0002-8658-1007](#)¹

[Nature](#) volume **604**, pages 451–456 (2022)

- 619 Accesses
- 1 Citations
- 34 Altmetric
- [Metrics details](#)

Subjects

- [Atomic and molecular physics](#)
- [Quantum information](#)
- [Quantum mechanics](#)
- [Quantum simulation](#)
- [Qubits](#)

Abstract

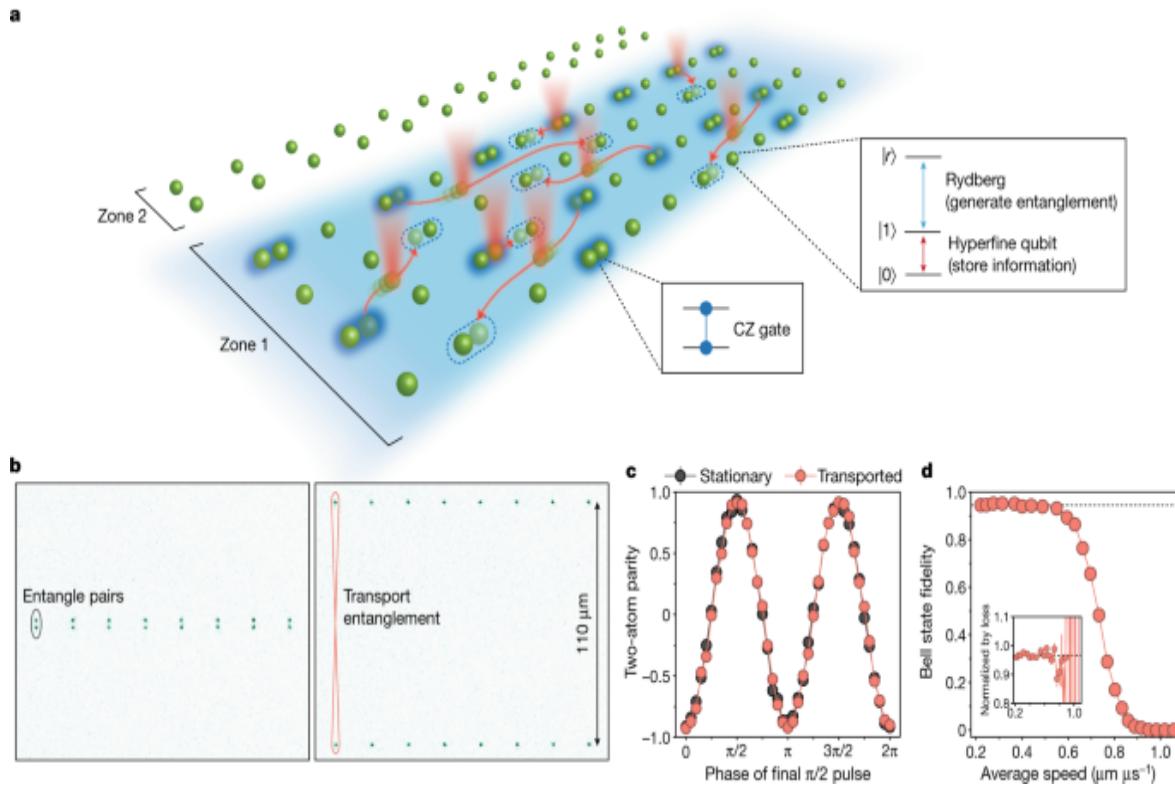
The ability to engineer parallel, programmable operations between desired qubits within a quantum processor is key for building scalable quantum information systems^{1,2}. In most state-of-the-art approaches, qubits interact locally, constrained by the connectivity associated with their fixed spatial layout. Here we demonstrate a quantum processor with dynamic, non-local connectivity, in which entangled qubits are coherently transported in a highly parallel manner across two spatial dimensions, between layers of single- and two-qubit operations. Our approach makes use of neutral atom arrays trapped and transported by optical tweezers; hyperfine states are used for robust quantum information storage, and excitation into Rydberg states is used for entanglement generation^{3,4,5}. We use this architecture to realize programmable generation of entangled graph states, such as cluster states and a seven-qubit Steane code state^{6,7}. Furthermore, we shuttle entangled ancilla arrays to realize a surface code state with thirteen data and six ancillary qubits⁸ and a toric code state on a torus with sixteen data and eight ancillary qubits⁹. Finally, we use this architecture to realize a hybrid analogue–digital evolution² and use it for measuring entanglement entropy in quantum simulations^{10,11,12}, experimentally observing non-monotonic entanglement dynamics associated with quantum many-body scars^{13,14}. Realizing a long-standing goal, these results provide a route towards scalable quantum processing and enable applications ranging from simulation to metrology.

Main

Quantum information systems derive their power from controllable interactions that generate quantum entanglement. However, the natural, local character of interactions limits the connectivity of quantum circuits and simulations. Non-local connectivity can be engineered via a global shared quantum data bus^{[15,16,17,18](#)}, but in practice these approaches have been limited in either control or size. A number of visionary architectures to address this challenge have been proposed theoretically over the past two decades. On the basis of coherent, dynamical transport of quantum information using movable traps or photonic links, these techniques have been the subject of intensive experimental explorations across different platforms^{[1,19,20,21,22,23,24](#)}. However, progress has been limited to small-scale, few-qubit systems lacking either full connectivity, programmability or true parallelism.

Our approach to address this long-standing challenge utilizes dynamically reconfigurable arrays of entangled neutral atoms, shuttled by optical tweezers in two spatial dimensions (Fig. [1a](#)). Hyperfine states are used for storing and transporting quantum information between quantum operations, and excitation into Rydberg states is used for generating entanglement. Highly parallel operations are enabled via selective qubit operations in distinct zones that qubits are dynamically shuttled between. Taken together, these ingredients enable a powerful quantum information architecture, which we employ to realize applications including entangled-state generation, the creation of topological surface and toric code states, and hybrid analogue–digital quantum simulations.

Fig. 1: Quantum information architecture enabled by coherent transport of neutral atoms.



a, In our approach, qubits are transported to perform entangling gates with distant qubits, enabling programmable and non-local connectivity. Atom shuttling is performed using optical tweezers, with high parallelism in two dimensions and between multiple zones allowing selective manipulations. Inset: the atomic levels used. The $|0\rangle$, $|1\rangle$ qubit states refer to the $m_F = 0$ clock states of ^{87}Rb , and $|r\rangle$ is a Rydberg state used for generating entanglement between qubits (Extended Data Fig. 1b). **b**, Atom images illustrating coherent transport of entangled qubits. Using a sequence of single-qubit and two-qubit gates, atom pairs are each prepared in the $|\Phi^+\rangle$ Bell state (Methods), and are then separated by $110\ \mu\text{m}$ over a span of $300\ \mu\text{s}$. **c**, Parity oscillations indicate that movement does not observably affect entanglement or coherence. For both the moving and the stationary measurements, qubit coherence is preserved using an XY8 dynamical decoupling sequence for $300\ \mu\text{s}$ (Methods). **d**, Measured Bell-state fidelity as a function of separation speed over the $110\ \mu\text{m}$, showing that fidelity is unaffected for a move slower than $200\ \mu\text{s}$ (average separation speed of $0.55\ \mu\text{m}\ \mu\text{s}^{-1}$). Inset: normalizing by atom loss during the move results in constant fidelity, indicating that atom loss is the dominant error mechanism (see Methods for details).

Entanglement transport in atom arrays

Our experiments utilize a two-dimensional (2D) atom array system described previously²⁵, with key upgrades to enable coherent transport and multiple layers of single-qubit and two-qubit gates. We store quantum information in magnetically insensitive clock states within the ground-state hyperfine manifold of ⁸⁷Rb atoms²⁰, and implement robust single-qubit Raman rotations (scattering error per π pulse of about 7×10^{-5})²⁶, realized by composite pulses that are robust to pulse errors (Extended Data Fig. 3)²⁷. High-fidelity controlled-Z (CZ) entangling gates in the hyperfine basis $\{|0\rangle, |1\rangle\}$ (Fig. 1a) are implemented in parallel using global Rydberg excitation pulses on the $|1\rangle \leftrightarrow |r\rangle$ Rydberg transition⁵. For dynamic reconfiguration, we initialize atoms into two sets of traps: static traps generated by a spatial light modulator (SLM) and mobile traps generated by a crossed 2D acousto-optic deflector (AOD). To execute a specific circuit, we arrange qubits into desired pairs, perform Rydberg-mediated CZ gates on each pair simultaneously and then move all mobile traps in parallel to dynamically change the connectivity into the next desired qubit arrangement.

Figure 1 shows our ability to transport qubits across large distances while preserving entanglement and coherence²⁰. We initialize pairs at an atom–atom distance of 3 μm (Fig. 1b) and then create a Bell state $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ in the hyperfine basis (Methods)⁵. To measure the resulting entangled-state fidelity, we apply a variable single-qubit phase gate before a final $\pi/2$ pulse, resulting in oscillations of the two-atom parity $\langle \sigma_1 z \sigma_2 z \rangle$ (Fig. 1c)⁵. We then repeat this experiment, but now move the atoms apart by 110 μm before applying the final $\pi/2$ pulse. Our transport protocol is optimized to suppress heating and loss by implementing cubic-interpolated atom trajectories (Methods), and is further accompanied by an eight-pulse XY8 robust dynamical decoupling sequence²⁸ to suppress dephasing. The resulting parity oscillations indicate that two-atom entanglement is unaffected by the transport process^{20,29}. Performing this experiment as a function of movement speed³⁰ shows that the fidelity remains unchanged until the total separation speed becomes more than 0.55

$\mu\text{m } \mu\text{s}^{-1}$, corresponding to the onset of atom loss (Fig. 1d). We note that the entanglement transport in Fig. 1b corresponds to moving quantum information across a region of space that can, in principle, host about 2,000 qubits (at an atom separation of 3 μm), on a timescale corresponding to $<10^{-3}$ of the coherence time T_2 (Extended Data Fig. 3), directly enabling applications in large-scale quantum information systems.

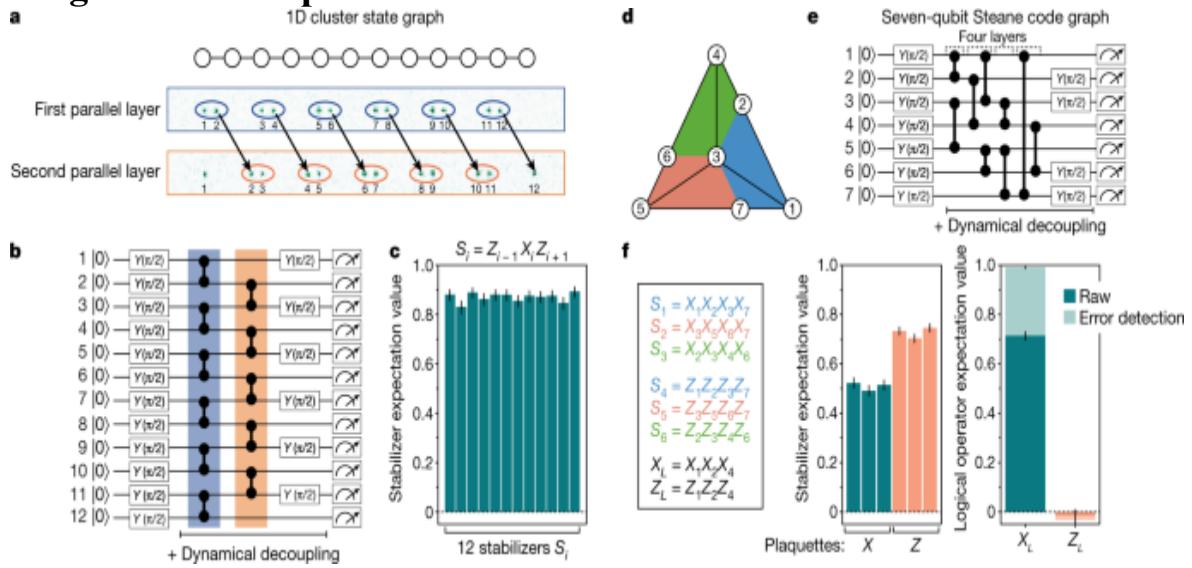
Programmable circuits and graph states

To exemplify the ability to generate non-local connectivity between qubit arrays in parallel, we carry out the preparation of entangled graph states: a large class of useful quantum information states, with examples ranging from Greenberger–Horne–Zeilinger states and cluster states to quantum error correction (QEC) codes³¹. Graph states are defined by initializing all qubits, located on the vertices of a geometric graph, in $\frac{(|+| - |-\rangle)}{\sqrt{2}}$ and then performing CZ gates on the links between qubits (corresponding to the edges of the graph)³¹. N -qubit graph states $|G\rangle$ are associated with a set of N stabilizers, defined by $\langle\langle S_i\rangle\langle X_i\rangle\langle\Pi_j|_{j\in u_i}\rangle\langle Z_j\rangle$, where X and Z are the Pauli matrices, u_i is the set of qubits (vertices) connected by an edge to qubit i , and Π denotes a product over qubit indices j (ref. 31). The stabilizers each have +1 eigenvalue for the graph state $|G\rangle$. Measuring these operators and their expectation values can be used to characterize the preparation of the target state.

As an example, Fig. 2a shows the preparation of a one-dimensional (1D) cluster state, a graph state defined by a linear chain of qubits^{32,33}. To realize this state, we perform one global, parallel layer of CZ gates on adjacent atom pairs, move half the atoms to form new pairs and then perform another parallel layer of CZ gates (Fig. 2a, b). To probe the resultant 12-qubit cluster state, we measure the stabilizer set $\{S_i\} = \{Z_{i-1}X_iZ_{i+1}\}$ through readout in two measurement settings, given by a local $\pi/2$ rotation on either the odd or the even sublattice before projective measurement³⁴. We achieve the local rotation by moving one sublattice of qubits to a separate zone and then performing a rotation on the unmoved qubits with a homogeneous beam illuminating the experiment zone (Fig. 1a, Methods). We measure $\langle S_i \rangle$ by

analysing the resulting bit-string outputs and plot the resulting raw stabilizer measurements (Fig. 2c). Across all 12 stabilizers, we find an average $\langle S_i \rangle = 0.87(1)$ (Fig. 2c) (accounting for state-preparation-and-measurement (SPAM) errors would yield $\langle S_i \rangle = 0.91(1)$), certifying biseparable entanglement in a cluster state (all $\langle S_i \rangle > 0.5$ (ref. 34)). The measured fidelities would correspond to a few-percent error per operation for a measurement-based quantum computation^{32,35}.

Fig. 2: One- and two-dimensional graph states using dynamic entanglement transport.



a, Generation of a 12-atom 1D cluster-state graph, created by initializing all qubits (vertices) in $|+\rangle$ and applying CZ gates on the links (edges) between qubits. The atom images show the configuration for the first and second gate layers. **b**, Quantum circuit representation of the 1D cluster-state preparation and measurement. Dynamical decoupling is applied throughout all quantum circuits ([Methods](#)). **c**, Raw measured stabilizers of the resulting 1D cluster state, given by $S_i = Z_{i-1}X_iZ_{i+1}$ (X_1Z_2 and $Z_{11}X_{12}$ for the edge qubits). **d**, Graph-state representation of the seven-qubit Steane code (colours represent stabilizer plaquettes). **e**, Circuit for preparing the Steane code logical $|+\rangle_L$ state, performed in four parallel gate layers. **f**, Measured stabilizers and logical operators after preparing $|+\rangle_L$. Error detection is done by postselecting on measurements where all stabilizers are +1. For both the 1D cluster state and the Steane code, the stabilizers and logical operators are

measured with two measurement settings (see text). Error bars represent 68% confidence intervals.

An important class of graph states are QEC codes, where the graph-state stabilizers manifest as the stabilizers of the QEC code and can be measured to correct errors on an encoded logical qubit. As an example, we prepare the seven-qubit Steane code^{6,7}, a topological colour code depicted by the graph in Fig. 2d, in the logical state $|+\rangle_L$. To prepare this state, we initialize all qubits in $|+\rangle$, apply CZs on the links between qubits (in four parallel layers; Extended Data Fig. 5) and then rotate either of the two sublattices for measuring stabilizers (Fig. 2e). After sublattice rotation, six of the graph-state stabilizers transform into the six Steane code stabilizers, given by four-body products of X_i or Z_i . Figure 2f shows the raw measured expectation values of these six stabilizers. The seventh graph-state stabilizer transforms into the logical qubit operator X_L and has eigenvalue +1 for the graph state $|G\rangle$, while anticommuting with logical Z_L . Accordingly, in Fig. 2f, we find $\langle X_L \rangle = 0.71(2)$ and $\langle Z_L \rangle = -0.02(3)$, demonstrating the preparation of the logical qubit state $|+\rangle_L$. Moreover, we perform error detection by post-selecting on measurement outcomes where all measured stabilizers yield +1 (refs. 36,37; with 66(1)% probability of no detected errors). Using this procedure, we obtain corrected values of $\langle \bar{X} \rangle_L = \{0.991\}_{-0.007}^{+0.004}$ and $\langle \bar{Z} \rangle_L = -0.03(3)$ demonstrating the error-detecting properties of the Steane code graph (see Extended Data Fig. 7 for error correction and logical operations).

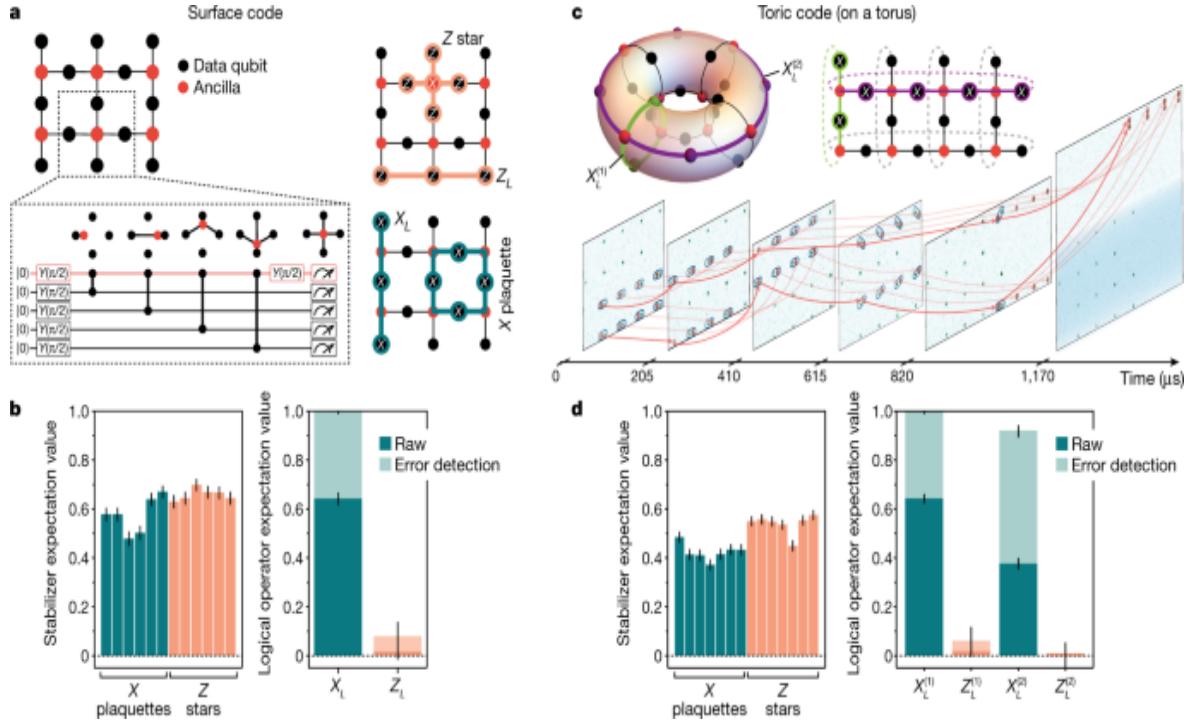
Topological states with ancilla arrays

We next make use of transportable ancillary qubit arrays to mediate quantum operations between remote qubits¹. Owing to the ability to quickly move arrays of atoms across the entire system, the use of ancillary qubits naturally complements our movement capabilities. Specifically, we employ ancillas for state preparation by mediating entanglement between physical qubits that never directly interact, followed by projective measurement of the ancilla array (performed simultaneously with the measurement of the data qubits), a form of measurement-based quantum computation^{32,35}. In particular, we

prepare topological surface code and toric code states^{8,9,38,39}, whose states are more difficult to construct by direct CZ gates between physical qubits (requiring an extensive number of layers^{8,40}). For these codes, the measured values of the ancilla qubits simply redefine the stabilizers and are handled in-software for practical QEC operation³⁸. As the redefinition is applied in-software, without physical intervention, the projective measurements on the ancillae commute with all operations on the data qubits and can be done at any time, and so we measure all qubits simultaneously.

Figure 3a shows the preparation of a 19-qubit graph state creating the $|+\rangle_L$ logical state of the surface code^{8,38}. The surface code is defined by X -plaquette and Z -star stabilizers, and logical operators X_L (Z_L) are defined as strings of X (Z) products across the height (width) of the graph. To prepare this state, ancillas are moved to perform CZ gates with each of their four neighbours and are then measured, projecting the data qubits into the surface code state. The graph-state stabilizers now transform into the X plaquettes, the Z stars (with value ± 1 for a measurement outcome of ± 1 of the central ancilla) and the logical X_L operator^{35,41}. Remarkably, this procedure creates a topologically ordered state in a constant-depth circuit^{35,40}, where measured ancilla values can be used for redefining stabilizers, which can be handled in-software for practical QEC operation³⁸. Figure 3b shows the measured expectation values of the 12 resulting stabilizers, as well as the logical operator expectation values with and without error detection. We find a raw value of $\langle X_L \rangle = 0.64(3)$, and a corrected value of $\langle \bar{X} \rangle = 1_{\pm 0.01}^{+0}$ using the measured stabilizers for error detection (with 35(1)% probability of no detected errors), demonstrating the preparation of this topological QEC state (see also Extended Data Fig. 7, showing the expected attributes for all prepared error-protected logical states).

Fig. 3: Topological surface code and toric code states using mobile ancilla qubit arrays.



a, Graph state realizing the surface code. Left: the circuit depicts formation of the graph state by use of mobile ancilla qubits; each move corresponds to performing a CZ gate with a neighbouring data qubit (illustrated in box). The logical $|+\rangle_L$ state is created upon projective measurement of the ancilla qubits in the X basis. Right: stabilizers and logical operators of the code. **b**, Measured X -plaquette and Z -star stabilizers of the resultant surface code, along with logical operators with and without error detection (implemented in postselection). **c**, Implementation of the toric code. Top: graph state realizing the two logical-qubit product state $\langle|+\rangle_L\langle L^{(1)}|+\rangle_L\langle L^{(2)}|\rangle$ of the toric code upon projective measurement of the ancilla qubits in the X basis. Bottom: images showing the movement steps implemented in creating and measuring the toric code state ([Supplementary Video 1](#)). The blue shading in the final image represents a local rotation on the data qubit zone. **d**, Measured X -plaquette and Z -star stabilizers, along with logical operators for the two logical qubits with and without error detection (implemented in postselection).

Although surface code states have previously been prepared with other methods, our transport capabilities allow us to use the full range of motion of ancilla qubits across the entire qubit array to enable periodic boundary conditions and realize the toric code state on a torus⁹. To this end, we create

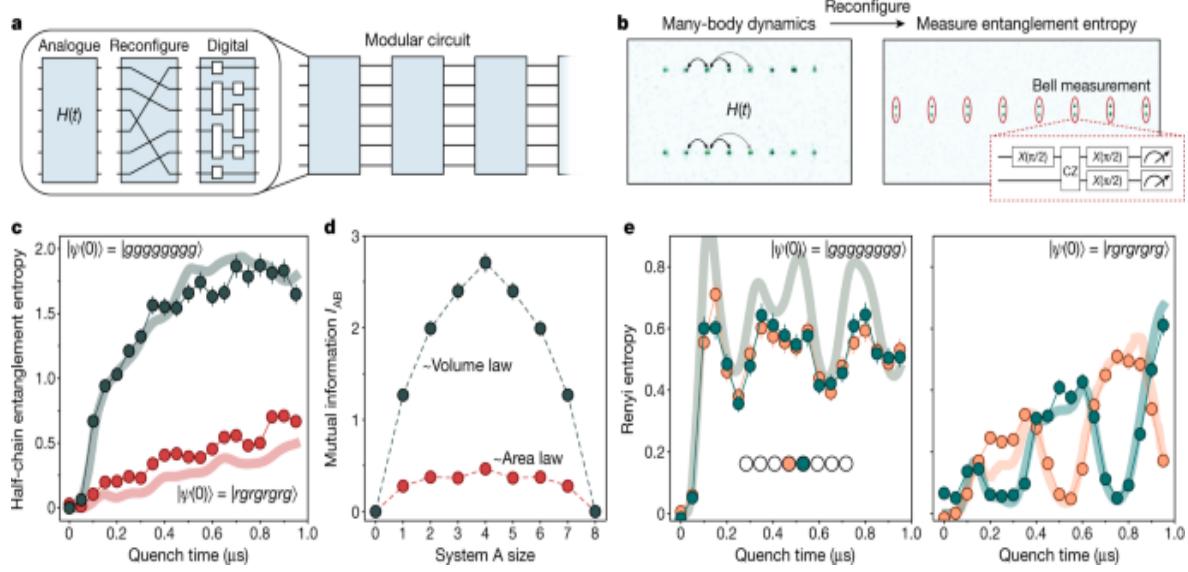
the 24-qubit graph state shown in Fig. 3c by performing five layers of parallel gates and moving the ancillae to their separate zone for readout in a separate basis (see also [Supplementary Video 1](#) showing the full atom trajectory). The state we prepare has seven (owing to periodic boundary conditions) independent X plaquettes and seven independent Z stars. Moreover, owing to the topological properties of this graph, two independent logical qubits can be encoded with logical operators $\langle\{X\}_L^{\{1\}}, \{Z\}_L^{\{1\}}\rangle$ and $\langle\{X\}_L^{\{2\}}, \{Z\}_L^{\{2\}}\rangle$ that wrap around the entire torus along two topologically distinct directions⁹. Upon projective measurement of the ancilla qubits in the X basis we create the toric code state $\langle|+\rangle_L^{\{1\}}|+\rangle_L^{\{2\}}\rangle$. State preparation is verified in Fig. 3d by measuring the toric code stabilizers. For the two encoded logical qubits, we find raw logical qubit expectation values of $\langle X \rangle_L^{\{1\}} = 0.64(2)$ and $\langle X \rangle_L^{\{2\}} = 0.38(2)$, and error-detected values of $\langle \bar{X} \rangle_L^{\{1\}} = 1_{-0.01}^{+0}$ and $\langle \bar{X} \rangle_L^{\{2\}} = 0.92_{-0.03}^{+0.02}$ (with 20(1)% probability of no detected errors), demonstrating the preparation of the toric code. We note that the different expectation values of the corrected logical qubits originate from the aspect ratio of our torus, where $\langle\{X\}_L^{\{1\}}\rangle$ and $\langle\{X\}_L^{\{2\}}\rangle$ are protected to code distance $d=4$ and $d=2$, respectively (see also Extended Data Fig. 7). Our measured fidelities are in good agreement with numerical simulations of the circuit (Extended Data Fig. 6), wherein each qubit experiences a per-layer error rate independent of the number of qubits or the shuttling process, indicating that errors in CZ gates (fidelity of about 97.5%; Methods⁵) constitute our dominant error source.

Hybrid analogue–digital circuits

Having established atom movement for realizing digital circuits, we now explore the applications to quantum simulation. In particular, we perform hybrid, modular quantum circuits composed of analogue Hamiltonian evolution, reconfiguration and digital gates (Fig. 4a). Together, these tools open a wide variety of possibilities in quantum simulation and many-body physics. As a specific example, we measure the Renyi entanglement entropy

after a quantum quench by effectively interfering two copies of a many-body system^{10,11}.

Fig. 4: Dynamic reconfigurability for hybrid analogue–digital quantum simulation.



a, Hybrid quantum circuit combining coherent atom transport with analogue Hamiltonian evolution and digital quantum gates. **b**, Measuring entanglement entropy in a many-body Rydberg system via two-copy interferometry. **c**, Measured half-chain Renyi entanglement entropy after many-body dynamics following quenches on two eight-atom systems. Quenching from $|gggg\dots\rangle$ ($|g\rangle \equiv |1\rangle$) results in rapid entropy growth and saturation, signifying quantum thermalization. Quenching from $|rgrg\dots\rangle$ reveals a significantly slower growth of entanglement entropy. **d**, Measuring the mutual information at 0.5- μs quench time reveals a volume-law scaling for the thermalizing $|gggg\dots\rangle$ state, and an area-law scaling for the scarring $|rgrg\dots\rangle$ state. **e**, The single-site Renyi entropies for sites in the middle of the chain quickly increase and saturate for the $|gggg\dots\rangle$ quench, but show large oscillations for the $|rgrg\dots\rangle$ quench. The solid curves are results of exact numerical simulations for the isolated quantum system under H_{Ryd} with no free parameters (see [Methods](#) for details of data processing). Error bars represent one standard deviation.

Figure 4b illustrates the experimental procedure. After initializing both copies with all qubits in $|1\rangle$, we independently evolve each copy under the

Rydberg Hamiltonian H_{Ryd} for a time t , generating an entangled many-body state in the $\{|1\rangle, |r\rangle\}$ basis (Methods)¹³. Raman and Rydberg π pulses then map $|1\rangle \rightarrow |0\rangle$ and $|r\rangle \rightarrow |1\rangle$, transferring the entangled many-body state into the long-lived and non-interacting $\{|0\rangle, |1\rangle\}$ basis⁴². Finally, we measure entanglement entropy by rearranging the system and interfering each qubit in the first copy with its identical twin in the second copy, by use of a Bell measurement circuit. Measuring twins in the Bell basis detects occurrences of the antisymmetric singlet state $\langle|\{\Psi\}^{\wedge\{-}\rangle| = \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle)$, the presence of which indicates that subsystems of the two copies were in different states owing to entanglement with the rest of the many-body system^{10,11}. Quantitatively, analysing the number parity of observed singlets within subsystem A yields the purity $\langle(\text{Tr}\{\rho_A\})^2\rangle$ of the reduced density matrix ρ_A , and thus yields the second-order Renyi entanglement entropy $S_2(A) = -\log\langle(\text{Tr}\{\rho_A\})^2\rangle$ (Methods). This measurement circuit provides the Renyi entropy of any constituent subsystem of our whole closed quantum system, where the calculation over any desired subsystem A is simply performed in data processing^{10,11}.

We use this method to probe the growth of entanglement entropy produced by many-body dynamics (see [Methods](#) for additional benchmarking of the technique). Specifically, we study the evolution of two eight-atom copies under the Rydberg Hamiltonian, subject to the nearest-neighbour blockade constraint^{4,13}. Upon a rapid quench from an initial state with all atoms in the ground state $|g\rangle \equiv |1\rangle$, we observe that the half-chain Renyi entanglement entropy quickly grows and saturates (Fig. 4c), a process corresponding to quantum thermalization¹². By analysing the Renyi mutual information $I_{AB} = S_2(A) + S_2(B) - S_2(AB)$ between the leftmost n atoms in the chain (A) and the complement subsystem of the rightmost $8 - n$ atoms (B), we find a volume-law scaling in the resulting state (Fig. 4d)^{11,12}.

Although such thermalizing dynamics is generically expected in strongly interacting many-body systems, remarkably, it was demonstrated previously that for certain initial states this system can evade thermalization.

Underpinned by special, non-thermal eigenstates called quantum many-body scars^{13,14,43}, these states were theoretically predicted to feature dynamics

associated with a slow, non-monotonic entanglement growth. Figure 4 shows the measurement of entanglement properties of many-body scars following a rapid quench from the initial state $\langle |Z\rangle_2 \rangle$, initialized by applying local Rydberg π pulses (Methods). We find that the rate of entropy growth for this initial state is significantly suppressed, and the mutual information reveals an area-law scaling (Fig. 4d). Furthermore, Fig. 4e shows the single-site entropy in the middle of the chain, demonstrating rapid growth and saturation for the thermalizing $|gggg\dots\rangle$ state but large oscillations for the $|Z_2\rangle$ state^{13,14}.

Remarkably, the data show that when sites of one sublattice return to low entropy, the other sublattice goes to high entropy; this reveals that the scar dynamics entangle distant atoms (of the same sublattice) while disentangling nearest neighbours, even with only nearest-neighbour interactions (Methods). These measurements reveal non-trivial aspects of quantum many-body scars, and constitute the direct observation of exotic entanglement phenomena in a many-body system.

These observations are in excellent agreement with exact numerical simulations in the isolated system (lines plotted in Fig. 4c,e, Extended Data Fig. 10). Moreover, whereas the single-site purity approaches that of a fully mixed state, our global purity (a 16-body observable composed of three-level systems) remains more than 100 times that of a fully mixed state (Extended Data Fig. 9), altogether demonstrating the high accuracy and fidelity of our circuit-based technique. These results demonstrate that combining atom movement, many-body Hamiltonian evolution and digital quantum circuits yields powerful tools for simulating and probing the quantum physics of complex systems.

Discussion and outlook

Our experiments demonstrate highly parallel coherent qubit transport and entanglement enabling a powerful quantum information architecture. The present techniques can be extended along a number of directions. Local Rydberg excitation on subsets of qubit pairs would eliminate residual interactions from unintended atoms, allowing parallel, independent operations on arrays with significantly higher qubit densities. Two-qubit gate fidelity can be improved using higher Rydberg laser power or more efficient

delivery methods, as well as more advanced atom cooling⁴⁴. These technical improvements should allow for scaling to deep quantum circuits operating on thousands of neutral atom qubits. These upgrades can be additionally supplemented by more sophisticated local single-qubit control employing, for example, parallel Raman excitation through acousto-optic modulator arrays¹⁶. Mid-circuit readout can be implemented by moving ancillas into a separate zone and imaging using, for example, avalanche photodiode arrays within a few hundred microseconds⁴⁵.

Our method has a clear potential for realizing scalable QEC⁴⁶. For example, the procedure demonstrated in Fig. 3c can be used for syndrome extraction in a practical QEC sequence, wherein ancillas are entangled with their data qubit neighbours and then moved to a separate zone for mid-circuit readout. We estimate that an entire QEC round can be implemented within a millisecond, much faster than the measured $T_2 > 1$ s, and with projected fidelity improvements theoretically surpassing the surface code threshold (Methods). We emphasize that such a mid-circuit readout is essential for realizing scalable fault-tolerant quantum computation. Furthermore, the ability to reconfigure and interlace our arrays will allow efficient, parallel execution of transversal entangling gates between many logical qubits^{38,47}. In addition, these techniques also enable implementation of higher-dimensional or non-local error-correcting codes with more favourable properties^{48,49}. Together, these ingredients could enable an approach to universal, fault-tolerant quantum computing with thousands of physical qubits.

Our dynamically reconfigurable architecture also opens many opportunities for digital and analogue quantum simulations. For example, our hybrid approach can be extended to probing the entire entanglement spectrum⁵⁰, simulating wormhole creation⁵¹, performing many-body purification⁵² and engineering non-equilibrium states⁵³. Entanglement transport could also empower metrological applications such as creating distributed states for probing gravitational gradients⁵⁴. Finally, our approach can facilitate quantum networking between separated arrays, paving the way towards large-scale quantum information systems^{29,55} and distributed quantum metrology^{54,56}.

Methods

Dynamic reconfiguration in 2D tweezer arrays

Our experiments utilize the same apparatus described previously in ref. [25](#). Inside our vacuum cell, ^{87}Rb atoms are loaded from a magneto-optical trap into a backbone array of programmable optical tweezers generated by an SLM^{[57](#)}. Atoms are rearranged in parallel into defect-free target positions in this SLM backbone^{[57](#)} by additional optical tweezers generated from a crossed 2D AOD. Following the rearrangement procedure, we transfer selected atoms from the static SLM traps back into the mobile AOD traps, and then move these mobile atoms to their starting positions in the quantum circuit. During this entire process, the atoms are cooled with polarization-gradient cooling. Before running the quantum circuit, we take a camera image of the atoms in their initial starting positions, and following the circuit we take a final camera image to detect qubit states $|0\rangle$ (atom presence) and $|1\rangle$ (atom loss, following resonant pushout). We postselect all data on finding perfect rearrangement of the AOD and SLM atoms before running the circuit. In all experiments here, each atom remains in a single static or single mobile trap throughout the duration of the quantum circuit^{[20,58,59](#)}.

The crossed AOD system is composed of two independently controlled AODs (AA Opto Electronic DTSX-400) for x and y control of the beam positions. Both AODs are driven by independent arbitrary waveforms, which are generated by a dual-channel arbitrary waveform generator (M4i.6631-x8 by Spectrum Instrumentation) and then amplified through independent MW amplifiers (Minicircuits ZHL-5W-1). The time-domain arbitrary waveforms are composed of multiple frequency tones corresponding to the x and y positions of columns and rows, which are independently changed as a function of time for steering around the AOD-trapped atoms dynamically; the full x and y waveforms are calculated by adding together the time-domain profile of all frequency components with a given amplitude and phase for each component. For running quantum circuits, we program the positions of the AOD atoms at each gate location and then smoothly interpolate (with a cubic profile) the AOD frequencies as a function of time between gate positions. The cubic profile enacts a

constant jerk onto the atoms, which allows us to move roughly five-to-ten-times faster (without heating and loss) than if we move at a constant velocity (linear profile). In our movement protocol, we only do stretches, compressions and translations of the AOD trap array: that is, the AOD rows and columns never cross each other to avoid atom loss and heating associated with two frequency components crossing each other.

We homogenize the AOD tweezer intensity throughout the whole atom trajectory to minimize dephasing induced by a time-varying magnitude of differential light shifts. To this end, we use a reference camera in the image plane to gauge the intensity of each AOD tweezer at each gate location and homogenize by varying the amplitude of each frequency component; during motion between two locations, we interpolate the amplitude of each individual frequency component.

The SLM tweezer light (830 nm) and the AOD tweezer light (828 nm) are generated by two separate, free-running titanium:sapphire lasers (M Squared, 18-W pump). Projected through a 0.5 numerical aperture objective, the SLM tweezers have a waist of roughly 900 nm (roughly 1,000 nm for AODs). When loading the atoms, the trap depths are about $2\pi \times 16$ MHz, with radial trap frequencies of about $2\pi \times 80$ kHz, and when running quantum circuits the trap depths are about $2\pi \times 4$ MHz, with radial trap frequencies of about $2\pi \times 40$ kHz.

Raman laser system

Fast, high-fidelity single-qubit manipulations are critical ingredients of the quantum circuits demonstrated in this work. To this end, we use a high-power 795-nm Raman laser system for driving global single-qubit rotations between magnetic sublevel $m_F = 0$ clock states. This Raman laser system is based on dispersive optics, developed and described in ref. [26](#). The 795-nm light (Toptica TA pro, 1.8 W) is phase-modulated by an electro-optic modulator (Qubig), which is driven by microwaves at 3.4 GHz (Stanford Research Systems SRS SG384) that are doubled to 6.8 GHz and amplified. The laser phase modulation is converted to amplitude modulation for driving Raman transitions through use of a chirped Bragg grating (Optigrate) [26](#). IQ (in-phase and quadrature) control of the SG384 is used for frequency and

phase control of the microwaves, which are imprinted onto the laser amplitude modulation and thus give us direct frequency and phase control over the hyperfine qubit drive.

The Raman laser illuminates the atom plane from the side in a circularly polarized elliptical beam with waists of 40 μm and 560 μm on the thin axis and the tall axis, respectively, with a total average optical power of 150 mW on the atoms. The large vertical extent ensures $<1\%$ inhomogeneity across the atoms, and shot-to-shot fluctuations in the laser intensity are also $<1\%$. For Figs. 1–3, we operate our Raman laser at a blue-detuned intermediate-state detuning of 180 GHz, resulting in two-photon Rabi frequencies of 1 MHz and an estimated scattering error per π pulse of 7×10^{-5} (that is, 1 scattering event per 15,000 π pulses)²⁶. For Fig. 4, to shorten the duration of the coherent mapping pulse sequence, we increase the Raman laser power and operate at a smaller blue-detuned intermediate-state detuning of 63 GHz, with a corresponding two-photon Rabi frequency of 3.2 MHz and an estimated scattering error per π pulse of 2×10^{-4} .

Robust single-qubit rotations

For almost all single-qubit rotations in this work (other than XY8 and XY16 self-correcting sequences), we implement robust single-qubit rotations in the form of composite pulse sequences. These composite pulse sequences are well known in the NMR community^{27,60} and can be highly insensitive to pulse errors such as amplitude or detuning miscalibrations. Our dominant source of coherent single-qubit errors arise from $\lesssim 1\%$ amplitude drifts and inhomogeneity across the array; as such, we primarily use the ‘BB1’ (broadband 1) pulse sequence, which is a sequence of four pulses that implements an arbitrary rotation on the Bloch sphere while being insensitive to amplitude errors to sixth order^{27,60}. We benchmark the performance of these robust pulses in Extended Data Fig. 3a. Furthermore, by applying a train of BB1 pulses, we find an accumulated error consistent with the estimated scattering limit (not plotted here), suggesting that the scattering limit roughly represents our single-qubit rotation infidelities (about 3×10^{-4} error per BB1 pulse owing to the increased length of the composite pulse sequence). Randomized benchmarking⁶¹ can be applied in future studies to further study single-qubit rotation fidelity.

Qubit coherence and dynamical decoupling

In our 830-nm traps, hyperfine qubit coherence is characterized by inhomogeneous dephasing time $\langle\{T\}_2\rangle^{\text{ast}} = 4 \text{ ms}$ (not plotted here), $T_2 = 1.5 \text{ s}$ (XY16 with 128 total π pulses) and relaxation time $T_1 = 4 \text{ s}$ (including atom loss) (Extended Data Fig. 3b,c). All of our experiments in this work are performed in a d.c. magnetic field of 8.5 G. Coherence can be further improved by using further-detuned optical tweezers (with trap depth held constant, the tweezer differential lightshifts decrease as $1/\Delta$ and $1/T_1$ decreases as $1/\Delta^3$ (ref. 62), where Δ is the detuning of the trap wavelength) and shielding against magnetic field fluctuations. For practical QEC operation, atom loss can be detected in a hardware-efficient manner⁴⁶ and the atom then replaced from a reservoir, which could in principle be continuously reloaded by a magneto-optical trap for reaching arbitrarily deep circuits.

All of our transport sequences^{20,58,59} are accompanied by dynamical decoupling sequences. The number of pulses we use is a trade-off between preserving qubit coherence while minimizing pulse errors. We interchange between two types of dynamical decoupling sequence: XY8 and XY16 sequences, composed of phase-alternated individual π pulses that are self-correcting for amplitude and detuning errors^{28,63}, and Carr–Purcell–Meiboom–Gill (CPMG)-type dynamical decoupling sequences composed of robust BB1 pulses. The CPMG-BB1 sequence is more robust to amplitude errors but incurs more scattering error. We empirically optimize for any given experiment by choosing between these different sequences and a variable number of decoupling π pulses, optimizing on either single-qubit coherence (including the movement) or the final signal. Typically, our decoupling sequences are composed of a total 12–18 π pulses.

Movement effects on atom heating and loss

We study here the effects of movement on atom loss and heating in the harmonic oscillator potential given by the tweezer trap. Motion of the trap potential is equivalent to the non-inertial frame of reference where the harmonic oscillator potential is stationary but the atom experiences a fictitious force given by $F(t) = -ma(t)$, where m is the mass of the particle

and $a(t)$ is the acceleration of the trap as a function of time^{64,65}. By following ref. ⁶⁶ (equation 5.4), we find the average vibrational quantum number increase ΔN is given by

$$\text{Delta } N = \frac{\tilde{a}(\omega_0) |x_{\text{zpf}}|}{\omega_0^2}, \quad (1)$$

where $\tilde{a}(\omega_0)$ is the Fourier transform of $a(t)$ evaluated at the trap frequency ω_0 , and the zero-point size of the particle $|x_{\text{zpf}}| \equiv \sqrt{\hbar/(2m\omega_0)}$, where \hbar is the reduced Planck constant. ΔN is the same for all initial levels of the oscillator⁶⁶. Experimentally, we apply an acceleration profile $a(t) = j t$ to the atom, from time $-T/2$ to $+T/2$ to move a distance D with constant jerk j . We calculate $|\tilde{a}(\omega_0)|$, simplify using $\omega_0 T \gg 1$, and assume a small range of trap frequencies to average the oscillatory terms, resulting in

$$\text{Delta } N = \frac{1}{2} \left(\frac{6D}{|x_{\text{zpf}}|} \right)^2 \omega_0^2. \quad (2)$$

Several relevant insights can be gleaned from this formula. First, this expression indicates our ability to move large distances D with comparably small increases in time T . Furthermore, to maintain a constant ΔN , the movement time $T \propto \omega_0^{-3/4}$. Moreover, to perform a large number of moves k for a deep circuit, we can estimate $\Delta N \propto k/T^4$, suggesting that we can increase our number of moves from, for example, 5 to 80 by slowing each move from 200 μs to 400 μs . Move speed could be further improved with different $a(t)$ profiles, but inevitably with finite resources such as trap depth, quantum speed limits will eventually prevent arbitrarily fast motion of qubits across the array³⁰.

We now compare equation (2) with our experimental observations. In Fig. 1d we start to observe atom loss when we move 55 μm in 200 μs under a constant negative jerk. This speed limit is consistent with our above estimates: using $\omega_0 = 2\pi \times 40 \text{ kHz}$ and $x_{\text{zpf}} = 38 \text{ nm}$, we predict $\Delta N \approx 6$ for

this move, corresponding to the onset of tangible heating at this move speed. More quantitatively, we assume a Poisson distribution with mean N and variance N and integrate the population above some critical N_{\max} upon which the atom will leave the trap. From this analysis we find atom retention is given by $\frac{1}{2}(1 + \text{erf}(\frac{N - N_{\max}}{\sqrt{2N}}))$.

Extended Data Fig. [2a, b](#) measures the atom retention as a function of move time T and trap frequency $\omega_0/2\pi$. Using the functional form above, for both sets of measurements, we extract an N_{\max} of about 30, corresponding to adding about 30 excitations before exciting the atom out of the trap. Such a limit is physically reasonable as the absolute trap depth of 4 MHz implies only about 100 levels, the atom starts at finite temperature, and moreover the effective trap frequency reduces once the anharmonicity of the trap starts to play a role. We note that these estimates are only approximate (we roughly estimate ω_0 for the trap depths used during the motion), but nonetheless suggests our motion limit is consistent with physical limits for our chosen $a(t)$. Our analysis here also neglects the acoustic lensing effects associated with ramping the AOD frequency, which causes astigmatism by focusing one axis to a different plane and thus deforms the trap and reduces the peak trap intensity (and ω_0) as given by the Strehl ratio.

Additional heating and loss during the circuit can also be caused by repeated short drops for performing two-qubit gates, where the tweezers are briefly turned off to avoid anti-trapping of the Rydberg state and light shifts of the ground–Rydberg transition. However, drop–recapture measurements in Extended Data Fig. [2c](#) suggest that the 500-ns drops we use experimentally have a negligible effect until hundreds of drops per atom (corresponding to hundreds of CZ gates). We find that atom loss and heating as a function of number of drops are well described by a diffusion model, which would then predict that reducing atom temperature by a factor of 2× (reducing thermal velocity by $\sqrt{2}$ times) and reducing the drop time t_{drop} by 2×, together would increase the number of possible CZ gates per atom to thousands.

Two-qubit CZ gates implementation

We implement our two-qubit gates and calibrations following ref. [5](#). Specifically, the two-qubit CZ gate is implemented by two global Rydberg pulses, with each pulse at detuning Δ and length τ , and with a phase jump ξ between the two pulses. The pulse parameters are chosen such that qubit pairs, adjacent and under the Rydberg blockade constraint, will return from the Rydberg state back to the hyperfine qubit manifold with a phase depending on the state of the other qubit. The numerical values for these pulse parameters are:

$$\begin{aligned} \$\$ \{\Delta\} &= -0.377371 \{\Omega\} \\ \$\$ \xi &= -0.621089 \times (2 \{\text{rm}\{\pi\}\}) \\ \$\$ \tau &= 0.683201 / [\{\Omega\} / (2 \{\text{rm}\{\pi\}\})] \end{aligned}$$

For our experiments in Figs. [1–3](#), we operate with a two-photon Rydberg Rabi frequency of $\Omega/2\pi = 3.6$ MHz, giving a theoretical $\tau = 190$ ns and a theoretical $\Delta/(2\pi) = -1.36$ MHz. We choose the negative detuning sign to help minimize excitation into the $m_j = +1/2$ Rydberg state (m_j denotes magnetic sublevel of the $70S_{1/2}$ Rydberg state), which is detuned by about 24 MHz under the field of 8.5 G (and experiences a three-times lower coupling to the Rydberg laser than the desired $m_j = -1/2$ state owing to reduced Clebsch–Gordan coefficients). In this work, we operate with strong blockade between adjacent qubits, with Rydberg–Rydberg interactions $V_0/2\pi$ ranging from 200 MHz to 1 GHz. In Fig. [4](#), we operate with $\Omega/2\pi = 4.45$ MHz for the two-qubit gates.

Managing spurious phases during CZ gates

The two-qubit gate from ref. [5](#) induces both an intrinsic single-qubit phase, as well as spurious phases that are primarily induced by the differential light shift from the 420-nm laser. Under certain configurations, the 420-nm-induced differential light shift on the hyperfine qubit can be exceedingly large (>8 MHz), yielding phase accumulations on the hyperfine qubit of about 6π . Small, percent-level variations of the 420-nm intensity can thus lead to significant qubit dephasing.

Reference [5](#) addresses this 420-induced-phase issue by performing an echo sequence: after the CZ gate, the 1,013-nm Rydberg laser is turned off, a

Raman π pulse is applied and then the 420-nm laser is pulsed again to cancel the phase induced by the 420 light during the CZ gate. This method echoes out the 420-induced phase, but comes at a cost of a factor of two increase in the 420-induced scattering error, which is the dominant source of error in our two-qubit CZ gates.

Echo between CZ gates

To address these various issues, here we perform a Raman π pulse between each CZ gate to echo out spurious gate-induced phases on the hyperfine qubit (Extended Data Fig. 1). This approach has several advantages. The 420-induced phase is now cancelled by pairs of CZ gates, without explicitly applying additional 420-nm pulses to echo each individual CZ gate, thereby reducing the scattering error of the CZ gate in this work by a factor of approximately two. We estimate that this echo technique, having reduced the scattering error incurred during each gate, roughly compensates the increased scattering rate incurred by spreading our optical power over more space in 2D, thereby giving us comparable gate fidelities to the two-qubit CZ gate fidelities of $\geq 97.4(2)\%$ reported in ref. 5. Furthermore, the echo between CZ gates also cancels the intrinsic single-qubit phase of the CZ gate, removing errors in the calibration of this parameter, as well as cancelling any other gate-induced spurious single-qubit phases such as a roughly 0.01-rad phase induced by pulsing the traps off for 500 ns for the two-qubit gate (Extended Data Fig. 1). In instances where the number of CZ gates we apply is odd, we perform the echo for the final CZ gate.

Sign of intermediate-state detuning

To further suppress the effect of the spurious, 420-induced phase, we operate our 420-nm laser to be red-detuned (by 2 GHz) from the $6P_{3/2}$ transition. For red detunings, the light shift on the $|0\rangle$ state and the $|1\rangle$ state are of the same sign, minimizing the differential light shift, whereas for blue detunings < 6.8 GHz, the light shift on the $|0\rangle$ state and the $|1\rangle$ state have opposite signs and amplify the differential light shift.

Sensitivity to axial trap oscillations

In typical Rydberg excitation timescales with optical tweezers, the axial trap oscillation frequencies of several kilohertz are inconsequential. Here with our circuits running as long as 1.2 ms, with Rydberg pulses throughout, we find that the axial trap oscillations can have important effects. In particular, the axial oscillations cause the atoms to make oscillations in and out of the Rydberg beams: at our estimated axial temperature of about 25 μK and axial oscillation frequency of 6 kHz, we estimate an axial spread $\sqrt{\langle z^2 \rangle} \approx 1 \text{ mm} \cdot \sqrt{\mu \text{ m}} \text{ m}$. For our 20- μm -waist beams, the effect of this positional spread is relatively small on the pulse parameters of the CZ gate, but can be significant on the sensitive 420-induced phase we seek to cancel by echoing out the phase induced by CZ gates separated by about 200 μs (see previous section). When using 20- μm -waist beams, and a 2.5-GHz blue detuning of our 420-nm laser, we find that the dephasing due to the axial trap oscillations is significant (Extended Data Fig. 4). To remedy this deleterious effect, we increase the beam waist of our 420-nm laser to 35 μm (while maintaining constant intensity) and change the laser frequency to be 2-GHz red-detuned, together resulting in a significant reduction in the dephasing associated with improper echoing of the 420-nm pulse.

Bell-state preparation and fidelity

In Fig. 1, we prepare the $|\Phi^+\rangle$ Bell state in the same way that is done in ref. 5. After initializing a pair of qubits in $|00\rangle$, we apply $X(\pi/2)$ pulse–CZ gate– $X(\pi/4)$ pulse. We calculate and plot the raw resulting fidelity of this $|\Phi^+\rangle$ Bell state as the sum of populations in $|00\rangle$ and $|11\rangle$, averaged with the fitted amplitude of parity oscillations (example in Fig. 1c), which measures the off-diagonal coherences. In Fig. 1d, upon significant loss from movement, this fidelity estimate becomes skewed because we begin measuring an artificially large population in $|11\rangle$ (as state $|1\rangle$ is detected as loss); accordingly, we estimate the $|\Phi^+\rangle$ population as two times the population of $|00\rangle$ once the population difference between $|11\rangle$ and $|00\rangle$ becomes greater than 0.1 (an arbitrary cut-off where the effects of loss start to become significant). In Fig. 1d, for moves slower than 300 μs , we achieve an average raw Bell-state fidelity after the moving of 94.8(2)%. If we do not move or attempt to preserve coherence for 500 μs (that is, if we measure immediately

after preparing the Bell state), then we measure a raw Bell-state fidelity of 95.2(1)% (not plotted here), consistent with the results in ref. [5](#).

Analysis of error sources

We detail here some of our measured and estimated sources of error for an entire sequence (toric code preparation in particular, our deepest circuit). We find the total single-qubit fidelity after performing the entire sequence is roughly 96.5% for the toric code circuit, which we measure by embedding the entire experiment in a Ramsey sequence: that is, we perform a Raman $\pi/2$ pulse, do all motion and decoupling, and then do a final $\pi/2$ pulse with variable phase to measure total contrast. We are able to account for our single-qubit fidelity quantitatively as being composed of our known single-qubit errors in Extended Data Fig. [6c](#).

Estimated contributions to two-qubit gate error are summarized in Extended Data Fig. [6c](#). These estimates come from numerical simulations in QuTiP (version 4.5.0) with experimental parameters. The effects of intermediate state scattering and Rydberg decay are included via collapse operators in the Lindblad master equation solver. Other error contributions include finite-temperature random Doppler shifts and position fluctuations, as well as laser pulse-to-pulse fluctuations, all of which are simulated using classical Monte Carlo sampling of experiment parameters. Experimental parameters used for the simulations are as follows: blue and red Rabi frequencies (Ω_b, Ω_r) = $2\pi \times (160, 90)$ MHz, $6P_{3/2}$ intermediate state detuning of 2 GHz, intermediate state lifetime of 110 ns, $70S_{1/2}$ Rydberg state lifetime of 150 μ s, Rydberg blockade energy of 500 MHz, splitting to second Rydberg state of 24 MHz, radial and axial trap frequencies (ω_r, ω_z) = $2\pi \times (40, 6)$ kHz, and temperature $T = 20 \mu$ K. We can also use this modelling to project for future performance; by assuming a 10 times increase in available 1,013-nm intensity and that atoms are cooled to a temperature of 2 uK, we project a possible CZ gate fidelity of 99.7%, beyond the surface code threshold^{[38,67](#)}. Alkaline-earth atoms could also offer other routes to high-fidelity operations for QEC^{[68,69,70](#)}.

To understand how our various single-qubit and two-qubit errors contribute to our graph-state fidelities, we perform a stochastic simulation of the

quantum circuit used for graph-state preparation (Extended Data Fig. 6a, b). We utilize the Clifford properties of our circuit, allowing for efficient numerical evaluation and random sampling of many possible error realizations. The simulation is performed under a realistic error model, where the rates of ambient depolarizing noise and atom loss are measured in the experiment (Extended Data Fig. 6c). The resulting stabilizer and logical qubit expectation values agree well with those measured experimentally.

Rydberg beam shaping and homogeneity

We shape our Rydberg beams into tophats of variable size through wavefront control using the phase profile on an SLM²⁵. This ability allows us to match the height of our beam profile to the experiment zone size of any given experiment, thereby maximizing our 1,013-nm light intensity and CZ gate fidelities. We optimize our Rydberg beam homogeneity until peak-to-peak inhomogeneities are below <1%. To this end, we correct all aberrations up to the window of our vacuum chamber, as done in ref. 25, which yields an inhomogeneity on the atoms of several per cent that we attribute to imperfections of the final window. To further optimize the homogeneity, we empirically tune aberration corrections on the tophat through Zernike polynomial corrections to the phase profile in the SLM plane (Fourier plane). With this procedure, we reduce peak-to-peak inhomogeneities to <1% over a range of 40–50 μm in the atom plane.

Creation and optimization of graph layouts

We outline here a description of how we optimize our graph layouts for the cluster state, Steane code, surface code and toric code preparation. Our optimization in this work is heuristic, and future work can develop appropriate algorithms for designing optimal circuits through atom spatial arrangement and AOD trajectories. Extended Data Fig. 5 shows all of the graphs we create and the process for creating them. There are several parameters we optimize for. (1) Minimize the number of parallel two-qubit gate layers. (2) Minimize the total move distance for the moving atoms. (3) Have all moving atoms in one sublattice (all graphs realized here are bipartite) to facilitate the final local rotation of one sublattice. (4) Minimize the vertical extent of the array and the number of distinct rows (to maximize

1,013-nm light intensity and minimize sensitivity to beam inhomogeneity between the rows). (5) When ordering gates, apply two-qubit gates as early as possible in the circuit. If a gate layer induces a bit-flip (X error) then that error can propagate during subsequent gates (becoming a Z error on the other qubit), so gates should be in the earliest layer possible.

Local (sublattice) hyperfine rotations

We perform local rotations in the hyperfine basis by use of our horizontally propagating 420-nm beam, which imposes a differential light of several megahertz on the hyperfine qubit and can thus be used for realizing a fast Z rotation. To realize the local $Y(\pi/2)$ rotation used throughout this work, we move one sublattice of atoms out of the 420-nm beam, then apply [global $Y(\pi/4)$]–[local $Z(\pi)$]–[global $Y(\pi/4)$]. This realizes a $Y(\pi/2)$ rotation on one sublattice and a $Z(\pi)$ rotation on the other sublattice (which is inconsequential as it then commutes with the immediately following measurement in the Z basis). To apply a $Y(\pi/2)$ on the other sublattice of atoms, we add an additional global $Z(\pi)$ (implemented by jumping the Raman laser phase) between the two $Y(\pi/4)$ pulses. Future experiments will benefit from an additional set of locally focused beams for performing local Raman control of hyperfine qubit states, but we find that moving atoms works so efficiently (even for moving $>50\ \mu\text{m}$ to move out of the 420-nm beam) that this approach is well suited for producing a high-fidelity, homogeneous rotation on roughly half the qubits.

Local Rydberg initialization

We perform local Rydberg control to initialize the $|\mathbb{Z}_2\rangle = |rgrg\dots\rangle \equiv |r1r1\dots\rangle$ state for studying the dynamics of many-body scars. We achieve this local initialization by applying approximately 50-MHz light shifts between $|1\rangle$ and $|r\rangle$ using 810-nm tweezers generated by an SLM onto a desired subset of sites, and then apply a global Rydberg π pulse, which excites the non-light-shifted atoms. We use this approach here to prepare every other atom in each chain into $|r\rangle$, but emphasize that as the locations of the SLM tweezers are fully programmable, this technique can be used to prepare any initial blockade-satisfying configuration of atoms in $|1\rangle$ and $|r\rangle$.

The 50-MHz biasing light shift is significantly larger than the Rydberg Rabi frequency $\Omega/2\pi = 4.45$ MHz, leading to a Rydberg population on undesired sites of <1%. The $t = 0$ time point of Extended Data Fig. 10b shows the high-fidelity preparation of the $|\mathbb{Z}_2\rangle$ state using this approach. We note that with 810-nm light, even though the achieved biasing light shift is significant, the Raman-scattering-induced T_1 (of the hyperfine qubit) is still about 50 ms and thus leads to a scattering error $\leq 4 \times 10^{-6}$ during the 200-ns pulse of the light-shifting tweezers. There can also be a motional effect from the biasing tweezers, with an estimated radial trapping frequency of 150 kHz, which we also deem to be negligible during the 200-ns pulse.

Rydberg Hamiltonian

In Fig. 4, we study dynamics under the many-body Rydberg Hamiltonian

$$\frac{\hbar}{\Omega} \sum_i (\sigma_i^x - \Delta) n_i + \sum_{i < j} V_{ij} n_i n_j \quad (3)$$

where \hbar is the reduced Planck constant, Ω is the Rabi frequency, Δ is the laser frequency detuning, $n_i = |r_i\rangle\langle r_i|$ is the projector onto the Rydberg state at site i and $(\sigma_i^x)^2 = |1\rangle\langle r_i| + |r_i\rangle\langle 1|$ flips the atomic state. For the entanglement entropy measurements in this work, we choose lattice spacings where the nearest-neighbour interaction $V_0 > \Omega$ results in the Rydberg blockade, preventing adjacent atoms from simultaneously occupying $|r\rangle$. In particular, the many-body experiments are performed on eight-atom chains, quenching to a time-independent H_{Ryd} with $V_0/2\pi = 20$ MHz, $\Omega/2\pi = 3.1$ MHz and $\Delta/2\pi = 0.3$ MHz. Quenching to small, positive $\Delta = 0.0173V_0$ partially suppresses the always-positive long-range interactions and thereby is optimal for scar lifetime, as derived and shown experimentally in ref. 71.

Coherent mapping protocol

As described in the text, we implement a coherent mapping protocol to transfer a generic many-body state in the $\{|1\rangle, |r\rangle\}$ basis to the long-lived and non-interacting $\{|0\rangle, |1\rangle\}$ basis. To achieve this mapping, immediately following the Rydberg dynamics we apply a Raman π pulse to map $|1\rangle \rightarrow |0\rangle$, and then a subsequent Rydberg π pulse to map $|r\rangle \rightarrow |1\rangle$ (ref. [72](#)).

Even for perfect Raman and Rydberg π pulses (on isolated atoms), there are three key sources of infidelity associated with this mapping process. (1) Any population in blockade-violating states (that is, two adjacent atoms both in $|r\rangle$) will be strongly shifted off-resonance for the final Rydberg π pulse. As such, this atomic population will be left in the Rydberg state and lost. (2) Long-range interactions, for example, from next-nearest neighbours, will detune the final Rydberg π pulse from resonance and thus reduce pulse fidelity. As the long-range interactions are not the same for all many-body microstates, this effect cannot be mitigated by a simple shift of the detuning. (3) Dephasing of the state occurs throughout the duration of the Raman π pulse, predominantly from Doppler shifts between the ground states $|0\rangle$ and $|1\rangle$ and the Rydberg state $|r\rangle$. Although these random on-site detunings are also present during the many-body dynamics, turning the Rydberg drive Ω off allows the system to freely accumulate phase and makes us particularly sensitive to dephasing errors.

We now detail our mitigation of the above error mechanisms. To minimize errors from (1), we perform our many-body dynamics with $\sqrt{\Omega^2/(2V_0^2)} \approx 0.01$. This minimizes the probability of an atom to violate blockade to be of order 1%. To help minimize errors from (2), we increase the amplitude of the 420-nm laser for the final π pulse by a factor of 2 \times , such that $(V_{\text{NNN}}/\Omega)^2 = 0.005$ (where V_{NNN} is the interactions with next-nearest neighbours), reducing pulse errors from long-range interactions to order 1%. Finally, to reduce errors from (3), we perform a fast Raman π pulse and leave only 150 ns between ending the many-body Rydberg dynamics and beginning the Rydberg π pulse. The 150-ns gap is comparably short relative to the $T_2^{\text{last}} \approx 3-4 \mu\text{s}$ of the $\{|g\rangle, |r\rangle\}$ basis, leading to a random phase accumulation of the order of about $0.02 \times 2\pi$ rad per particle, but is further compounded by having entangled states of N particles in one copy accumulating a random phase relative to entangled

states of N particles in the second copy. We study these various effects numerically in Extended Data Fig. 9c.

Finally, we note that the global Raman beam induces a light-shift-induced phase shift of about π on $|0\rangle$ and $|1\rangle$ relative to $|r\rangle$ during the Raman π pulse. Similarly, the global 420-nm laser also induces a light-shift-induced phase shift of about π between $|0\rangle$ and $|1\rangle$ during the Rydberg π pulse. Although the measurements we perform here are interferometric (in other words, the singlet state we measure is invariant under global rotations) and thus not affected by these global phase shifts, in future work these phase shifts can be measured and accounted for where relevant.

Measuring entanglement entropy

The second-order Renyi entanglement entropy is given by $\langle\langle S \rangle\rangle_2 = -\log \text{Tr}[\rho_A^2]$, where $\text{Tr}[\rho_A^2]$ is the state purity of reduced density matrix ρ_A on subsystem A. The purity can be measured with two copies by noticing that $\text{Tr}[\rho_A^2] = \text{Tr}[\hat{\text{SWAP}}]$ is the expectation value of the many-body SWAP operator $\hat{\text{SWAP}}$ (ref. 10,11). The many-body SWAP operator is composed of individual SWAP operators $\hat{\text{SWAP}}_i$ on each twin pair, that is, $\hat{\text{SWAP}} = \prod_i \hat{\text{SWAP}}_i$ (with $i \in A$). Measuring this expectation value amounts to probing occurrences of the singlet state $|01\rangle - |10\rangle / \sqrt{2}$ (with eigenvalue -1 under $\hat{\text{SWAP}}_i$), as all other $\hat{\text{SWAP}}_i$ eigenstates have eigenvalue $+1$. Occurrences of the singlet state in each twin pair, that is, the Bell state $|\Psi^-\rangle$, is extracted by a Bell measurement circuit (with an additional local $Z(\pi)$, see next paragraph), which maps $|\Psi^-\rangle \rightarrow |00\rangle$ and can thereafter be measured in the computational basis. As such, after performing the Bell measurement circuit, we analyse the resulting bit-string outputs and calculate the purity of any subsystem A by calculating $\langle\langle \text{Pi}_i \rangle\rangle$: that is, we measure purity as the average parity = $\langle (-1)^{\text{observed } |00\rangle \text{ pairs}} \rangle$ within A. In the absence of experimental imperfections, the purity will equal 1 for the whole

system, and be less than 1 for subsystems that are entangled with the rest of the system.

A Bell measurement circuit can be decomposed into applying an $X(\pi/2)$ rotation on one atom of the twin pair, then applying a CZ gate and then a global $X(\pi/2)$ rotation. In other measurements, we realize a local $X(\pi/2)$ by doing a global $X(\pi/4)$ rotation, then local $Z(\pi)$ rotation and then global $X(\pi/4)$. However, we note that for this singlet measurement circuit, the first $X(\pi/4)$ is redundant as the singlet state is invariant under global rotations, and so for the local $X(\pi/2)$ we only apply the local $Z(\pi)$ and then the second global $X(\pi/4)$. This effectively realizes the $X(\pi/2)$ on one qubit, up to a $Z(\pi)$ on the other qubit (not shown in circuit diagram in Fig. 4). Under this simplification, the Bell measurement circuit to map $|\Psi^-\rangle \rightarrow |00\rangle$ can be roughly understood as the reverse of the Bell-state preparation circuit in ref. 5, which is precisely how we calibrate the parameters of the Bell measurement.

Calibrating and benchmarking the interferometry

To validate the interferometry measurement (and check for proper calibration), we benchmark it separately from the many-body dynamics and coherent mapping protocol. We perform this benchmarking by preparing independent qubits in identical, variable single-qubit superpositions (through a global Raman pulse of variable time) and ensuring that the interferometry rarely results in $|00\rangle$ for all the variable initial product states (Extended Data Fig. 8a). We find this is an important benchmarking step, because we find that small miscalibrations of the Bell measurement can lead to lower fidelity (that is, higher entropy) for different initial product states and thereby result in additional spurious signals in an entanglement entropy measurement. We note that this measurement is particularly sensitive to the single-qubit phase immediately before the final $X(\pi/2)$ pulse (induced by the CZ gate and cancelled by a global $Z(\theta)$ pulse).

Additional many-body data and details

To benchmark our method of measuring entanglement entropy in a many-body system, in Extended Data Fig. 8b we study the entanglement dynamics

after initializing two proximal atoms in $|1\rangle$ and resonantly exciting to the Rydberg state for a variable time t . Under conditions of Rydberg blockade, this excitation results in two-particle Rabi oscillations between $|11\rangle$ and the entangled state $\frac{1}{\sqrt{2}}(|1r\rangle + |r1\rangle)$ (top panel of Extended Data Fig. 8b)^{3,13,72}. The state purity of this two-particle system is measured by performing Bell measurements on atom pairs from two identical copies. Locally, the measured purity of the one-particle subsystem reduces to a value of about 0.5 when the system enters the maximally entangled $|W\rangle$ state, at which point the reduced density matrix of each individual atom is maximally mixed. In contrast, the purity of the global, two-particle state remains high. The observation that the global-state purity is higher than the local-subsystem purity is a distinct signature of quantum entanglement^{11,12}.

For the data shown in Fig. 4c,e, we subtract the data by an extensive classical entropy as is done in ref. 12. This fixed, time-independent offset is given by the entropy per particle, that is, (global entropy at quench time $t=0$) \times (subsystem size)/(global system size). In Extended Data Fig. 9a, we show the raw entanglement entropy measurements alongside numerics, to indicate the size of the extensive classical entropy contribution. In plotting, we also delay the theory curves by 10 ns to account for the fact that the Raman π pulse cuts off the final 10 ns of the Rydberg evolution, which is done to keep the coherent mapping gap as short as possible and minimize Doppler dephasing. Furthermore, in Extended Data Fig. 9b we plot the measured global purity and compare it with numerical simulations incorporating experimental errors (Extended Data Fig. 9c).

In Extended Data Fig. 10, we show additional many-body data on the eight-atom chain system, with the same parameters as those used in the main text. We show the measured single-site entropy of each site in the eight-atom chain for the $|\mathbb{Z}_2\rangle$ quench in Extended Data Fig. 10a. Furthermore, in Extended Data Fig. 10b, we plot the global Rydberg population, measured in both the $\{|1\rangle, |r\rangle\}$ basis and the $\{|0\rangle, |1\rangle\}$ basis.

Data availability

The data that supports the findings of this study are available from the corresponding author on reasonable request.

References

1. Cirac, J. I. & Zoller, P. A scalable quantum computer with ions in an array of microtraps. *Nature* **404**, 579–581 (2000).
2. Preskill, J. Quantum Computing in the NISQ era and beyond. *Quantum* **2**, 79 (2018).
3. Jaksch, D. et al. Fast quantum gates for neutral atoms. *Phys. Rev. Lett.* **85**, 2208–2211 (2000).
4. Urban, E. et al. Observation of Rydberg blockade between two atoms. *Nat. Phys.* **5**, 110–114 (2009).
5. Levine, H. et al. Parallel implementation of high-fidelity multiqubit gates with neutral atoms. *Phys. Rev. Lett.* **123**, 170503 (2019).
6. Nigg, D. et al. Quantum computations on a topologically encoded qubit. *Science* **345**, 302–305 (2014).
7. Ryan-Anderson, C. et al. Realization of real-time fault-tolerant quantum error correction. *Phys. Rev.* **11**, 041058 (2021).
8. Satzinger, K. J. et al. Realizing topologically ordered states on a quantum processor. *Science* **374**, 1237–1241 (2021).
9. Kitaev, A. Y. Fault-tolerant quantum computation by anyons. *Ann. Phys.* **303**, 2–30 (2003).
10. Daley, A. J., Pichler, H., Schachenmayer, J. & Zoller, P. Measuring entanglement growth in quench dynamics of bosons in an optical

lattice. *Phys. Rev. Lett.* **109**, 020505 (2012).

11. Islam, R. et al. Measuring entanglement entropy in a quantum many-body system. *Nature* **528**, 77–83 (2015).
12. Kaufman, A. M. et al. Quantum thermalization through entanglement in an isolated many-body system. *Science* **353**, 794–800 (2016).
13. Bernien, H. et al. Probing many-body dynamics on a 51-atom quantum simulator. *Nature* **551**, 579–584 (2017).
14. Turner, C. J., Michailidis, A. A., Abanin, D. A., Serbyn, M. & Papić, Z. Weak ergodicity breaking from quantum many-body scars. *Nat. Phys.* **14**, 745–749 (2018).
15. Molmer, K. & Sørensen, A. Multiparticle entanglement of hot trapped ions. *Phys. Rev. Lett.* **82**, 1835 (1999).
16. Wright, K. et al. Benchmarking an 11-qubit quantum computer. *Nat. Commun.* **10**, 5464 (2019).
17. Periwal, A. et al. Programmable interactions and emergent geometry in an array of atom clouds. *Nature* **600**, 630–635 (2021).
18. Majer, J. et al. Coupling superconducting qubits via a cavity bus. *Nature* **449**, 443–447 (2007).
19. Mandel, O. et al. Coherent transport of neutral atoms in spin-dependent optical lattice potentials. *Phys. Rev. Lett.* **91**, 010407 (2003).
20. Beugnon, J. et al. Two-dimensional transport and transfer of a single atomic qubit in optical tweezers. *Nat. Phys.* **3**, 696–699 (2007).
21. Monroe, C. et al. Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. *Phys. Rev. A* **89**, 022317 (2014).
22. Wu, T.-Y., Kumar, A., Giraldo, F. & Weiss, D. S. Stern–Gerlach detection of neutral-atom qubits in a state-dependent optical lattice.

Nat. Phys. **15**, 538–542 (2019).

23. Pompili, M. et al. Realization of a multinode quantum network of remote solid-state qubits. *Science* **372**, 259–264 (2021).
24. Pino, J. M. et al. Demonstration of the trapped-ion quantum CCD computer architecture. *Nature* **592**, 209–213 (2021).
25. Ebadi, S. et al. Quantum phases of matter on a 256-atom programmable quantum simulator. *Nature* **595**, 227–232 (2021).
26. Levine, H. et al. Dispersive optics for scalable Raman driving of hyperfine qubits. *Phys. Rev. A* **105**, 032618 (2022).
27. Vandersypen, L. M. K. & Chuang, I. L. NMR techniques for quantum control and computation. *Rev. Mod. Phys.* **76**, 1037 (2005).
28. Gullion, T., Baker, D. B. & Conradi, M. S. New, compensated Carr–Purcell sequences. *J. Magn. Reson.* **89**, 479–484 (1990).
29. Dordević, T. et al. Entanglement transport and a nanophotonic interface for atoms in optical tweezers. *Science* **373**, 1511–1514 (2021).
30. Lam, M. R. et al. Demonstration of quantum brachistochrones between distant states of an atom. *Phys. Rev. X* **11**, 011035 (2021).
31. Hein, M. et al. Entanglement in graph states and its applications. In *Proc. International School of Physics “Enrico Fermi”* (eds Casati, G. et al.) Vol. 162, 115–218 (IOS Press, 2006).
32. Raussendorf, R. & Briegel, H. J. A one-way quantum computer. *Phys. Rev. Lett.* **86**, 5188 (2001).
33. Schwartz, I. et al. Deterministic generation of a cluster state of entangled photons. *Science* **354**, 434–437 (2016).
34. Tóth, G. & Gühne, O. Entanglement detection in the stabilizer formalism. *Phys. Rev. A* **72**, 022340 (2005).

35. Raussendorf, R. & Harrington, J. Fault-tolerant quantum computation with high threshold in two dimensions. *Phys. Rev. Lett.* **98**, 190504 (2007).
36. Gottesman, D. An introduction to quantum error correction and fault-tolerant quantum computation. Preprint at <https://arxiv.org/abs/0904.2557> (2009).
37. Egan, L. et al. Fault-tolerant control of an error-corrected qubit. *Nature* **598**, 281–286 (2021).
38. Fowler, A. G., Mariantoni, M., Martinis, J. M. & Cleland, A. N. Surface codes: towards practical large-scale quantum computation. *Phys. Rev. A* **86**, 032324 (2012).
39. Semeghini, G. et al. Probing topological spin liquids on a programmable quantum simulator. *Science* **374**, 1242–1247 (2021).
40. Bravyi, S., Hastings, M. B. & Verstraete, F. Lieb–Robinson bounds and the generation of correlations and topological quantum order. *Phys. Rev. Lett.* **97**, 050401 (2006).
41. Bolt, A., Duclos-Cianci, G., Poulin, D. & Stace, T. Foliated quantum error-correcting codes. *Phys. Rev. Lett.* **117**, 070501 (2016).
42. Wilk, T. et al. Entanglement of two individual neutral atoms using Rydberg blockade. *Phys. Rev. Lett.* **104**, 010502 (2010).
43. Lesanovsky, I. & Katsura, H. Interacting Fibonacci anyons in a Rydberg gas. *Phys. Rev. A* **86**, 041601 (2012).
44. Kaufman, A. M., Lester, B. J. & Regal, C. A. Cooling a single atom in an optical tweezer to its quantum ground state. *Phys. Rev.* **2**, 041014 (2012).
45. Shea, M. E., Baker, P. M., Joseph, J. A., Kim, J. & Gauthier, D. J. Submillisecond, nondestructive, time-resolved quantum-state readout of a single, trapped neutral atom. *Phys. Rev. A* **102**, 053101 (2020).

46. Cong, I., Levine, H., Keesling, A., Bluvstein, D., Wang, S.-T. & Lukin, M. D. Hardware-efficient, fault-tolerant quantum computation with Rydberg atoms. Preprint at <https://arxiv.org/abs/2105.13501> (2021).
47. Wang, C., Harrington, J. & Preskill, J. Confinement-Higgs transition in a disordered gauge theory and the accuracy threshold for quantum memory. *Ann. Phys.* **303**, 31–58 (2003).
48. Bombín, H. Gauge color codes: optimal transversal gates and gauge fixing in topological stabilizer codes. *New J. Phys.* **17**, 083002 (2015).
49. Breuckmann, N. P. & Eberhardt, J. N. Quantum low-density parity-check codes. *PRX Quantum* **2**, 040101 (2021).
50. Pichler, H., Zhu, G., Seif, A., Zoller, P. & Hafezi, M. Measurement protocol for the entanglement spectrum of cold atoms. *Phys. Rev.* **6**, 041033 (2016).
51. Brown, A. R. et al. Quantum gravity in the lab: teleportation by size and traversable wormholes. Preprint at <https://arxiv.org/abs/1911.06314> (2019).
52. Koczor, B. Exponential error suppression for near-term quantum devices. *Phys. Rev.* **11**, 031057 (2021).
53. Hashizume, T., Bentsen, G. S., Weber, S. & Daley, A. J. Deterministic fast scrambling with neutral atom arrays. *Phys. Rev. Lett.* **126**, 200603 (2021).
54. Bothwell, T. et al. Resolving the gravitational redshift within a millimeter atomic sample. *Nature* **602**, 420–424 (2022).
55. Fowler, A. G. et al. Surface code quantum communication. *Phys. Rev. Lett.* **104**, 180503 (2010).
56. Zheng, X. et al. High precision differential clock comparisons with a multiplexed optical lattice clock. *Nature* **602**, 425–430 (2022).

57. Labuhn, H. et al. Tunable two-dimensional arrays of single Rydberg atoms for realizing quantum Ising models. *Nature* **534**, 667–670 (2016).
58. Lengwenus, A., Kruse, J., Schlosser, M., Tichelmann, S. & Birkl, G. Coherent transport of atomic quantum states in a scalable shift register. *Phys. Rev. Lett.* **105**, 170502 (2010).
59. Yang, J. et al. Coherence preservation of a single neutral atom qubit transferred between magic-intensity optical traps. *Phys. Rev. Lett.* **117**, 123201 (2016).
60. Wimperis, S. Broadband, narrowband, and passband composite pulses for use in advanced NMR experiments. *J. Magn. Reson. A* **109**, 221–231 (1994).
61. Xia, T. et al. Randomized benchmarking of single-qubit gates in a 2D array of neutral-atom qubits. *Phys. Rev. Lett.* **114**, 100503 (2015).
62. Ozeri, R. et al. Errors in trapped-ion quantum gates due to spontaneous photon scattering. *Phys. Rev. A* **75**, 042329 (2007).
63. Souza, A. M., Lvarez, G. A. & Suter, D. Robust dynamical decoupling. *Phil. Trans. R. Soc. A* **370**, 4748–4769 (2012).
64. Couvert, A., Kawalec, T., Reinaudi, G. & Guéry-Odelin, D. Optimal transport of ultracold atoms in the non-adiabatic regime. *Europhys. Lett.* **83**, 13001 (2008).
65. Reichle, R. et al. Transport dynamics of single ions in segmented microstructured Paul trap arrays. *Fortschr. Phys.* **54**, 666–685 (2006).
66. Carruthers, P. & Nieto, M. M. Coherent states and the forced quantum oscillator. *Am. J. Phys.* **33**, 537 (1965).
67. Chen, Z. et al. Exponential suppression of bit or phase errors with cyclic error correction. *Nature* **595**, 383–387 (2021).

68. Wilson, J. T. et al. Trapping alkaline earth rydberg atoms optical tweezer arrays. *Phys. Rev. Lett.* **128**, 033201 (2022).
69. Madjarov, I. S. et al. High-fidelity entanglement and detection of alkaline-earth Rydberg atoms. *Nat. Phys.* **16**, 857–861 (2020).
70. Schine, N., Young, A. W., Eckner, W. J., Martin, M. J. & Kaufman, A. M. Long-lived Bell states in an array of optical clock qubits. Preprint at <https://arxiv.org/abs/2111.14653> (2021).
71. Bluvstein, D. et al. Controlling quantum many-body dynamics in driven Rydberg atom arrays. *Science* **371**, 1355–1359 (2021).
72. Picken, C. J., Legaie, R., McDonnell, K. & Pritchard, J. D. Entanglement of neutral-atom qubits with long ground-Rydberg coherence times. *Quantum Sci. Technol.* **4**, 015011 (2018).
73. Steane, A. Multiple-particle interference and quantum error correction. *Proc. R. Soc. Lond. A* **452**, 2551–2577 (1996).
74. Erhard, A. et al. Entangling logical qubits with lattice surgery. *Nature* **589**, 220–224 (2021).
75. Ho, W. W., Choi, S., Pichler, H. & Lukin, M. D. Periodic orbits, entanglement, and quantum many-body scars in constrained models: matrix product state approach. *Phys. Rev. Lett.* **122**, 040603 (2019).

Acknowledgements

We thank I. Cong and X. Gao for their insight into QEC and for pointing us to the connections between graph states and QEC codes; M. Abobeih, H. Bernien, M. Cain, S. Choi, M. Endres, S. Evered, W. W. Ho, T. Manovitz, A. Omran, M. Serbyn and H. Zhou for reading of the manuscript and helpful discussions. We acknowledge financial support from the Center for Ultracold Atoms, the National Science Foundation, the Vannevar Bush Faculty Fellowship, the US Department of Energy (DE-SC0021013 and DOE Quantum Systems Accelerator Center, contract number 7568717), the

Office of Naval Research, the Army Research Office MURI (grant number W911NF-20-1-0082) and the DARPA ONISQ programme (grant number W911NF2010021). D.B. acknowledges support from the NSF Graduate Research Fellowship Program (grant DGE1745303) and The Fannie and John Hertz Foundation. H.L. acknowledges support from the National Defense Science and Engineering Graduate (NDSEG) fellowship. G.S. acknowledges support from a fellowship from the Max Planck/Harvard Research Center for Quantum Optics. N.M. acknowledges support by the Department of Energy Computational Science Graduate Fellowship under award number DE-SC0021110. H.P. acknowledges support by the Army Research Office (grant number W911NF-21-1-0367)

Author information

Author notes

1. Harry Levine

Present address: AWS Center for Quantum Computing, Pasadena, CA, USA

Affiliations

1. Department of Physics, Harvard University, Cambridge, MA, USA

Dolev Bluvstein, Harry Levine, Giulia Semeghini, Tout T. Wang, Sepehr Ebadi, Marcin Kalinowski, Alexander Keesling, Nishad Maskara, Markus Greiner & Mikhail D. Lukin

2. QuEra Computing Inc., Boston, MA, USA

Alexander Keesling

3. Institute for Theoretical Physics, University of Innsbruck, Innsbruck, Austria

Hannes Pichler

4. Institute for Quantum Optics and Quantum Information, Austrian Academy of Sciences, Innsbruck, Austria

Hannes Pichler

5. Department of Physics and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA

Vladan Vuletić

Contributions

D.B., H.L., G.S., T.T.W., S.E., M.K. and A.K. contributed to building the experimental setup, performed the measurements and analysed the data. M.K., N.M. and H.P. performed theoretical analysis. All work was supervised by M.G., V.V. and M.D.L. All authors discussed the results and contributed to the manuscript.

Corresponding author

Correspondence to [Mikhail D. Lukin](#).

Ethics declarations

Competing interests

M.G., V.V. and M.D.L. are co-founders and shareholders of QuEra Computing. A.K. is an executive at and shareholder of QuEra Computing. All other authors declare no competing interests.

Peer review

Peer review information

Nature thanks Hannah Williams and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. [Peer reviewer reports](#) are

available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 CZ gate echo, atomic level structure, and typical pulse sequence.

a, The two-qubit gates we apply, in addition to applying a controlled-Z operation between the two qubits, also induce a single-qubit phase $Z(\zeta)$ to both qubits, composed of the intrinsic phase of the CZ gate⁵ and additional spurious phases from the 420-nm Rydberg laser and pulsing the traps off. Since we apply all gates in parallel by global pulses of the Rydberg laser, if a qubit is not adjacent to another qubit, it does not perform a CZ gate but still acquires the same $Z(\zeta)$ (identical to being adjacent to another qubit in state $|0\rangle$, which is dark to the Rydberg laser). As diagrammed, we cancel the additional, undesired $Z(\zeta)$ by applying a π pulse between pairs of CZ gates. This echo procedure removes any need to calibrate the intrinsic phase from the CZ gate, and renders us insensitive to any spurious changes in $Z(\zeta)$ slower than $\sim 200\ \mu\text{s}$. The additional $Y(\pi)$ propagates in a known way through the CZ gates and multiplies certain stabilizers by a -1 sign, which simply redefines the sign of stabilizers and logical qubits. **b**, Level diagram showing key ^{87}Rb atomic levels used. Our Rydberg excitation scheme from $|1\rangle$ to $|r\rangle$ is composed of a two-photon transition driven by a 420-nm laser and a 1013-nm laser (see ref. ²⁵ for description of laser system). A DC magnetic field of $B = 8.5\ \text{G}$ is applied throughout this work. **c**, A typical pulse sequence for running a quantum circuit.

Extended Data Fig. 2 Movement characterization and multiple drop-recaptures.

a, Atom retention as a function of average separation speed $2D/T$ (as is plotted in Fig. 1d of the main text for separating Bell pairs), with subtracted background loss of 0.7%. The inset in Fig. 1d of the main text is normalized by (Atom retention)² (without subtracting background loss). Dark curve is calculated using experimental parameters and Eq. 2, matched to the experimental data by setting $N_{\max} = 26$ and averaging over a range of $\omega_0/2\pi$ of $\pm 15\%$ around an average $\omega_0/2\pi = 40$ kHz. **b**, Atom retention as a function of inverse trap frequency ($2\pi/\omega_0$) after the four moves of the surface code circuit. For calculating the atom loss here we set $N_{\max} = 33$ and average the trap frequencies over a range of $\pm 15\%$. We note that these quantitative estimates are sensitive to ω_0 which we roughly estimate. **c**, Atom loss as a function of drop time and number of drop loops, with $100 \mu\text{s}$ wait between each drop. When running quantum circuits we use 500-ns drops for each CZ gate (to avoid anti-trapping of the Rydberg state and light shifts of the transition), for which we observe here that hundreds of drops can be made (corresponding to hundreds of possible CZ gates per atom) before atom loss becomes significant. **d**, By rescaling the x-axis of the data to $\langle \{t\}_{\{\{\text{rm}\{\text{drop}\}\}}}\rangle \sqrt{N} \rangle$, we find the data of the various t_{drop} collapse onto a universal curve, suggesting a diffusion model for explaining the atom loss after a certain number of drops. By modeling such a diffusion process analytically we obtain the black curve with a temperature of $10 \mu\text{K}$ and a trapping radius of $1 \mu\text{m}$.

Extended Data Fig. 3 Robust single-qubit control and qubit coherence.

a, Robust BB1 single-qubit rotation in comparison to a normal single-qubit rotation, as a function of pulse area error. An arbitrary BB1(θ, ϕ) rotation on the Bloch sphere of angle θ about axis ϕ is realized with a sequence of four pulses: $(\pi)_{\phi + \frac{\theta}{4}}(2\pi)_{3\phi + \frac{\theta}{2}}(\pi)_{\phi + \frac{\theta}{4}}(\theta)_\phi$, where $\phi = \cos^{-1}(-\theta/4\pi)$ ⁶⁰. Pulse fidelity is measured here for a π pulse, defined such that the fidelity is the probability of successful transfer from $|0\rangle \rightarrow |1\rangle$, including SPAM correction. **b**, Preserving hyperfine qubit coherence using dynamical decoupling (XY16 with 128 total π pulses). Qubit coherence is observed on a timescale of seconds, with a fitted coherence time $T_2 = 1.49(8)\text{s}$. Data is

measured with either a $+\pi/2$ or $-\pi/2$ pulse at the end of the sequence, and these curves are then subtracted to yield the coherence y-axis. **c**, Hyperfine qubit T_1 , measured by the difference of final $F=2$ populations between measurements starting in $|F=2, m_F=0\rangle$ and $|F=1, m_F=0\rangle$. Atom loss without cooling is separately measured (predominantly arising from vacuum loss) and normalized to also measure the intrinsic spin relaxation time $\langle\{T\}_1'\rangle$ in the absence of atom loss. All data here is measured in 830-nm traps.

Extended Data Fig. 4 Effect of axial trap oscillations on echo fidelity of 420-nm Rydberg pulse.

a, Noise correlation measurement of the 420-nm Rydberg laser pulse intensity. In the blue-detuned configuration used in this figure only, the 420-nm laser induces an 8 MHz differential light shift on the hyperfine qubit, and consequently a phase accumulation of 32π during a 2- μ s pulse (our CZ gates are 400-ns total). Small fluctuations of the 420-nm laser intensity lead to large fluctuations in phase accumulation of the hyperfine qubit, and thus cause significant dephasing. The echo sequence diagrammed here probes the correlation of the accumulated phase between two 420-nm pulses separated by a variable time τ , and thus informs how far-separated in time the 420-nm pulses can be while still properly echoing out fluctuations in the 420-nm intensity. **b**, Hyperfine coherence (a proxy for echo fidelity) versus gap time τ between the two 420-nm pulses. The echo fidelity decreases initially due to a decorrelation of the 420-nm intensity, but then increases again, showing that the correlation of the 420-nm intensity is non-monotonic. The decaying oscillations are fit to a functional form of $y = y_0 + A \cos^2(\pi f \tau) \exp[-(\tau/T)^2]$. **c**, The fitted oscillation frequency f of the correlation / decorrelation of the noise follows a square-root relationship with the trap power, and is consistent with the expected axial trap oscillation frequency. These observations indicate that a significant portion of the correlation / decorrelation of the 420-nm noise arises from the several- μ m axial oscillations of the atom in the trap. For this measurement, we intentionally displace the 420-nm beam by several μ m in order to place the atom on a slope of the beam, increasing our sensitivity to this phenomenon. For the other experiments in our work, we minimize

sensitivity to these effects by operating in the center of a larger (35-micron-waist) 420-nm beam and operating red-detuned of the intermediate-state transition.

Extended Data Fig. 5 Movement schematics.

Schematics showing the gate-by-gate creation of **a** the 1D cluster state, **b** the Steane code, **c** the surface code, and **d** the toric code (see also [Supplementary Video 1](#)), in a side-by-side comparison. These various graph states are all generated in the same way, and encoding a desired circuit is a matter of positioning the atoms in different initial positions and applying an appropriate AOD waveform. To realize a desired circuit, atom layouts and trajectories are optimized heuristically in the way described in the Methods text. Panel c also shows the definition of surface code stabilizers as ordered in the main text.

Extended Data Fig. 6 Error simulations and tabulated single-qubit and two-qubit error estimates.

We compare our measured graph state fidelities to those from a stochastic Monte Carlo simulation of our system for **a**, the surface code and **b**, the toric code. We find that the simulated stabilizers agree well with the experimental data for this empirical depolarizing noise model. In addition, for the surface code (toric code) in the experiment we find 35% (20%) of measurements detect no stabilizer errors, compared to 40% (26%) in the simulation. We assume two-qubit errors are described by rates of 0.2% Y error, 0.2% X error, 0.5% Z error, and 0.5% loss per qubit per parallel layer (4 layers for surface code, 5 layers for toric code), corresponding to a 97.2% CZ-gate fidelity. We also add ambient, single-qubit errors at a rate of 0.1% Y error, 0.1% X error, 0.4% Z error, and 0.2% loss per qubit per parallel layer, as well as an initial 1% loss before the circuit begins (empirically factoring in SPAM errors). **c**, Tabulation of single-qubit (SQ) and two-qubit (TQ) gate errors that are measured, estimated, and extrapolated. Simulated TQ fidelities include the 0.6% scattering error from the 420-nm echo pulse. The estimated TQ fidelities are given for the experiments of the surface code and toric code, but is an underestimate of the TQ fidelities for the cluster state and Steane code measurements where

we increase the 1013-nm intensity by $2\times$ and reduce the 420-nm intensity by $2\times$, increasing gate fidelity. The Bell state estimate of CZ gate fidelity is similarly done with $2\times$ higher 1013 intensity, but includes the 420-nm echo pulse, and consequently yields a similar gate fidelity as the surface and toric code estimates.

Extended Data Fig. 7 Properties of encoded logical states.

a, Summary of logical error probabilities for the various error correcting graphs made in this work (all in logical state $|+\rangle_L$), for raw measurements as well as implementing error correction and error detection in postprocessing. Error correction for the Steane code is implemented with the Steane code decoder^{36,73} and is implemented with the minimum-weight-perfect-matching algorithm for the surface and toric codes³⁸. For the even-distance toric code, when correction is ambiguous we do not flip the logical qubit, and accordingly the distance $d = 2$ logical qubit does not change under the correction procedure. We remark that the observed fidelities are comparable to similar demonstrations in state-of-the-art experiments with other platforms^{8,74}. These will need to be improved to surpass the threshold for practical error correction³⁸ (see [Methods](#) text). **b**, Lifetime of the logical $|+\rangle_L$ state on the surface code, with correction and detection performed in postprocessing as in **a**. After state preparation, the $|+\rangle_L$ state is held for a variable time before projective measurement, with two π pulses applied for dynamical decoupling (lifetime can be extended significantly further by applying e.g. 128 π pulses as done in Extended Data Fig. [3b](#)). Some experimental parameters are slightly different here compared to those in **a**, hence the higher error rates here at the time 0 point. **c**, Logical $\pi/2$ rotation on the Steane code to prepare logical qubit state $|0_L\rangle$. The Steane code, surface code, and toric code all have transversal single-qubit Clifford operations on the logical qubit^{8,36} (including in-software rotations of the lattice), which is a high-fidelity operation in our system since the transversal rotations are implemented in parallel with our global Raman laser and the physical single-qubit fidelities are high. We show a logical $\pi/2$ rotation here for the Steane code as an example but emphasize that we can readily realize the various basis states for all of these codes.

Extended Data Fig. 8 Benchmarking the interferometry measurement.

a, To benchmark our gate-based interferometry technique, we prepare variable single-particle pure states (by applying a variable-length resonant Raman pulse) and then reconfigure the system and apply the interferometry circuit on twin pairs. The interferometry circuit converts the anti-symmetric singlet state $|\Psi^-\rangle$ to the computational basis state $|00\rangle$, while converting the symmetric triplet states to other computational states. We plot the resulting twin pair output states in the left panel. We rarely observe the $|00\rangle$ state (1.95(2)% of measurements), with a measurement fidelity independent of the initial state. This low probability P_{00} of observing $|00\rangle$ corresponds to a high extracted single-particle purity of $2P_{00} - 1 = 0.961(3)$ (right panel). We find this measurement to be a useful benchmark, as interferometry miscalibrations can result in significant state-dependence of the observed purity that would then compromise the validity of the many-body entanglement entropy measurement. **b**, Benchmarking the entanglement entropy measurement with Bell state arrays. (Top) Microstate populations during two-particle oscillations between $|11\rangle$ and $\frac{1}{\sqrt{2}}(|1r\rangle + |r1\rangle)$ under a Rydberg pulse of variable duration. Faint lines show measurement results in the $\{|1\rangle, |r\rangle\}$ basis, and dark lines show results in the $\{|0\rangle, |1\rangle\}$ basis after the coherent mapping process. (Bottom) Measured local and global purities by analyzing the number parity of observed $|00\rangle$ twin pairs in each measurement. For this two-particle data we use a gap of 230 ns in the coherent mapping sequence as opposed to the 150-ns gap used in the many-body data.

Extended Data Fig. 9 Raw many-body data and numerical modeling of errors.

a, Raw measured Renyi entropy without subtracting the extensive classical entropy, as a function of subsystem size for quenches from $|rgrgrgrg\rangle$ and $|gggggggg\rangle$. The Renyi entropy of the 4-atom subsystem is the same underlying data as the half-chain entanglement entropy plotted in Fig. 4d of the main text. In the main text, we subtract the data by a fixed offset given by the classical entropy-per-particle, corresponding to the time = 0 offset for

each subsystem size. The extensive, classical entropy offset is slightly larger for the $|rgrgrgrg\rangle$ quench due to non-unity fidelities both of preparing $|r\rangle$ and mapping $|r\rangle \rightarrow |1\rangle$. **b**, Raw global purity after the $|gggggggg\rangle$ quench. The global purity is a sensitive proxy for the fidelity of our entire process. We find this 16-body observable, composed of three-level systems, remains $> 100\times$ the purity expected for a fully mixed state of 8 qubits ($1/2^8$) (see inset). For comparison of scale we also plot single-particle purity to the 8th power, to indicate what the global purity would be if the measurement results on each twin were uncorrelated. **c**, Global purity for the 8-atom quench calculated through numerical modeling of the three-level system $\{|0\rangle, |1\rangle \equiv |g\rangle, |r\rangle\}$ with a variety of simulated error sources. We model the experimentally measured purity by calculating the expectation value of the SWAP operator in the $\{|0\rangle, |1\rangle\}$ basis between two independent chains, taking into account that residual population in $|r\rangle$ results in atom loss and measurement associated with the +1 eigenvalue of the SWAP operator (as the twin state $|00\rangle$ can no longer be detected). The top curve includes only errors from population left in $|r\rangle$ following the coherent mapping step (see [methods](#) text). The second-from-top curve includes single-site dephasing (\(\{T\}_2^{\{ast\}}\)) during the Rydberg dynamics and the coherent mapping gap, modeled by a random on-site detuning which is Gaussian-distributed with zero mean and standard deviation of 100 kHz. The third and fourth curves include multiplication by the experimentally observed raw global purity at quench time $t = 0$, and then further multiplying empirically by an exponential decay $\exp[-16 \times t/(70 \mu s)]$ as a simple model for scattering and decay errors with an experimentally estimated rate of roughly 70 μs for each of the 16 atoms between the two chains.

[Extended Data Fig. 10 Local observables and entanglement entropy for quantum many-body scars.](#)

a, Experimentally measured single-site entropy for each site in the 8-atom chain when quenching from the scarred $|\mathbb{Z}_2\rangle$ state, including the classical entropy subtraction. Solid curves plot exact, ideal (imperfection-free) numerics of H_{Ryd} (Eq. 3); excellent agreement between data and numerics is found for every atom in the chain. **b**, (Top) Same data as Fig. [4f](#) of the main text, showing single-site entropy of the middle two atoms in the chain,

for two different initial states⁷⁵. (Bottom) Measurements of the many-body state in the Z-basis with the interferometry circuit turned off. Characteristic of the scars from the $|\mathbb{Z}_2\rangle = |rgrgrgrg\rangle$ state, the Rydberg excitation probability on the sublattices exhibits periodic oscillations⁷¹. In the bottom row, the dark data points are measured in the $\{|1\rangle, |r\rangle\}$ basis, and the faint data points are measured in the $\{|0\rangle, |1\rangle\}$ basis after the coherent mapping sequence. Measurements in both bases agree well with exact numerics (solid lines), which we emphasize has no free fit parameters and does not account for any experimental imperfections, such as detection infidelity. Moreover, the data indicate the high fidelity of preparation into the $|\mathbb{Z}_2\rangle$ state by use of local Rydberg π pulses. In plotting, we delay the theory curves and the $\{|1\rangle, |r\rangle\}$ basis measurement by 10 ns to account for the fact that the Raman π pulse we apply cuts off the final 10 ns of the Rydberg evolution, when measuring in the $\{|0\rangle, |1\rangle\}$ basis. **c**, Numerical simulations of the single-site Renyi entropy on two adjacent sites in the idealized ‘PXP’ model of perfect nearest-neighbor blockade¹⁴. The system size is 24 atoms with periodic boundary conditions, showing the same out-of-phase oscillations in the entanglement entropy of the two sublattices. **d**, Diagram of the constrained Hilbert space of the system¹⁴. The early-time, out-of-phase entropy oscillations⁷⁵ of the scars can be understood in this constrained Hilbert space picture, where the scar dynamics are known to take the state from the left end ($|rgrgrgrg\rangle$) to the right end ($|grgrgrgr\rangle$) (dark circles represent $|r\rangle$ and white circles represent $|g\rangle$)¹⁴. Near these crystalline ends of this constrained Hilbert space, the Rydberg atoms can fluctuate (high entropy), but the ground state atoms are pinned (low entropy). Our analysis shows that entanglement between atoms on the same sublattice contributes to the eventual degradation of the revival fidelity of the $|\mathbb{Z}_2\rangle$ state.

Supplementary information

Peer Review File

41586_2022_4592_MOESM2_ESM.mp4

Supplementary Video 1 Generation of the toric code using coherent transport of entangled atom arrays. The video depicts the formation of the toric code state on a torus, realized by dynamically changing the quantum processor connectivity throughout the computation time of 1 ms. Ancillas are transported across the array and entangled with stationary data qubits using a two-qubit CZ gate (red ovals), and are then moved to a separate zone for local rotation before readout of all qubits.

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

[Reprints and Permissions](#)

About this article

Cite this article

Bluvstein, D., Levine, H., Semeghini, G. *et al.* A quantum processor based on coherent transport of entangled atom arrays. *Nature* **604**, 451–456 (2022). <https://doi.org/10.1038/s41586-022-04592-6>

- Received: 06 December 2021
- Accepted: 28 February 2022

- Published: 20 April 2022
- Issue Date: 21 April 2022
- DOI: <https://doi.org/10.1038/s41586-022-04592-6>

Share this article

Anyone you share the following link with will be able to read this content:

[Get shareable link](#)

Sorry, a shareable link is not currently available for this article.

[Copy to clipboard](#)

Provided by the Springer Nature SharedIt content-sharing initiative

Further reading

- [**Multi-qubit entanglement and algorithms on a neutral-atom quantum computer**](#)
 - T. M. Graham
 - Y. Song
 - M. Saffman

Nature (2022)

This article was downloaded by **calibre** from <https://www.nature.com/articles/s41586-022-04592-6>

[Download PDF](#)

- Article
- Open Access
- [Published: 20 April 2022](#)

Global seasonal forecasts of marine heatwaves

- [Michael G. Jacox](#) [ORCID: orcid.org/0000-0003-3684-0717](#)^{1,2,4},
- [Michael A. Alexander](#) [ORCID: orcid.org/0000-0001-9646-6427](#)²,
- [Dillon Amaya](#) [ORCID: orcid.org/0000-0001-6720-3265](#)²,
- [Emily Becker](#)³,
- [Steven J. Bograd](#) [ORCID: orcid.org/0000-0003-3872-9932](#)^{1,4},
- [Stephanie Brodie](#)^{1,4},
- [Elliott L. Hazen](#) [ORCID: orcid.org/0000-0002-0412-7178](#)^{1,4},
- [Mercedes Pozo Buil](#)^{1,4} &
- [Desiree Tommasi](#) [ORCID: orcid.org/0000-0003-4027-6047](#)^{4,5}

Nature volume **604**, pages 486–490 (2022)

- 3414 Accesses
- 255 Altmetric
- [Metrics details](#)

Subjects

- [Natural hazards](#)
- [Physical oceanography](#)

Abstract

Marine heatwaves (MHWs)—periods of exceptionally warm ocean temperature lasting weeks to years—are now widely recognized for their capacity to disrupt marine ecosystems^{1,2,3}. The substantial ecological and socioeconomic impacts of these extreme events present significant challenges to marine resource managers^{4,5,6,7}, who would benefit from forewarning of MHWs to facilitate proactive decision-making^{8,9,10,11}. However, despite extensive research into the physical drivers of MHWs^{11,12}, there has been no comprehensive global assessment of our ability to predict these events. Here we use a large multimodel ensemble of global climate forecasts^{13,14} to develop and assess MHW forecasts that cover the world’s oceans with lead times of up to a year. Using 30 years of retrospective forecasts, we show that the onset, intensity and duration of MHWs are often predictable, with skilful forecasts possible from 1 to 12 months in advance depending on region, season and the state of large-scale climate modes, such as the El Niño/Southern Oscillation. We discuss considerations for setting decision thresholds based on the probability that a MHW will occur, empowering stakeholders to take appropriate actions based on their risk profile. These results highlight the potential for operational MHW forecasts, analogous to forecasts of extreme weather phenomena, to promote climate resilience in global marine ecosystems.

Main

Marine heatwaves (MHWs) affect marine ecosystems around the globe, with reported impacts including altered primary productivity, proliferation of harmful algal blooms, displacement of ocean habitats, changes to distributions and populations of marine species, increased human–wildlife conflict and fishery disasters^{1,5,15,16,17,18,19}. Reliable forecasts of these climate extremes would help marine stakeholders to mitigate negative impacts and seize opportunities, thereby improving resilience through anticipatory decision-making⁷. A key step in that direction is the development of MHW predictions, which can be achieved by using operational global climate forecasts. Seasonal (that is, 1–12-month lead time) sea surface temperature (SST) forecasts are routinely used to predict the state of large-scale climate modes, such as the El Niño/Southern Oscillation (ENSO)^{20,21}, and for targeted applications, such as the NOAA

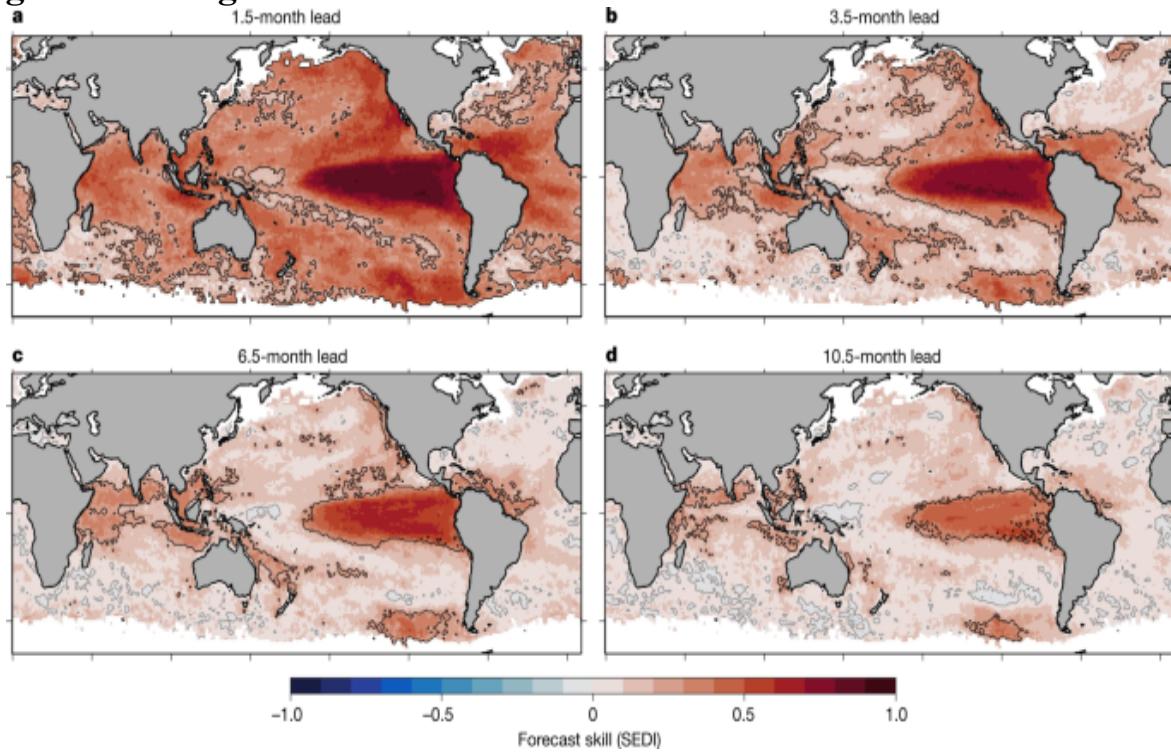
Coral Reef Watch coral bleaching outlook²². Here we use these climate forecast systems to develop global predictions of MHWs and evaluate their skill over the past three decades. In doing so, we highlight the feasibility of predicting MHWs and provide a foundation for a much-needed operational MHW forecast system.

MHW forecast skill

The MHW forecasts developed here show considerable skill on seasonal time scales (Fig. 1). Relative to random forecasts ([Methods](#)), the model MHW forecasts have significant skill nearly everywhere at shorter lead times (up to approximately 2 months), over large areas of the global ocean at lead times of 3–6 months and in some areas at even longer lead times (6–12 months). Forecast MHW probability is also related to MHW intensity, with low probabilities preceding non-MHW periods and higher probabilities preceding stronger MHWs (Extended Data Figs. 1 and 2). The degree of forecast skill is highly dependent on region, with the highest skill found in the tropics (particularly the eastern tropical Pacific) and portions of the extratropical Pacific (off the west coasts of North America and Patagonia, east of Australia). The most predictable regions are not necessarily those with the most intense MHWs; relatively poor MHW forecast skill occurs in much of the Southern Ocean and in Western Boundary Current regions, in which highly energetic and variable currents produce intense but relatively short-lived MHWs^{12,19}. As forecast lead time increases, the global pattern of forecast skill is retained, but forecast skill degrades; at 10.5-month lead time significant skill is confined primarily to the Eastern Tropical Pacific and portions of the Indian Ocean, Indo-Pacific region and high-latitude Eastern Pacific. Similarly, whereas the patterns in skill described above generally hold throughout the year, there is a seasonal modulation of our ability to predict MHWs for specific regions (Extended Data Fig. 3). For example, in some regions 3.5-month lead forecasts are most skilful when initialized in boreal winter (for example, Northeast Pacific, Indian Ocean), whereas for other regions forecasts tend to be more skilful when initialized in boreal spring (for example, tropical Atlantic) or summer (for example, Coral Triangle, Eastern Tropical Pacific). As the forecasts are built on monthly data, their skill reflects an ability to predict longer-lived warming events (as opposed to those lasting just days or weeks), although see the Methods

(section ‘Sensitivity to defining MHWs from daily versus monthly SST’) for an exploration of forecasts using daily data.

Fig. 1: Skill of global MHW forecasts.

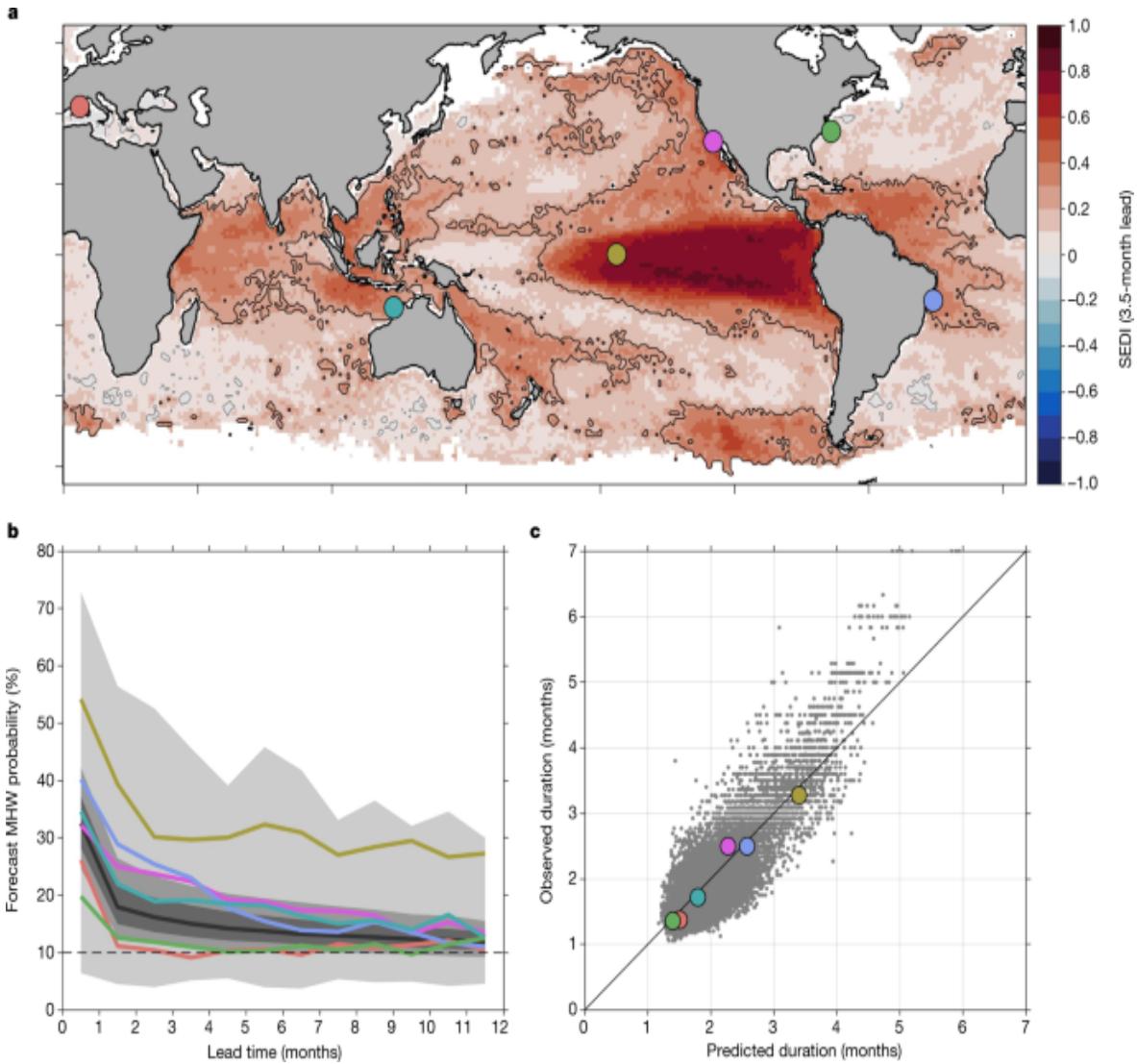


Maps indicate MHW forecast skill, as measured using the SEDI, for the 73-member ensemble of forecasts obtained from six global climate forecast systems for the period 1991–2020. SEDI scores range from -1 (no skill) to 1 (perfect skill). Scores above (below) zero, indicated by grey contours, indicate skill better (worse) than chance, and skill that is significantly better than random forecasts at the 95% confidence level is indicated by black contours. MHW forecasts were initialized every month, with lead times up to 11.5 months; a subset of lead times is shown here. **a–d**, 1.5 months (**a**), 3.5 months (**b**), 6.5 months (**c**) and 10.5 months (**d**). Areas with permanent or seasonal sea ice coverage are masked in white.

From an end-user perspective, it is useful to quantify not just the overall forecast skill for MHWs (that is, whether there will be an active MHW in any given month; Figs. 1 and 2a), but also our ability to predict different characteristics of MHWs. Specifically, we focus on forecasting aspects of MHW evolution including their onset (that is, the month when a MHW first arises) and duration (how long a MHW persists once initiated). When

averaged globally, forecast probabilities indicate elevated MHW likelihood ($>10\%$) even 11.5 months before the observed start of MHWs (Fig. 2b). For shorter lead times (<3 – 4 months), forecast probabilities on average exceed 15% (that is, 50% higher than the base rate of occurrence) and increase rapidly as lead time decreases (Fig. 2b). However, there are substantial regional differences in the predictability of MHW onset. In regions with MHWs driven by rapid atmospheric or oceanic fluctuations, such as the Mediterranean Sea and western boundary currents like the Gulf Stream¹¹, skilful forecast lead times are often limited to two months or less (Fig. 2b) and even intense MHWs are unpredictable at longer lead times (Extended Data Fig. 2). By contrast, for regions in which MHWs result from predictable ocean evolution, such as the Eastern Tropical Pacific²¹, highly elevated MHW probability ($>20\%$, more than double the climatological probability) is forecast up to a year ahead of MHW onset. Regions influenced by atmospheric and oceanic teleconnections also show relatively high forecast skill; on average, the onset of MHWs in areas such as the Indo-Pacific region north of Australia, the California Current System and the northern Brazil Current are presaged by forecast MHW probabilities exceeding 20% approximately 3–6 months in advance (Fig. 2b and Extended Data Fig. 2).

Fig. 2: Predicting the onset and persistence of MHWs.



a, SEDI for 3.5-month lead forecasts (as in Fig. 1). Example locations are indicated by coloured circles and are referred to in the text as Mediterranean Sea (red), Indo-Pacific (blue-green), Eastern Equatorial Pacific (gold), California Current System (pink), Gulf Stream (green) and Brazil Current (blue). **b**, Forecast MHW probability leading up to the initial appearance of observed MHWs. For each $1^\circ \times 1^\circ$ grid cell, forecast probabilities for each lead time preceding the first month of observed MHWs are averaged across all events from 1991 to 2020. Coloured lines correspond to individual locations in **a**, whereas the grey line and shading indicate the global median and the 25th–75th, 10th–90th and 0–100th percentiles. For reference, a horizontal dashed line at 10% marks the base rate of MHW occurrence; probabilities higher than 10% indicate that MHW forecasts correctly predict elevated MHW likelihood from 0.5 to 11.5 months in advance (for example,

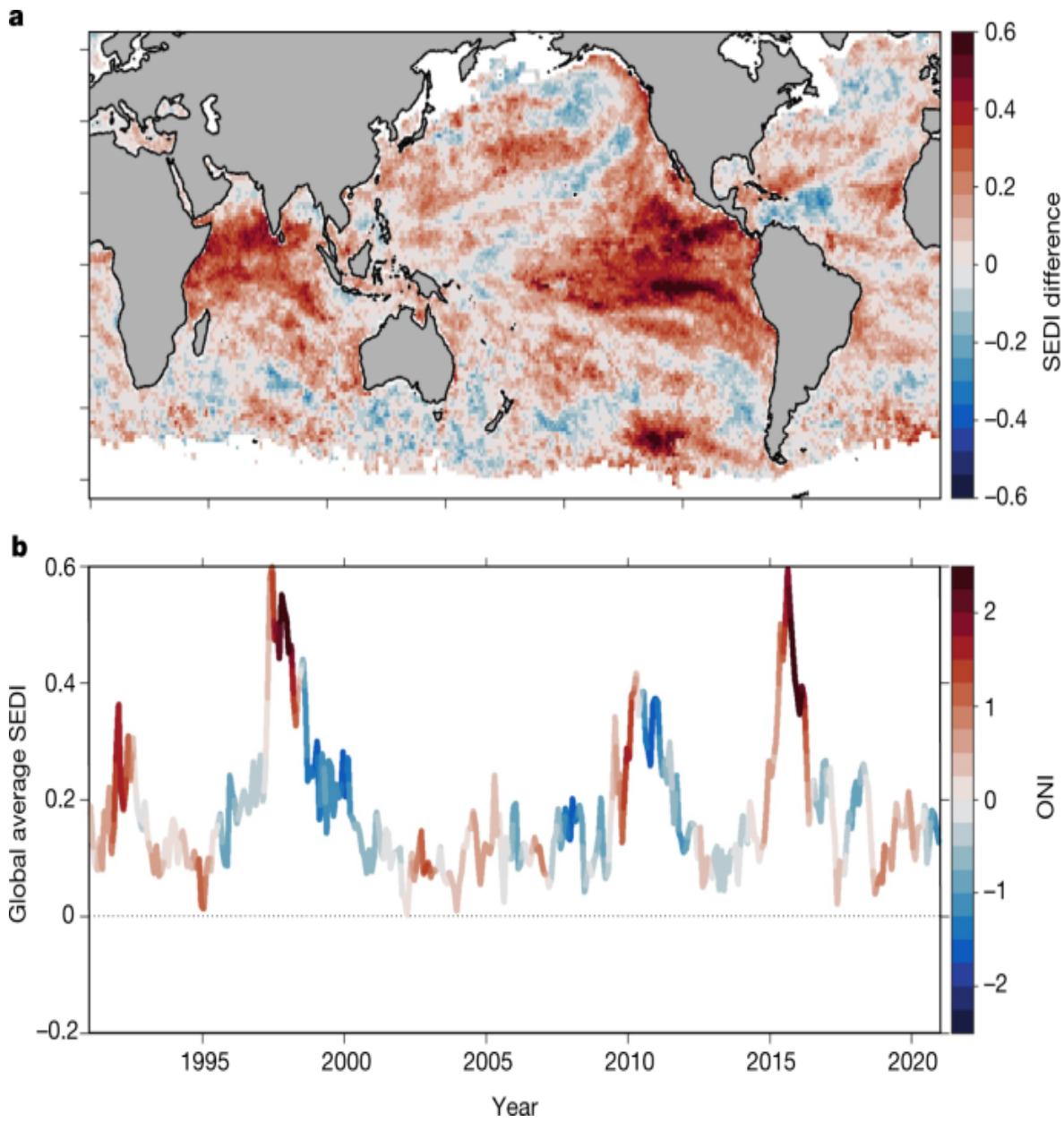
for 30% probability, forecasts are indicating that the likelihood of a MHW occurring has tripled). **c**, Comparison of observed and predicted mean MHW duration (that is, on average how long MHWs last once established at a given location). Each dot represents the mean duration of all events in a $1^\circ \times 1^\circ$ grid cell, with coloured markers corresponding to locations in **a**. The strong correlation ($r = 0.83$) shows that the global spatial pattern of mean MHW duration is reproduced well by forecasts. For temporal correlations of observed and predicted MHW durations at individual locations, see Extended Data Fig. 4.

MHW duration is highly variable across the world's oceans, with the mean length of events ranging from approximately 1 to 7 months globally (based on monthly SST data). We find that MHW forecasts reproduce these spatial patterns well; over the ice-free regions of the ocean, there is a strong correlation between the mean durations of forecast and observed MHWs (Pearson correlation coefficient $r = 0.83$; Fig. 2c). However, not all regions show the same potential for accurately predicting the durations of different MHWs at a specific location. Temporal correlations between observed and predicted MHW duration tend to be highest in regions with the highest overall MHW forecast skill (Extended Data Fig. 4; compare with Fig. 1). Regions of higher skill also tend to have longer MHWs on average. Forecast skill (symmetrical extremal dependence index, SEDI) and mean MHW duration are strongly positively correlated (Pearson correlation coefficient $r = 0.74$), as regions with shorter duration MHWs tend to be less predictable (for example, Gulf Stream, Mediterranean Sea), whereas longer MHW duration is associated with greater predictability (for example, Eastern Pacific).

In the patterns of MHW forecast skill described above, there is a clear imprint of large-scale climate variability. In particular, the dominant signal in global maps of forecast skill (Fig. 1) is the unmistakable signature of ENSO, which is consistent with ENSO effects on seasonal SST predictability more generally^{23,24,25}. Previous work has shown that ENSO is strongly tied to an increased or decreased frequency of MHW occurrence in many regions¹² and, although changes in the frequency of MHWs do not necessarily translate to changes in forecast skill (Methods and Extended Data Fig. 5), there is an ENSO-related modulation of MHW forecast skill. When ENSO is active at the time that forecasts are initialized (that is, during

an El Niño or La Niña event), MHW forecast skill is enhanced in many regions (Fig. 3). Thus, the ENSO state at forecast initialization can be used for a priori assessment of whether a forecast is more or less likely to be skilful. The most pronounced forecast skill increases occur in the Indian and Eastern Pacific Oceans, and the globally averaged MHW forecast skill is closely linked to ENSO. The highest global skill in our 30-year record occurred during the extreme 1997–98 and 2015–16 El Niño events, and additional periods of elevated skill occurred during the 1991–92 and 2009–10 El Niño events and the 1998–2000 and 2010–11 La Niña events. There is debate about how ENSO events will change under increased greenhouse gas forcing, with some studies suggesting they may become more frequent or extreme in the future²⁶, whereas others point to limitations of global climate models in the tropics²⁷ and argue that the ENSO amplitude is more likely to decrease^{28,29}. In any case, these studies should be extended to explore the potential impacts of ENSO changes on the predictability of MHWs and other extreme phenomena.

Fig. 3: Influence of ENSO on MHW forecast skill.



a, Difference in 3.5-month lead forecast skill (SEDI) between periods when ENSO is in an active state and when it is in a neutral state. Active states include both positive and negative phases, defined here as the upper and lower quartiles of the oceanic Niño index (ONI), respectively. **b**, Time series of globally averaged 3.5-month lead forecast skill, with ENSO state (as measured by the ONI) indicated by the colours. Although 3.5-month lead forecasts are shown here, the patterns of enhanced or suppressed skill also hold for other lead times.

MHW forecasts for ocean decision-making

Given the impacts of MHWs on ocean ecosystems, there is a need for operational MHW forecasts to help ocean users prepare for and adapt to these events. In particular, skilful forecasts of MHWs would provide an early warning to resource managers and ocean stakeholders who could act to mitigate potential ecosystem impacts or capitalize on new opportunities. MHW forecasts could also portend changes in the availability of target and bycatch species to recreational and commercial fisheries, giving both fishing fleets and managers forewarning so as to maximize sustainable practices^{11,30,31,32}. For example, proactive fishery closures may reduce both economic losses and ecological risk during events such as the 2014–16 MHW that led to increased baleen whale entanglements in the California Current System^{5,33}. In other cases, MHW forecasts could inform the allocation of increased resources to monitor sensitive sites^{34,35} or guide strategic planning to minimize aquaculture losses³⁶. To support such proactive, climate-ready management approaches, forecast time scales must match those required for end-users to manage climate risk and enact rapid adaptive responses^{37,38}. Here the lead times of skilful MHW forecasts match the time scales of many marine resource management decisions⁸, showcasing the potential for an operational MHW forecast system to be a highly effective decision support tool.

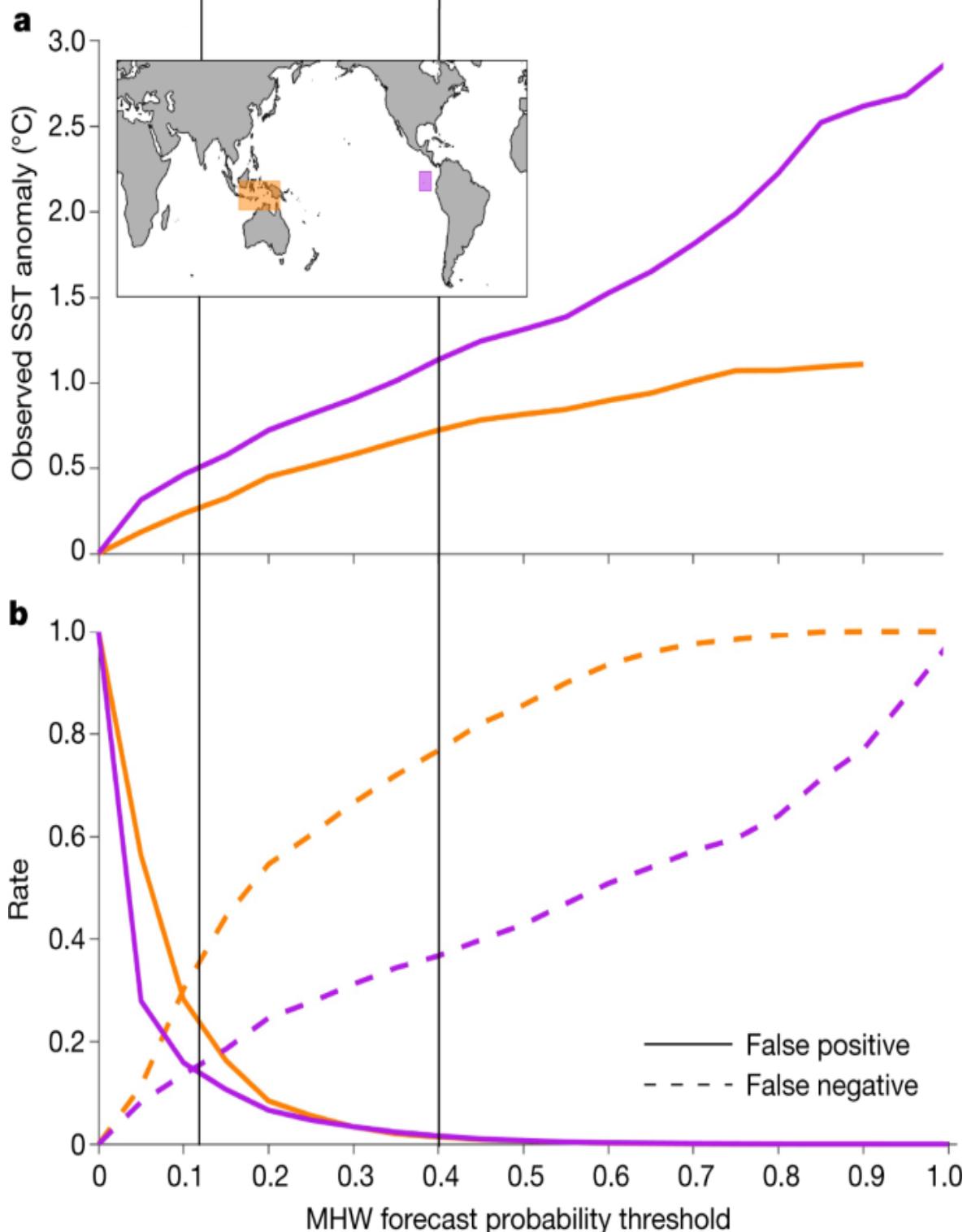
When utilizing probabilistic MHW forecasts, end-users will need to establish thresholds for the MHW probability at which decisions are triggered. The consequences of threshold choice are illustrated here for the Coral Triangle and the Galapagos Islands—two regions with coral reefs that are subject to bleaching or mortality during MHWs (Fig. 4). Enacting a lower threshold means that action is taken more often, which is a risk-averse approach that minimizes the rate of false negatives (failing to take action when a MHW occurs) but also leads to more false positives (taking action when a MHW does not occur). By contrast, higher thresholds limit action to more intense MHWs and a higher certainty of MHW occurrence (Fig. 4 and Extended Data Fig. 1) at the expense of an increased false negative rate. Individual users must balance the risk of inaction with the risk of unnecessary action—for example, trading off potentially adverse ecological impacts of unchanged ocean use during MHWs (false negatives) against

economic consequences of excessive restrictions or excessive monitoring during non-MHWs (false positives). In this context, special consideration should be given to the handling of long-term SST trends in a forecast system, as the decision to retain or remove trends when defining MHWs will alter MHW frequency and consequently the statistics of forecast hits and misses (Methods and Extended Data Fig. 5).

Fig. 4: Adjusting thresholds to support decision-making based on risk tolerance.

Low risk tolerance
 Minimize false negatives
 Avoid inaction when MHW occurs
 Concerned with all MHWs

High risk tolerance
 Minimize false positives
 Avoid action when MHW does not occur
 Concerned with most intense MHWs



a, Observed MHW intensity (SST anomaly) shown as a function of MHW forecast probability threshold for 3.5-month lead forecasts in the Coral Triangle (orange) and Galapagos Islands (purple) regions. For a given threshold, SST anomalies are averaged over all times when the forecast probability was at or above that threshold. **b**, As in **a**, but for rates of false positives (solid lines) and false negatives (dashed lines). Note, **a** and **b** have the same *x* axis.

Operational MHW forecasts

The analysis here provides a template for, and demonstrates the feasibility of, an operational MHW forecast system to be used by ocean decision-makers. Because the MHW forecasts are built on the existing infrastructure of operational climate forecast systems, their transition from research to operations is relatively straightforward. In addition, the analyses performed here can be tailored to specific locations to provide site-specific decision support, including quantification of MHW forecast skill (Fig. 1, 2), its dependence on large-scale climate variability (Fig. 3) and appropriate decision thresholds (Fig. 4). In the future, our MHW forecasts could be expanded upon, with coupled climate forecasts from additional modelling centres and international collaborations (for example, from the Copernicus Climate Change Service; <https://climate.copernicus.eu/seasonal-forecasts>) as well as statistical forecasting methods such as linear inverse modelling^{25,39} or machine learning techniques. In addition, whereas the monthly resolution of seasonal forecast output limits its application to the longer-lived MHWs (>1 month) that tend to be more predictable, forecasts of short-lived events may be useful and viable especially at short (for example, subseasonal) lead times. Regionally tailored MHW predictions can also be generated, either with statistical methods or by downscaling global forecasts, and may provide enhanced skill for specific areas. However, we anticipate that they would supplement, rather than replace, global forecasts. Ensuring global coverage facilitates equitable access to information about ocean extremes that may disproportionately affect regions and communities that lack the resources to develop regionally tailored MHW forecast systems⁴⁰. Likewise, a global operational MHW forecast system can facilitate scientific collaboration to address the impacts of these extreme events on marine social–ecological systems. Given the pressing need for the forewarning of

MHWs, the skilful predictions described here represent a key advance towards improved climate adaptation and resilience for marine-dependent communities around the globe.

Methods

MHW observation

MHWs were identified based on v.2.1 of NOAA's Optimum Interpolation Sea Surface Temperature (OISST v.2.1)^{41,42}. OISST v.2.1 was released in April 2020, and is identical to v.2.0 for data up until 2015 but includes significant quality improvements starting in 2016 (<https://www.ncdc.noaa.gov/oisst/optimum-interpolation-sea-surface-temperature-oisst-v21>). We obtained SST data at daily frequency and 0.25° horizontal resolution from NOAA's Physical Sciences Laboratory (<https://psl.noaa.gov/data/gridded/data.noaa.oisst.v2.highres.html>).

The bulk of our analysis was performed using monthly SST data for both observations and forecasts, but see the section ‘Sensitivity to defining MHWs from daily versus monthly SST’ for a discussion of the implications and practicality of using daily instead of monthly output. Daily 0.25° OISST data were averaged to monthly temporal resolution and 1° spatial resolution for consistency with the forecasts being evaluated (see the next section).

MHWs were identified based on methods proposed in a previous study⁴³ and adapted for monthly data as described in ref. ¹⁹. First, SST anomalies at each grid cell were computed by subtracting the 1991–2020 monthly climatology. MHW thresholds specific to each month of the year were then calculated as the 90th percentile of observed SST anomalies in a 3-month window (for example, for January MHWs, the 90th percentile of all December to February SST anomalies). SST anomalies were then converted to binary time series (MHW or no MHW) depending on whether they were above or below their respective thresholds.

Global climate forecasts

Underlying the MHW forecasts described in this study are seasonal SST forecasts obtained from six global climate models contributing to the North

American Multimodel Ensemble^{13,14}. For each of the six models, an ensemble of forecasts is initialized each month, with the number of ensemble members and the forecast lead time varying between models (Extended Data Table 1). In addition to real time forecasts, a multidecadal set of reforecasts has been performed for each model. Reforecasts, also sometimes referred to as retrospective forecasts or hindcasts, are forecasts simulated for past periods using only information available at the time of forecast initialization (that is, ignoring information that has subsequently become available). The long historical suite of (re)forecasts is necessary to rigorously evaluate the skill and biases of the forecast systems. Here we obtained monthly averaged SST forecast output for 1991–2020, which is a period that is available from all six models, from the IRI/LDEO climate data library (<https://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/>). Output from all models is served on a common grid with 1° resolution in longitude and latitude.

MHW forecasts

To develop MHW forecasts based on the SST forecasts described above, a series of steps were performed for each model. First, the reforecast and forecast periods were concatenated to produce a single set of forecasts for analysis. For models that have more ensemble members in the real time forecasts than in the retrospective forecasts (Extended Data Table 1), we kept the same number of ensemble members as the retrospective forecasts to maintain consistency throughout the analysis period. Next, the model mean forecasts were calculated by averaging together the individual ensemble members of each model. The model mean forecasts were used to calculate model-specific monthly forecast climatologies for each initialization month and lead time, as is customary in climate forecast skill evaluation^{44,45}, and forecast anomalies were calculated for each individual ensemble member by subtracting the model mean climatology. Next, seasonally varying MHW thresholds for each model, lead time and initialization month were calculated with the same methodology described above for SST observations. Forecasts with SST anomalies at or above their respective thresholds were classified as MHWs, resulting in an ensemble of forecasts for binary outcomes (MHW or no MHW). The above steps were repeated for each of the six models, resulting in a multimodel ensemble of 73

members that was used to generate probabilistic monthly MHW forecasts. As forecasts are initialized at the beginning of the month, and we report monthly averages, lead times range from 0.5 months (for example, forecasts of January MHWs, initialized at the beginning of January) to 11.5 months (for example, forecasts of December MHWs, initialized at the beginning of January).

Sensitivity to defining MHWs from daily versus monthly SST

In general, the time scale of predictable events increases with forecast lead time, such that one might look at daily output from weather-scale forecasts (for example, 1–2 weeks lead time) whereas monthly output is more appropriate for seasonal forecasts (up to a year). However, although the most impactful MHWs are overwhelmingly longer-lived events (>1 month)⁴⁶, there is also interest in more ephemeral warm extremes (lasting days to weeks) that may be missed in monthly averaged SST. To illustrate the influence of using daily rather than monthly SST forecasts for MHW prediction, we compare forecasts of MHWs identified based on daily and monthly output from CCSM4 for the locations highlighted in Fig. 2. For these locations, we obtained daily output of forecast SST for the entire 1991–2020 period from the CCSM4 model. We then repeated the analysis of observed and forecast MHWs using daily data with the definition described previously⁴³, which requires MHW thresholds to be exceeded for at least five days. Skill metrics for MHW forecasts generated from daily SST output were calculated using the same methods as those applied to monthly SST forecasts (see ‘MHW forecast evaluation’ below).

Relative to MHW forecasts defined from monthly data, forecasts at daily resolution show shorter mean MHW durations and often slightly lower skill, but no change in the reported patterns in MHW forecast skill (Extended Data Figs. 6 and 7). The consistency between the monthly and daily forecast skill is not surprising given that MHWs defined with daily data are still strongly driven by low frequency variability. However, it is important to note that even though seasonal forecasts can predict the enhanced or reduced likelihood of MHWs on daily time scales, this does not mean that one can predict the details of a specific short (for example, five-day) warming event months in advance. Rather, the skill in MHW forecasts provided at daily

resolution is still reflective of predictable longer-lived SST anomalies, and forecast skill tends to be lower for shorter-lived events (Extended Data Fig. 6).

We also note that daily output is often not publicly available for seasonal forecasts (for example, NMME output is provided as monthly averages). Fortunately, we were able to get daily output from the CCSM4 model to conduct the comparison shown here, but at least in the near term a global MHW forecast system will necessarily be based on monthly output. The same may not be true for subseasonal forecasts (for example, 45 days or less), for which daily MHW forecasts would be more appropriate and daily model output would more likely be available.

Accounting for warming trends

Owing to long-term warming trends in the world's oceans, the rate of MHW occurrence increases over time if fixed thresholds are used to identify them. This effect is prominent even over the relatively short 30-year period examined here, with MHW occurrence increasing two- to threefold if the warming trend is not accounted for (Extended Data Fig. 5). There has been debate in the literature about whether (or when) it is appropriate to retain or remove warming trends in MHW research^{3,19,47}. Here we present results in the main text for MHWs calculated from detrended SST anomalies, but all analyses have been conducted using both methods. For the detrended analysis, we removed linear trends over the 1991–2020 period from the observed SST anomalies and the lead-time-dependent forecast SST anomalies at each grid cell. In the context of MHW forecasts, warming trends may be removed or included depending on the user and the application, but it is important to understand the implications of how trends are handled. Some forecast skill metrics are sensitive to the rate of events, so if trends are retained (and MHW frequency increases over time), those skill metrics will also show trends that are unrelated to the actual capabilities of the model^{48,49} (Extended Data Fig. 5). In the following section we expand on this point in the context of specific forecast skill metrics.

MHW forecast evaluation

Our MHW forecast assessment follows common methods for evaluating climate and weather forecast skill, particularly for extreme events, which present challenges because of their relatively rare occurrence. For forecast verification, we first classify each ensemble member at each time step according to its position in the 2×2 contingency table: true positives (MHW is forecast and occurs), true negatives (no MHW is forecast and MHW does not occur), false positives (MHW is forecast but does not occur) and false negatives (no MHW is forecast but MHW occurs). From the contingency table we calculate two skill metrics, the forecast accuracy and the SEDI, described below. We also calculate the Brier skill score, which is derived from the MHW forecast probability (that is, the average of the binary forecasts from all ensemble members for a given forecast). Below, each of these metrics is described further. All three skill metrics show similar spatial patterns (Extended Data Fig. 8).

Of the many skill metrics proposed for forecasts of extreme events, SEDI⁴⁹ has several desirable qualities⁵⁰, including (1) it is non-degenerate, meaning that it does not trend towards a meaningless limit (for example, zero or infinity) as event rarity increases, (2) it is base-rate independent, meaning that it is not influenced by changes in the frequency of events, and (3) it is equitable, meaning its expected value is the same (zero) for random forecasts, regardless of what method is used to generate the random forecasts⁵¹. SEDI is calculated as

$$\text{SEDI} = \frac{\log F - \log H - \log(\log(1-F)) + \log((1-H)\{F + \log(1-F)\})}{\log(\log(1-F)) + \log((1-H)\{F + \log(1-F)\})},$$

where H is the hit rate (ratio of true positives to total observed events) and F is the false alarm rate (ratio of false positives to total observed non-events). The maximum SEDI score is one and scores above (below) zero indicate forecasts better (worse) than random chance.

For completeness, we also calculate two additional forecast skill metrics: the Brier skill score (BSS) and forecast accuracy. The Brier score is an estimate of the mean square error of the probabilistic forecast:

$$\$ \$ \{ \mathrm{BrS} \} = \frac{1}{N} \sum_{i=1}^N (\{f\}_i - \{o\}_i)^2 \$ \$$$

where N is the total number of forecasts being evaluated, f_i is the forecast probability computed from all ensemble members (that is, the fraction of forecasts predicting a MHW) for forecast i and o_i is the observed probability, which is either zero (no MHW) or one (MHW). The Brier skill score normalizes the Brier score relative to the skill of a reference forecast ($\mathrm{BrS}_{\mathrm{ref}}$):

$$\$ \$ \{ \mathrm{BSS} \} = 1 \{ \mathrm{mbox}\{--\} \} \\ \{ \mathrm{BrS} \} / \{ \{ \mathrm{BrS} \} \} \text{ } \{ \mathrm{ref} \} . \$ \$$$

Here the reference forecast is simply the climatological rate of MHW occurrence (that is, always predicting a 10% chance of a MHW occurring). The BSS ranges from one (perfect skill) to negative infinity (no skill); as for SEDI, scores above (below) zero indicate forecasts better (worse) than random chance.

Forecast accuracy is included as a common and easily understandable skill metric; it is simply the fraction of forecasts that are correct:

$$\$ \$ \{ \mathrm{forecast}; \mathrm{accuracy} \} = (\{ \mathrm{true}; \mathrm{positives} \} + \{ \mathrm{true}; \mathrm{negatives} \}) / N . \$ \$$$

For events that occur on average 10% of the time, the forecast accuracy for random forecasts is 0.82. Thus, MHW forecast accuracy above (below) 0.82 indicates skill better (worse) than random chance.

Significance of forecast skill metrics is quantified using a Monte Carlo simulation with block bootstrapping. Specifically, for a given grid cell we (1) calculate the MHW decorrelation time scale, τ (that is, the lag at which autocorrelation drops below 1/e), and then (2) randomly sample (with replacement) blocks of length τ from the observed MHW time series and concatenate them to create a forecast of length 360 months (the same as the model forecast verification period). This process is repeated to create 1,000 random forecasts, and forecast skill is calculated for each one. The 95% confidence intervals are then calculated from the skill values of the random

forecasts, with significance defined as forecast skill exceeding the 97.5th percentile of the random forecast skill distribution.

When calculating time series of forecast skill (Fig. 3b and Extended Data Fig. 5), skill metrics are calculated over all grid cells at each time, rather than over all times at each grid cell. For example, the forecast accuracy for a given month in Extended Data Fig. 5b is the fraction of the ice-free global ocean for which the MHW state that month was correctly predicted.

Temporal patterns in skill are similar between different metrics (Extended Data Fig. 5), with the exception that there is a base rate dependence in the forecast accuracy and in the individual components of the contingency table (true/false positives/negatives). That dependency is apparent during the strongest El Niño events (when SEDI and BSS increase but forecast accuracy declines), and also in the influence of long-term warming (Extended Data Fig. 5). If SST data are not detrended and consequently the rate of MHWs increases, then forecast accuracy declines, true and false positives increase, and true and false negatives decrease. These trends simply reflect changes in the frequency of events, whereas the performance of the forecast system (for example, as measured by SEDI) does not show a long-term trend (Extended Data Fig. 5). Thus, whether long-term temperature trends are removed or retained during MHW identification and forecasting, one must understand the implications for skill assessment.

Data availability

NOAA High Resolution OISST v.2.1 data^{41,42} were obtained from the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, at their website (<https://www.esrl.noaa.gov/psd/>). Global climate forecasts from the NMME^{13,14} were obtained from the IRI/LDEO climate data library (<https://iridl.ldeo.columbia.edu/SOURCES/Models/.NMME/>). The MHW forecasts described here can be accessed at the NOAA PSL MHWs page (<https://psl.noaa.gov/marine-heatwaves/>).

Code availability

All analyses were performed using MATLAB. Codes can be accessed at https://github.com/mjacox/MHW_Forecasts.

References

1. Smale, D. A. et al. Marine heatwaves threaten global biodiversity and the provision of ecosystem services. *Nat. Clim. Change* **9**, 306–312 (2019).
2. Frölicher, T. L. & Laufkötter, C. Emerging risks from marine heat waves. *Nat. Commun.* **9**, 1–4 (2018).
3. Oliver, E. C. et al. Marine heatwaves. *Annu. Rev. Mar. Sci.* **13**, 313–342 (2021).
4. Mills, K. E. et al. Fisheries Management in a Changing Climate: Lessons from the 2012 Ocean Heat Wave in the Northwest Atlantic. *Oceanography* **26**, 191–195 (2013).
5. Santora, J. A. et al. Habitat compression and ecosystem shifts as potential links between marine heatwave and record whale entanglements. *Nat. Commun.* **11**, 650 (2020).
6. Barbeaux, S. J., Holsman, K. & Zador, S. Marine heatwave stress test of ecosystem-based fisheries management in the Gulf of Alaska Pacific Cod Fishery. *Front. Mar. Sci.* **7**, 703 (2020).
7. Smith, K. E. et al. Socioeconomic impacts of marine heatwaves: global issues and opportunities. *Science* **374**, eabj3593 (2021).
8. Hobday, A. J., Spillman, C. M., Paige Eveson, J. & Hartog, J. R. Seasonal forecasting for decision support in marine fisheries and aquaculture. *Fish. Oceanogr.* **25**, 45–56 (2016).
9. Tommasi, D. et al. Managing living marine resources in a dynamic environment: the role of seasonal to decadal climate forecasts. *Prog. Oceanogr.* **152**, 15–49 (2017).

10. Jacox, M. G. et al. Seasonal-to-interannual prediction of North American coastal marine ecosystems: forecast methods, mechanisms of predictability, and priority developments. *Prog. Oceanogr.* **183**, 102307 (2020).
11. Holbrook, N. J. et al. Keeping pace with marine heatwaves. *Nat. Rev. Earth Environ.* **1**, 482–493 (2020).
12. Holbrook, N. J. et al. A global assessment of marine heatwaves and their drivers. *Nat. Commun.* **10**, 2624 (2019).
13. Kirtman, B. P. et al. The North American Multimodel Ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Am. Meteorol. Soc.* **95**, 585–601 (2014).
14. Becker, E., Kirtman, B. P. & Pegion, K. Evolution of the North American Multi-Model Ensemble. *Geophys. Res. Lett.* **47**, e2020GL087408 (2020).
15. Sen Gupta, A. et al. Drivers and impacts of the most extreme marine heatwave events. *Sci Rep.* **10**, 1–15 (2020).
16. McCabe, R. M. et al. An unprecedented coastwide toxic algal bloom linked to anomalous ocean conditions. *Geophys. Res. Lett.* **43**, 10,366–10,376 (2016).
17. Cavole, L. M. et al. Biological impacts of the 2013–2015 warm-water anomaly in the Northeast Pacific: winners, losers, and the future. *Oceanography* **29**, 273–285 (2016).
18. Bellquist, L., Saccomanno, V., Semmens, B. X., Gleason, M. & Wilson, J. The rise in climate change-induced federal fishery disasters in the United States. *PeerJ* **9**, e11186 (2021).
19. Jacox, M. G., Alexander, M. A., Bograd, S. J. & Scott, J. D. Thermal displacement by marine heatwaves. *Nature* **584**, 82–86 (2020).
20. Barnston, A. G., Tippett, M. K., Ranganathan, M. & L'Heureux, M. L. Deterministic skill of ENSO predictions from the North American

Multimodel Ensemble. *Clim. Dyn.* **53**, 7215–7234 (2019).

21. L'Heureux, M. L. et al. in *El Niño Southern Oscillation in a Changing Climate* (eds McPhaden, M. J. et al.) 227–246 (Wiley-American Geophysical Union, 2020).
22. Liu, G. et al. Predicting heat stress to inform reef management: NOAA Coral Reef Watch's 4-month coral bleaching outlook. *Front. Mar. Sci.* **5**, 57 (2018).
23. Jacox, M. G., Tommasi, D., Alexander, M. A., Hervieux, G. & Stock, C. A. Predicting the evolution of the 2014–2016 California Current System marine heatwave from an ensemble of coupled global climate forecasts. *Front. Mar. Sci.* **6**, 497 (2019).
24. Jacox, M. G., Alexander, M. A., Stock, C. A. & Hervieux, G. On the skill of seasonal sea surface temperature forecasts in the California Current System and its connection to ENSO variability. *Clim. Dyn.* **53**, 7519–7533 (2019).
25. Shin, S.-I. & Newman, M. Seasonal predictability of global and North American coastal sea surface temperature and height anomalies. *Geophys. Res. Lett.* **48**, e2020GL091886 (2021).
26. Cai, W. et al. ENSO and greenhouse warming. *Nat. Clim. Change* **5**, 849–859 (2015).
27. Seager, R. et al. Strengthening tropical Pacific zonal sea surface temperature gradient consistent with rising greenhouse gases. *Nat. Clim. Change* **9**, 517–522 (2019).
28. Callahan, C. W. et al. Robust decrease in El Niño/Southern Oscillation amplitude under long-term warming. *Nat. Clim. Change* **11**, 752–757 (2021).
29. Wengel, C. et al. Future high-resolution El Niño/Southern Oscillation dynamics. *Nat. Clim. Change* **11**, 758–765 (2021).

30. Payne, M. R. et al. Lessons from the first generation of marine ecological forecast products. *Front. Mar. Sci.* **4**, 289 (2017).
31. Brodie, S. et al. Seasonal forecasting of dolphinfish distribution in eastern Australia to aid recreational fishers and managers. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **140**, 222–229 (2017).
32. Le Bris, A. et al. Climate vulnerability and resilience in the most valuable North American fishery. *Proc. Natl Acad. Sci. USA* **115**, 1831–1836 (2018).
33. Samhouri, J. F. et al. Marine heatwave challenges solutions to human–wildlife conflict. *Proc. R. Soc. B Biol. Sci.* **288**, 20211607 (2021).
34. Heron, S. F. et al. Validation of reef-scale thermal stress satellite products for coral bleaching monitoring. *Remote Sens.* **8**, 59 (2016).
35. Eakin, C. M. et al. Ding, dong, the witch is dead (?)—three years of global coral bleaching 2014–2017. *Reef Encount.* **32**, 33–38 (2017).
36. Spillman, C. M. & Hobday, A. J. Dynamical seasonal ocean forecasts to aid salmon farm management in a climate hotspot. *Clim. Risk Manag.* **1**, 25–38 (2014).
37. Hobday, A. J. et al. A framework for combining seasonal forecasts and climate projections to aid risk management for fisheries and aquaculture. *Front. Mar. Sci.* **5**, 137 (2018).
38. Holsman, K. K. et al. Towards climate resiliency in fisheries management. *ICES J. Mar. Sci.* **76**, 1368–1378 (2019).
39. Xu, T., Newman, M., Capotondi, A. & Di Lorenzo, E. The continuum of Northeast Pacific marine heatwaves and their relationship to the Tropical Pacific. *Geophys. Res. Lett.* **48**, 2020GL090661 (2021).
40. Hobday, A. J. et al. Ethical considerations and unanticipated consequences associated with ecological forecasting for marine resources. *ICES J. Mar. Sci.* **76**, 1244–1256 (2019).

41. Banzon, V., Smith, T. M., Chin, T. M., Liu, C. & Hankins, W. A long-term record of blended satellite and in situ sea-surface temperature for climate monitoring, modeling and environmental studies. *Earth Syst. Sci. Data* **8**, 165–176 (2016).
42. Reynolds, R. W. et al. Daily high-resolution-blended analyses for sea surface temperature. *J. Clim.* **20**, 5473–5496 (2007).
43. Hobday, A. J. et al. A hierarchical approach to defining marine heatwaves. *Prog. Oceanogr.* **141**, 227–238 (2016).
44. Stock, C. A. et al. Seasonal sea surface temperature anomaly prediction for coastal ecosystems. *Prog. Oceanogr.* **137**, 219–236 (2015).
45. *Data and Bias Correction for Decadal Climate Predictions* ICPO Publication Series 150 5 (ICPO, 2011).
46. Hobday, A. J. et al. Categorizing and naming marine heatwaves. *Oceanography* **31**, 162–173 (2018).
47. Jacox, M. G. Marine heatwaves in a changing climate. *Nature* **571**, 485–487 (2019).
48. Hogan, R. & Mason, I. in *Forecast Verification: A Practitioner's Guide in Atmospheric Science* 31–59 (Wiley, 2012).
49. Ferro, C. A. & Stephenson, D. B. Extremal dependence indices: improved verification measures for deterministic forecasts of rare binary events. *Weather Forecast.* **26**, 699–713 (2011).
50. Marshall, A. G. et al. Intra-seasonal drivers of extreme heat over Australia in observations and POAMA-2. *Clim. Dyn.* **43**, 1915–1937 (2014).
51. Gandin, L. S. & Murphy, A. H. Equitable skill scores for categorical forecasts. *Mon. Weather Rev.* **120**, 361–370 (1992).
52. Merryfield, W. J. et al. The Canadian seasonal to interannual prediction system. Part I: Models and initialization. *Mon. Weather Rev.* **141**, 2910–

2945 (2013).

53. Lin, H. et al. The Canadian Seasonal to Interannual Prediction System Version 2 (CanSIPSv2). *Weather Forecast.* **35**, 1317–1343 (2020).
54. Delworth, T. L. et al. SPEAR: the next generation GFDL modeling system for seasonal to multidecadal prediction and projection. *J. Adv. Model. Earth Syst.* **12**, e2019MS001895 (2020).
55. Molod, A. et al. GEOS-S2S version 2: the GMAO high-resolution coupled model and assimilation system for seasonal prediction. *J. Geophys. Res. Atmos.* **125**, e2019JD031767 (2020).
56. Infant, J. M. & Kirtman, B. P. Prediction and predictability of land and atmosphere initialized CCSM4 climate forecasts over North America. *J. Geophys. Res. Atmos.* **121**, 12690–12701 (2016).
57. Saha, S. et al. The NCEP climate forecast system version 2. *J. Clim.* **27**, 2185–2208 (2014).

Acknowledgements

We thank N. Mantua for comments on an earlier version of the manuscript as well as A. Capotondi, C. Deser, J. Dias, K. Karnauskas, A. Phillips and J. Scott for discussions in the early stages of the study. This work was supported by funding from the NOAA Climate Program Office's Modeling, Analysis, Predictions and Projections program and the NOAA Fisheries Office of Science and Technology.

Author information

Affiliations

1. NOAA Southwest Fisheries Science Center, Monterey, CA, USA

Michael G. Jacox, Steven J. Bograd, Stephanie Brodie, Elliott L. Hazen & Mercedes Pozo Buil

2. NOAA Physical Sciences Laboratory, Boulder, CO, USA

Michael G. Jacox, Michael A. Alexander & Dillon Amaya

3. University of Miami, Miami, FL, USA

Emily Becker

4. University of California Santa Cruz, Santa Cruz, CA, USA

Michael G. Jacox, Steven J. Bograd, Stephanie Brodie, Elliott L. Hazen, Mercedes Pozo Buil & Desiree Tommasi

5. NOAA Southwest Fisheries Science Center, La Jolla, CA, USA

Desiree Tommasi

Contributions

M.G.J. conceived the study, performed the analysis and wrote the first draft of the manuscript. E.B. helped with downloading and processing the climate forecasts. All authors contributed to the design of the study, interpretation and presentation of results, and writing and revision of the manuscript.

Corresponding author

Correspondence to [Michael G. Jacox](#).

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Villy Christensen, Markus Donat and Neil Holbrook for their contribution to the peer review of this work. [Peer reviewer reports](#) are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

[Extended Data Fig. 1 Forecast MHW probability varies with MHW intensity.](#)

Maps show the mean 3.5-month lead forecast MHW probability associated with periods of **a**, no observed MHW (<90th percentile of SST anomalies) and observed MHWs that are **b**, “weak” (90th–95th percentile of SST anomalies) or **c**, “strong”(>95th percentile). Forecast probabilities higher (lower) than 10% indicate an elevated (reduced) likelihood of MHW occurrence. A positive relationship between MHW forecast probability and observed MHW strength is indicative of forecast skill.

[Extended Data Fig. 2 Observed and predicted MHWs for sample locations.](#)

a, Mean observed MHW intensity (the average SST anomaly during MHWs), with markers corresponding to locations in Fig. 2. **b–g**, Time series of 3.5-month lead forecast MHW probability (blue bars) and observed SST anomalies (black, with MHWs indicated in red). Panel letters correspond to locations shown in **a**.

[Extended Data Fig. 3 MHW forecast skill as a function of season.](#)

Maps show 3.5-month lead forecast skill, as measured by the SEDI, for forecasts initialized in each season: **a**, December–February, **b**, March–May, **c**, June–August, **d**, September–November.

Extended Data Fig. 4 Comparison of observed and predicted MHW duration.

Maps show the correlation (Pearson correlation coefficient) between observed and predicted MHW duration at each location.

Extended Data Fig. 5 Temperature trends can influence MHW frequency and forecast skill metrics.

a, Time series show the global frequency of MHW occurrence (percentage of the ice-free global ocean in a MHW state at each monthly time step) calculated from SST anomalies with linear 1991–2020 trends removed (solid lines) and with trends retained (dashed lines). **b**, Time series of 3.5-month lead forecast skill metrics (Symmetrical Extremal Dependence Index, SEDI; Brier Skill Score, BSS; and Forecast Accuracy, FA). Skill metrics are calculated using globally aggregated forecasts each month (for example, forecast accuracy for a given month is the fraction of the ice-free global ocean for which the MHW state that month was correctly predicted). **c**, As in **b**, but for individual components of the 2x2 contingency table.

Extended Data Fig. 6 MHW forecast skill as a function of MHW duration for forecasts based on daily and monthly SST data.

For locations in **a** (which are the same as those in Fig. 2 and Extended Data Fig. 2), 3.5-month lead MHW forecast skill (SEDI) is shown as a function of mean MHW duration calculated from **b**, daily and **c**, monthly CCSM4 output.

Extended Data Fig. 7 Comparison of lead time dependent MHW forecast skill for forecasts based on daily and monthly SST data.

a–f, For locations in Extended Data Fig. 6, forecast skill (SEDI) is shown as a function of lead time calculated from daily (lines) and monthly (circles) CCSM4 output. Daily skill is smoothed with a 30-day running mean for plotting.

Extended Data Fig. 8 Comparison of MHW forecast skill metrics.

Maps show **a**, SEDI, **b**, Brier Skill Score (BSS), and **c**, forecast accuracy (FA) for 3.5-month lead MHW forecasts. Perfect forecasts would yield a score of one for all three metrics, while the skill expected from random forecasts is 0 for SEDI and BSS, and 0.82 for FA (indicated by gray contours).

Extended Data Table 1 Summary of NMME Forecasts

Supplementary information

Peer Review File

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Reprints and Permissions

About this article

Cite this article

Jacox, M.G., Alexander, M.A., Amaya, D. *et al.* Global seasonal forecasts of marine heatwaves. *Nature* **604**, 486–490 (2022).
<https://doi.org/10.1038/s41586-022-04573-9>

- Received: 29 October 2021
- Accepted: 23 February 2022
- Published: 20 April 2022
- Issue Date: 21 April 2022
- DOI: <https://doi.org/10.1038/s41586-022-04573-9>

Share this article

Anyone you share the following link with will be able to read this content:

[Get shareable link](#)

Sorry, a shareable link is not currently available for this article.

[Copy to clipboard](#)

Provided by the Springer Nature SharedIt content-sharing initiative

[**Marine heatwaves are reliably forecast by climate models**](#)

- Markus G. Donat

Nature News & Views 20 Apr 2022

| [Section menu](#) | [Main menu](#) |

[Download PDF](#)

- Article
- Open Access
- [Published: 13 April 2022](#)

Climate effects on archaic human habitats and species successions

- [Axel Timmermann](#) ORCID: orcid.org/0000-0003-0657-2969^{1,2},
- [Kyung-Sook Yun](#)^{1,2},
- [Pasquale Raia](#) ORCID: orcid.org/0000-0002-4593-8006³,
- [Jiaoyang Ruan](#) ORCID: orcid.org/0000-0003-4733-1125^{1,2},
- [Alessandro Mondanaro](#) ORCID: orcid.org/0000-0003-0325-7066⁴,
- [Elke Zeller](#) ORCID: orcid.org/0000-0003-0244-5882^{1,2},
- [Christoph Zollikofer](#) ORCID: orcid.org/0000-0003-2523-3700⁵,
- [Marcia Ponce de León](#) ORCID: orcid.org/0000-0003-2597-9852⁶,
- [Danielle Lemmon](#)^{1,2},
- [Matteo Willeit](#) ORCID: orcid.org/0000-0003-3998-6404⁷ &
- [Andrey Ganopolski](#)⁷

Nature volume **604**, pages 495–501 (2022)

- 100k Accesses
- 1 Citations
- 722 Altmetric
- [Metrics details](#)

Subjects

- [Archaeology](#)
- [Palaeoclimate](#)

Abstract

It has long been believed that climate shifts during the last 2 million years had a pivotal role in the evolution of our genus *Homo*^{1,2,3}. However, given the limited number of representative palaeo-climate datasets from regions of anthropological interest, it has remained challenging to quantify this linkage. Here, we use an unprecedented transient Pleistocene coupled general circulation model simulation in combination with an extensive compilation of fossil and archaeological records to study the spatiotemporal habitat suitability for five hominin species over the past 2 million years. We show that astronomically forced changes in temperature, rainfall and terrestrial net primary production had a major impact on the observed distributions of these species. During the Early Pleistocene, hominins settled primarily in environments with weak orbital-scale climate variability. This behaviour changed substantially after the mid-Pleistocene transition, when archaic humans became global wanderers who adapted to a wide range of spatial climatic gradients. Analysis of the simulated hominin habitat overlap from approximately 300–400 thousand years ago further suggests that antiphased climate disruptions in southern Africa and Eurasia contributed to the evolutionary transformation of *Homo heidelbergensis* populations into *Homo sapiens* and Neanderthals, respectively. Our robust numerical simulations of climate-induced habitat changes provide a framework to test hypotheses on our human origin.

Main

During the past 5 million years (Ma), a gradual transition in climate conditions has occurred from the warmer and wetter Pliocene (5.3–2.6 Ma) to the colder and drier Pleistocene (2.6–0.011 Ma). During this time, tropical savannahs and open grasslands expanded in central–eastern Africa⁴, which, according to the savannah hypothesis⁵ and variants thereof⁶, contributed to the early evolution of our human ancestors. Milankovitch cycles in solar insolation and climate (Extended Data Figs. 1–3), particularly the

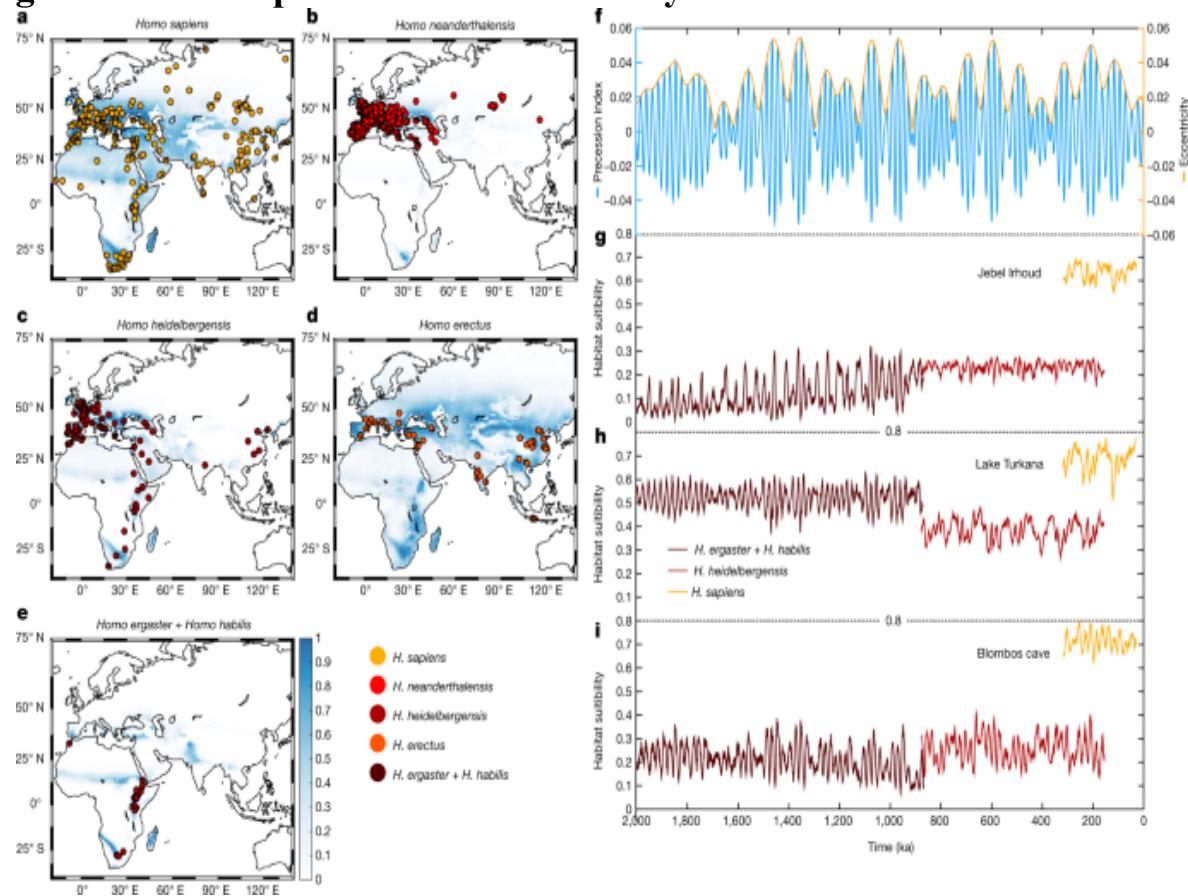
eccentricity-modulated precessional cycle (Extended Data Fig. 1a), further created multiple human migration corridors from sub-Saharan Africa into northern Africa, the Arabian Peninsula and Eurasia^{7,8,9,10}. The existence of these corridors is well supported by fossil, archaeological⁹ and genetic¹¹ evidence. A possible effect of astronomical forcings on early hominin evolution has been suggested in the context of the variability selection hypothesis^{3,12,13}, which posits that early hominin evolution, selection and speciation were influenced by alternating periods of high and low variability in climate and resources.

To better quantify the impact of spatially heterogenous orbital-scale climate variability¹⁴ (Extended Data Fig. 4) on human evolutionary transitions, we conducted an unprecedented transient global coupled general circulation model (CGCM) simulation covering the global climate history of the last 2 Ma (henceforth referred to as the 2Ma simulation). 2Ma is based on the state-of-the-art Community Earth System Model version 1.2 (CESM1.2) in $3.75^\circ \times 3.75^\circ$ horizontal resolution forced with ice-sheet distribution and elevation as well as CO₂ evolution obtained from another transient intermediate-complexity model simulation¹⁵ and astronomical insolation changes¹⁶ ([Methods](#)). 2Ma, which uses an orbital acceleration¹⁷ factor of 5, reproduces key palaeo-climate records such as tropical sea surface temperatures, Antarctic temperatures, the eastern African hydroclimate and the East Asian summer monsoon in close agreement with palaeo-reconstructions (Extended Data Figs. 1 and 2), which supports the realism of our CGCM-based simulation. Glacial–interglacial variability is characterized by a global mean temperature amplitude of approximately 2–3 °C (5–6 °C) during the Early (Late) Pleistocene (Extended Data Fig. 2a), which is consistent with estimates from an Earth system model of intermediate complexity¹⁵ and palaeo-climate data constraints^{18,19,20}.

To quantify the relationship between climate and the presence of hominin species, we built a climate envelope model (CEM; [Methods](#) and Extended Data Fig. 5). This CEM was derived from an extended version of a previously published species database (SDB)^{21,22} composed of geochronologically constrained hominin fossils and archaeological layers containing lithic industries (Fig. 1a–e and [Methods](#)) and topographically downscaled ($1^\circ \times 1^\circ$ grid) 1,000-year averaged data of climate variables

from 2Ma, which are relevant for human survival. These factors included annual mean precipitation, temperature, yearly minimum precipitation and net primary productivity (NPP; [Methods](#)). The 3,245 data entries of the extended SDB (Supplementary Table 1) contain information about location, age, age uncertainty and hypothesized species, selected among early African *Homo* (combining *Homo habilis* and *Homo ergaster* as one group), Eurasian *Homo erectus*, *Homo heidelbergensis*, *Homo neanderthalensis* and *Homo sapiens*. The spatiotemporal climate fields of the 2Ma simulation (see Extended Data Figs. 3 and 4 for select locations) were extracted for the species-presence locations and ages in the SDB and were then statistically aggregated as a CEM. Subsequently, using the Mahalanobis metric^{23,24} and the spatiotemporal climate evolution in 2Ma, we derived a habitat suitability model (HSM; [Methods](#)) for each species, which quantifies the probability of finding fossil and/or archaeological evidence of the species at a given time and geographical location.

Fig. 1: Hominin species habitat suitability.



a–e, African–Eurasian species distribution calculated from a Mahalanobis distance model using four-dimensional climate envelope data of topographically downscaled temperature, precipitation and NPP changes simulated by 2Ma ([Methods](#)) and the locations and ages of fossil and archaeological sites (Supplementary Table 1). The time-averaged habitat suitability (blue to white shading) covering the period of respective hominin presence can be interpreted in terms of probability ([Methods](#)), with values ranging from 0 (habitat unsuitable) to 1 (habitat extremely suitable). Coloured circles represent the locations of fossils and/or archaeological artefacts associated with the five hominin groups. **f–i**, Time series for precession (blue) and eccentricity (**f**) and simulated regional habitat suitability at selected sites of archaeological interest for *H. habilis* and *H. ergaster* (treated jointly as early African *Homo*), *H. heidelbergensis* and *H. sapiens* (**g–i**). The centre locations of a $4^\circ \times 4^\circ$ average area include Jebel Irhoud (34° N, 4° W), a region near Lake Turkana (0° N, 34° E) and Blombos cave (34° S, 21° E).

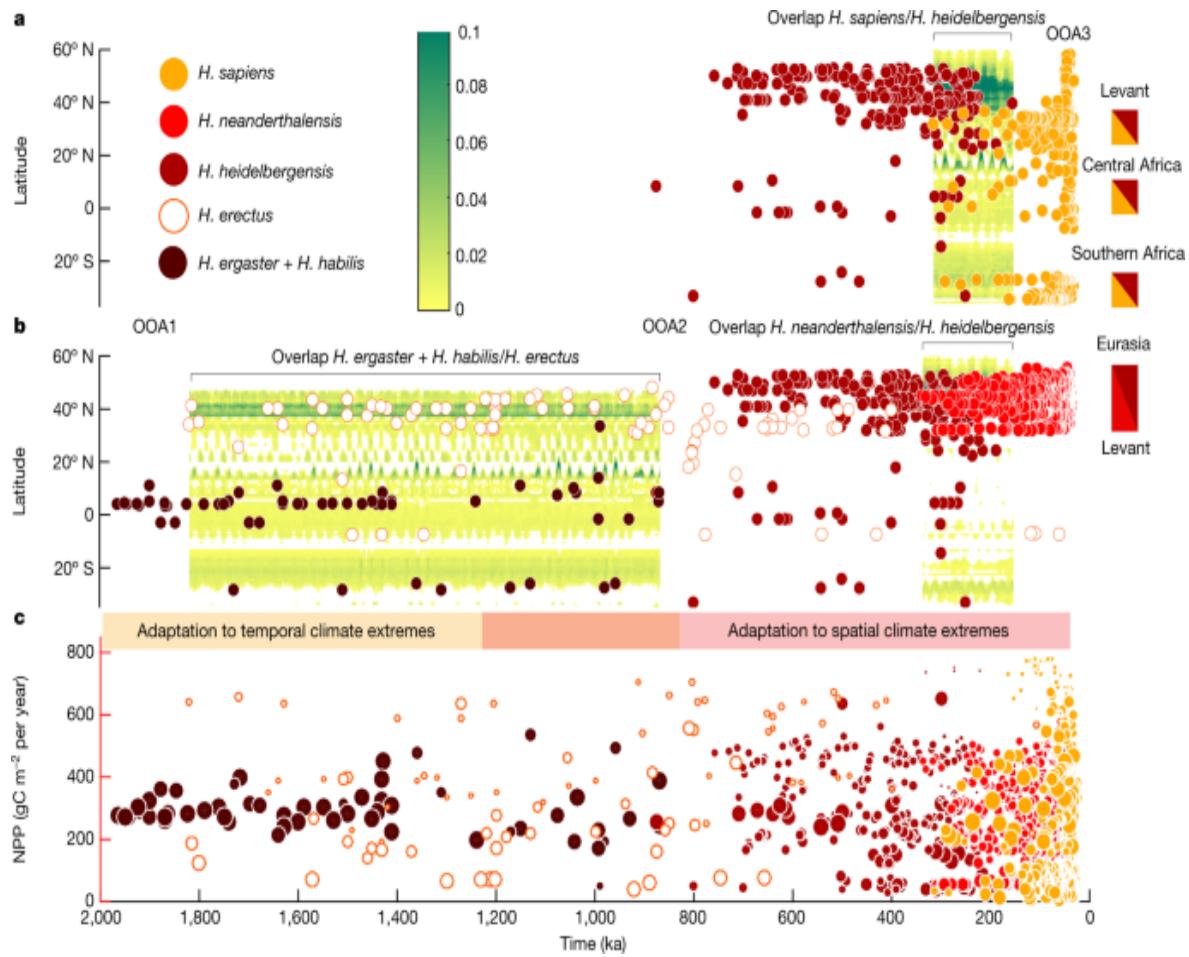
The key goals of our study were (1) to address how past climate changes have affected archaic human habitats; (2) to test whether the current fossil and archaeological records (location and age of each hominin species) have been affected by the orbital-scale evolution of our climate system; (3) to identify common climate envelopes and therefore potential contact zones of hominin groups; and (4) to identify linkages between regional climate shifts and evolutionary diversification.

Time-averaged habitats

To illustrate the connection between climate and the temporal and geographical extent of hominin species, we focused on habitat suitability calculated from the CEM. The simulated time-averaged maps of hominin habitat suitability (Fig. 1a–e) exhibit several interesting features. In particular, the suitable habitat for early African *Homo* (Fig. 1e) is composed of relatively narrow corridors that begin in southern Africa and run northward throughout the rift valley, straddle the Intertropical Convergence Zone and cut across southern Africa in a northwest–southeast direction. Such a limited range and high spatiotemporal heterogeneity of habitat suitability are consistent with high levels of environmental specialization

and sensitivity to regional environmental perturbations, such as eccentricity-modulated precessional cycles (Fig. 1f–i). Even though we included only Eurasian fossils and artefacts for *H. erectus* in the HSM, the predicted global habitat suitability of this species is far more extensive than that of any other hominin species analysed here (Fig. 1d). This is consistent with the concept that *H. erectus* was, on an evolutionary timescale, a flexible generalist who roamed Earth for more than 1 Ma and inhabited a wide range of different environmental conditions (Extended Data Fig. 6). Even though *H. erectus* and early African *Homo* fossil records are treated as geographically disjunct (Fig. 1d,e), there is still regional overlap in their climateenvelopes inside Africa (Fig. 2b and Extended Data Fig. 7d), which is consistent with a deeper ancestral linkage between these two groups²⁵. For *H. heidelbergensis*, we observed a time-averaged habitat suitability pattern that was qualitatively similar to that of *H. neanderthalensis* (Fig. 1b,c). By comparing the climate niches of *H. sapiens* (Fig. 1a) with those of other hominin species, we determined that our own species was best equipped to cope with dry conditions (Fig. 1g and Extended Data Fig. 6c). This extended climatic tolerance of *H. sapiens* was introduced into the CEM by a group of fossils and archaeological artefacts located in northeastern Africa, the Arabian Peninsula and the Levant (Fig. 1a and Supplementary Table 1). This tolerance of dry conditions greatly enhanced the mobility of *H. sapiens*, which may have further facilitated the documented multiple-wave dispersals into Eurasia across the Sinai passage or Bab-el Mandeb strait into the Levant (Extended Data Fig. 6c) and the Arabian Peninsula⁹, respectively.

Fig. 2: Habitat overlap, succession and geographical distribution of fossils.



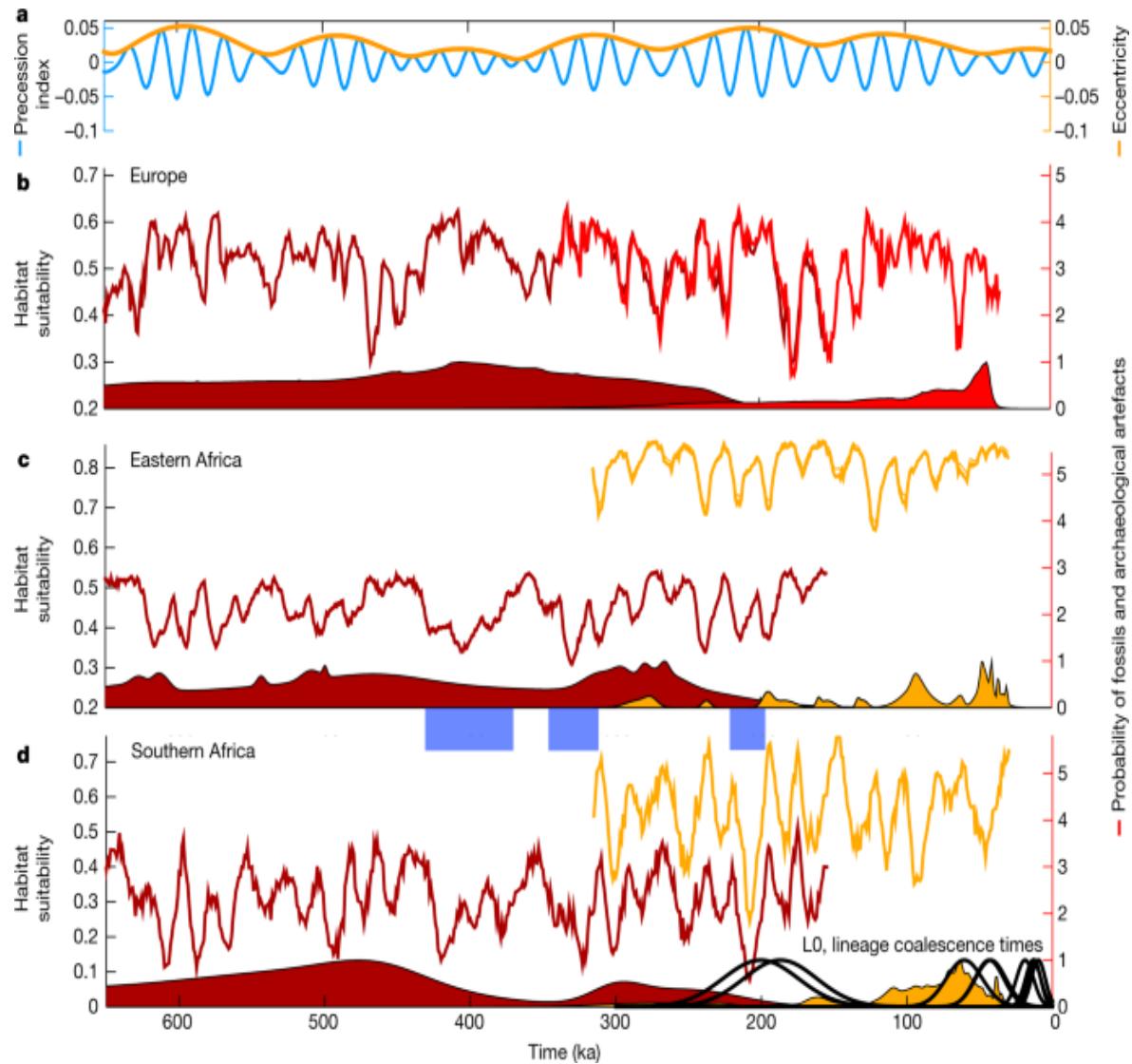
a, Green shading represents a Hovmöller (time–latitude) diagram of the zonal mean of the spatial scalar product of habitat suitability for *H. sapiens* and *H. heidelbergensis*. Circles represent the corresponding average age and latitudinal distributions of fossils and archaeological artefacts. High values of habitat overlap coinciding with joint presence of fossils indicate possible locations of hominin interaction, diversification and possibly speciation. **b**, Same as **a** but for *H. neanderthalensis* and *H. heidelbergensis* (right side) and *H. erectus* and *H. ergaster–habilis*. ‘Out-of-Africa’ migration periods are marked as OOA. Potential regions for gradual diversification and transformation are indicated by dual-coloured boxes. **c**, NPP ([Methods](#)) for each fossil and archaeological site (coloured circles), selected by averaging the 2Ma NPP data in a $6^\circ \times 6^\circ$ vicinity around the fossil sites and for their respective fossil ages. The size of the circle represents the great-circle (Haversine) distance to a grid point in central–eastern Africa (4° N, 36° E), with larger circles indicating closer proximity to this location.

Climate impacts on species distributions

The temporal evolution of our HSM exhibits pronounced Milankovitch cycles (Figs. 1g–i and 3, Extended Data Fig. 6 and Supplementary Videos 1–5). Tropical regions are characterized mostly by precessional cycles, which are modulated by eccentricity cycles of 80–120 thousand years (ka) and 405 ka (Fig. 1f), whereas extratropical locations show a stronger component of 80–120 ka due to CO₂ and ice-sheet forcings (Extended Data Fig. 6b,d).

Notably, regional climate changes and the resulting habitat changes were driven not only by the interplay of local forcings but also by remote effects such as eastern equatorial Pacific temperature changes, as suggested recently²⁶ by a synthesis of African hydroclimate proxy records and tropical sea surface temperature reconstructions.

Fig. 3: Hominin species succession in Europe and southern Africa.



a, Eccentricity (orange) and precession (blue) indices from Fig. 1f. **b**, Habitat suitability calculated from the CEM for *H. heidelbergensis* (dark red curves) and *H. neanderthalensis* (red curves) in Europe ($4^\circ \times 4^\circ$ average centred around 45° N, 6° E). **c**, Same as **b** but for *H. heidelbergensis* (dark red) and *H. sapiens* (orange) in eastern–central Africa ($4^\circ \times 4^\circ$ average centred near 5° S, 36° E). **d**, Same as **c** but for southern Africa ($4^\circ \times 4^\circ$ average centred near 24° S, 24° E). The shaded curves represent probability estimates of the occurrence of respective fossil and/or archaeological data obtained from the ages and age uncertainties of the fossils in the respective broader regions. The thick black curves in **d** represent the probability of the coalescence times³² of the mitochondrial DNA lineages L0, L0d'k, L0a'b'f'g, and L0k as a genetic manifestation of deep-rooted modern human ancestry

in southern Africa. Light-blue shaded bars indicate key periods of reduced habitat suitability in southern Africa. The robustness of these calculations against uncertainties in species attribution and dating of archaeological layers ([Methods](#)) is further documented in Extended Data Fig. [10](#).

To further test whether orbital-scale climate variability influenced the observed spatiotemporal distribution of hominin species, we recalculated the CEM for each species using the fossil and archaeological data in combination with a time-scrambled trajectory of the CESM1.2 simulation ([Methods](#)). The resulting new CEM is different from the original one in that it assigns different, randomized temporal climate states to the fossil and archaeological data while maintaining the overall regional co-variability of the climate components and the long-term mean state. By comparing the long-term mean difference in the habitat suitability projections of the null-hypothesis model with the original one, we could then ascertain whether Milankovitch cycles influenced the distribution of fossils and archaeological sites on a regional level. The results for *H. sapiens*, *H. neanderthalensis* and *H. heidelbergensis* (Extended Data Fig. [8a–c](#)) showed statistically significant differences ($P < 0.05$, paired *t*-test) in the calculated habitat suitability, with values of more than 0.05 in magnitude attained when comparing the unshuffled and shuffled models over parts of Asia, Europe and Africa. This documents that the orbital-scale trajectory had an important role in determining where and when hominin species lived.

Species successions

To identify locations where potential succession or speciation of hominin groups may have taken place, we calculated the species overlap as the covariance of habitat suitability between the different hominin groups (Fig. [2](#) and Extended Data Fig. [7](#)). We assumed that species that interacted with or emerged from each other probably shared similar regional climate envelopes, at least during their transition time.

For *H. neanderthalensis* and *H. heidelbergensis*, the highest values of niche overlap were found in Europe (Fig. [2b](#) and Extended Data Fig. [7b](#)), which also hosts archaeological artefacts and fossils from both species^{[27](#)} (Fig. [3b](#)) and has been regarded as the ‘birthplace’ of Neanderthals^{[28](#)}. By comparing

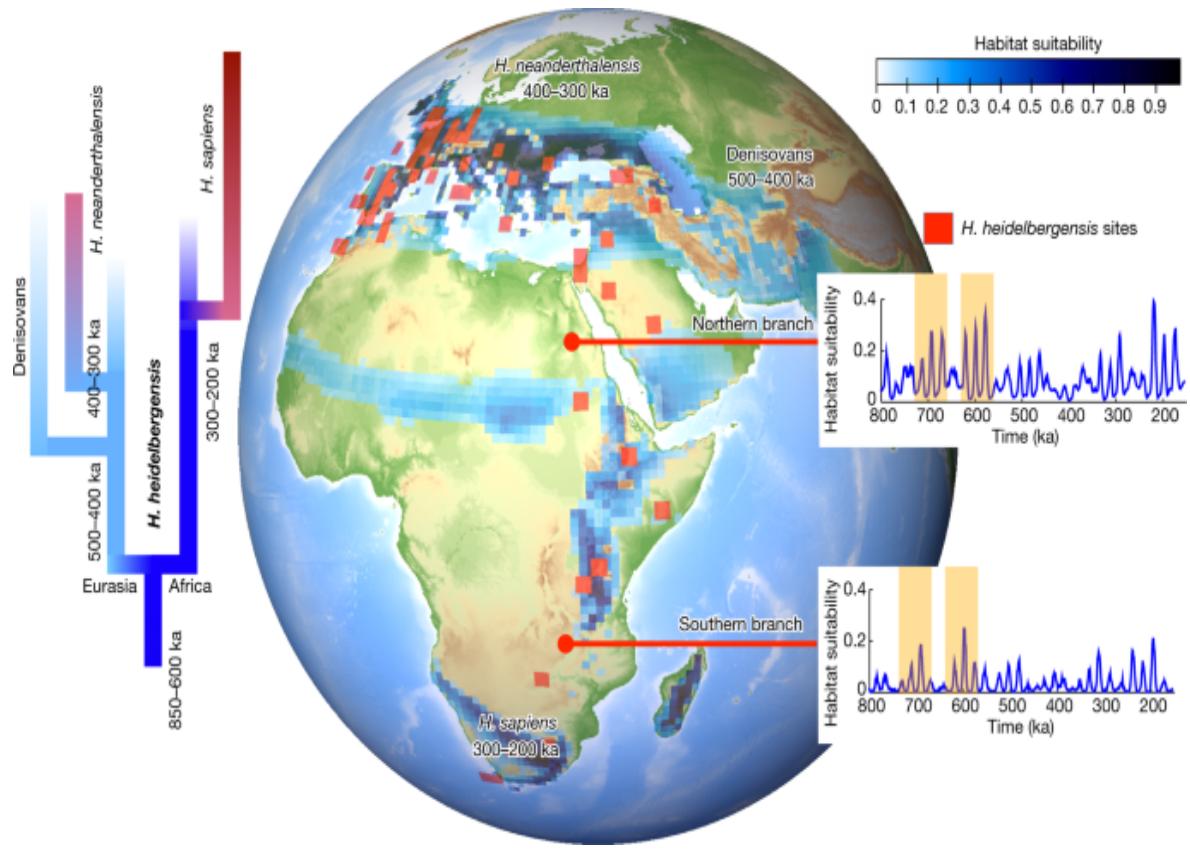
the zonal mean overlap for *H. sapiens* and *H. heidelbergensis* with their respective fossil and archaeological sites, we identified two key areas with climatic conditions that were suitable for joint occupancy outside Europe: central–eastern Africa and southern Africa (Fig. 3). In addition to habitat overlap (Fig. 2), we calculated the regional habitat similarity as an indicator for potential evolutionary transitions such as baseline evolution or speciation events (Extended Data Fig. 7a). A more detailed analysis into the simulated regions of orbitally varying species overlap indicated two pronounced periods of reduced habitat suitability in southern Africa for *H.*

heidelbergensis at 415–360 ka and 340–310 ka (Fig. 3d). These prolonged eras of climatic stress were further characterized by low probabilities for fossil and archaeological records in subequatorial Africa. Subsequently, from 310 ka to 200 ka, high values of habitat suitability correlated with the first evidence of *H. sapiens* in southern Africa in terms of both fossils and archaeological artefacts^{29,30,31} (Figs. 2a and 3d) as well as presence of the earliest mitochondrial DNA lineage (L0) of southern African origin³². The disappearance of *H. heidelbergensis* from Africa could potentially be explained by progressive evolution of *H. heidelbergensis* into *H. sapiens*, which would be consistent with the presence of their respective fossils and archaeological artefacts at about 200–300 ka (Supplementary Table 1) and their similar values for regional habitat suitability (Fig. 3d). By contrast, a larger habitat discrepancy between *H. heidelbergensis* and *H. sapiens* (Fig. 3c) occurred in central Africa, indicating that gradual species transition or diversification is less likely to have occurred in this region than in southern Africa, at least from a climate envelope perspective. Another major climate disruption in southern Africa around 210–200 ka (Fig. 3d) during the austral summer perihelion (Fig. 3a) could have increased the regional environmental stress on *H. sapiens*, leading to dispersal and, subsequently, genetic diversification. This timing is consistent with the presence of the first known mutation event that occurred in our reconstructed common mitochondrial ancestry³², even though considerable uncertainties in dating and methodology still exist³³. Overall, our analysis suggests that the emergence of *H. sapiens* and the gradual disappearance of *H. heidelbergensis* in southern Africa coincided with long-term climatic anomalies during Marine Isotope Stages 11 and 9.

Speciation and dispersal

We combined a transient Pleistocene climate model simulation with an extensive compilation of hominin fossils and archaeological artefacts to study the environmental context of hominin evolution. On the basis of the resulting HSM and palaeogenetic evidence^{34,35}, we propose the following scenario (Fig. 4): about 850–600 ka, *H. heidelbergensis*, which may have originated from *H. ergaster* in eastern Africa (Extended Data Fig. 7e), split into southern and northern African branches, the latter of which included northern African and Eurasian populations. The intensified dispersal into off-equatorial regions may have occurred during periods of high eccentricity around 680 ka and 580 ka, which increased habitat suitability in otherwise inhospitable regions (Fig. 4, insets). The southern branch experienced considerable climatic stress in southern Africa during Marine Isotope Stages 11 and 9, which could have accelerated either a gradual or a cladogenetic transition into *H. sapiens*³⁶. The Eurasian populations of the northern branch further bifurcated around 430 ka, possibly giving rise to Denisovans, which populated parts of central and eastern Asia. Inside central Europe, *H. heidelbergensis*, which experienced strong local climatic stress due to eccentricity-modulated ice-age cycles (Fig. 3b), gradually evolved into *H. neanderthalensis* between 400 ka and 300 ka. Side branches to northwestern Africa, back-propagation, multiple dispersals³⁷, interbreeding³⁸ and subsequent speciation³⁹ may have further complicated the picture.

Fig. 4: Schematics of *H. heidelbergensis* succession.



On the basis of fossil ages, we propose a split of *H. heidelbergensis* into northern and southern branches (blue shading, habitat suitability) around 850–650 ka. The gradual transition at 300–200 ka of *H. heidelbergensis* into *H. sapiens* in southern Africa is supported by fossil and archaeological data in this region and habitat overlap estimates (Fig. 2a). The proposed divergence at 400–300 ka of *H. heidelbergensis* into Neanderthals in Europe is consistent with recent genetic estimates³⁴. This scenario is also in agreement with Neanderthal whole-genome data⁴⁴ that suggest a population split between Neanderthal–Denisovan and modern human lineages between 550 and 765 ka and a divergence between Neanderthals and Denisovans around 445–473 ka. Possible eccentricity-modulated windows for early non-coastal north–south migrations occurred around 700 ka and 600 ka during periods of high eccentricity, according to the calculated HSM (see inset time series for $4^\circ \times 4^\circ$ averages centred near 21° N, 31° E and 20° S, 31° E).

Recent studies have suggested that the sequence of hominin speciation events and the long-term positive trend in brain size may have been linked to past climatic shifts in Africa⁴⁰. Our analysis supports the notion of strong

Milankovitch cycles in early hominin habitat suitability in central Africa (Fig. 1h). Moreover, during the early Pleistocene (2–1 Ma), early African *Homo* populations occupied two main habitats: one in central–eastern Africa and the other in southern Africa (Fig. 2b). On average, these groups preferred geographical regions that were characterized by relatively stable NPP values of 200–380 gC m⁻² per year (Fig. 2c). Within Africa, early African *Homo* populations adapted mostly to local orbital-scale variations in climate and NPP (Fig. 1g–i and Extended Data Figs. 3, 4 and 9), as reflected also in their habitat suitability. After the mid-Pleistocene transition and with the emergence of *H. heidelbergensis* between approximately 885 and 865 ka, the dynamics again changed remarkably. *H. heidelbergensis* began to migrate into Eurasia and other regions, encountering along their journey a much wider spatial range of NPP, from extremely low values of 20 gC m⁻² per year to values exceeding 600 gC m⁻² per year (Fig. 2c). These migrating groups crossed large spatial gradients in climate and NPP that far exceeded the temporal ranges in NPP experienced by their more stationary Early Pleistocene predecessors. This transition to global wanderers about 0.8–0.6 Ma must have required *H. heidelbergensis* to acquire new adaptation skills, which in turn also strengthened their ability to further expand their geographical range, thereby providing a strong positive migration–climate adaptation feedback. Our analysis clearly shows that *H. erectus* had already undergone such a transition from regional dweller to early global wanderer before 1.8 Ma (Fig. 2c). Together with the *H. heidelbergensis* evidence, this indicates that dispersals from Africa always involved an adaptive shift, either biological or cultural, to wider climate envelopes. Therefore, to understand hominin evolution during the Pleistocene, the full spatial and temporal complexity of the climate signal and the corresponding habitat suitability must be considered.

Discussion

The main conclusions of our analysis are robust with respect to the existing uncertainties in species attribution, particularly for the period from 1 to 0.3 Ma, and the dating of archaeological layers, as demonstrated by key HSM calculations with four different scenarios that accounted for these factors (Methods and Extended Data Fig. 10). Although our study is based on species-stratified fossil and archaeological input data, our calculation of

species overlap as HSM co-variability allowed us to treat potential species transitions and successions in human evolutionary history quantitatively and to identify their spatiotemporal characteristics. To the best of our knowledge, such research has not been reported thus far. The HSM captures regionally distributed patchworks of habitable areas in agreement with a general multiregional perspective⁴¹ (Figs. 1 and 4). According to our CEM, southern and eastern Africa as well as the region north of the Intertropical Convergence Zone emerge as potential long-term refugia for various types of archaic humans. As the climate changed on orbital timescales, these refugia shifted geographically, creating population patterns with greater complexity. Further analysis of the pan-African connectivity of refugia in our HSM dataset, as shown in the inset in Fig. 4, will increase understanding of hominin dispersal, interbreeding and cladogenetic transitions as well as potential cultural exchanges.

In summary, we demonstrated that astronomically forced climate shifts were a key factor in driving hominin species distributions⁴² and dispersal and were probably important for diversification⁴³.

Methods

2Ma simulation

We conducted the 2Ma simulation with the Community Earth System Model (CESM), version 1.2, at an ocean and atmosphere resolution of approximately $3.75^\circ \times 3.75^\circ$. The model uses bathymetry of the Last Glacial Maximum and time-varying forcings of greenhouse gases¹⁵, ice sheets¹⁵ and astronomical insolation conditions¹⁶. CESM1.2 has a relatively low standard equilibrium climate sensitivity (ECS) of 2.4°C per CO_2 doubling, which lies outside the likely range of estimates⁴⁵ ($3.7 \pm 1.2^\circ\text{C}$) obtained with other climate model simulations conducted as part of the Coupled Model Intercomparison Project, phase 6. However, this value is within the lower range of recent estimates compiled by the Intergovernmental Panel on Climate Change sixth assessment report⁴⁶ of Working Group 1, which identifies a very likely ECS range of $2\text{--}5^\circ\text{C}$. To obtain a more realistic response to past long-wave radiative forcings in our palaeo-climate model

simulation and to implicitly capture radiative effects of other CO₂-correlated forcings⁴⁷ from dust, vegetation, N₂O or CH₄, we therefore scaled the range of the applied CO₂ forcing¹⁵ by a factor of 1.5. The resulting effective ECS, which includes non-CO₂ greenhouse gas forcings, was in our case approximately 3.8 °C. Our result is in reasonable agreement with the Coupled Model Intercomparison Project phase 6 estimate and previous palaeo-climate estimates^{18,19} of 3.2 °C, which were obtained from reconstructions of the global mean surface temperature and radiative forcings covering the last 784,000 years. Amplification of the CO₂ forcing in CESM1.2 led to a realistic representation of the amplitude of global mean, tropical and Antarctic temperature changes (Extended Data Figs. 1b, 2a, b) and to a simulated temperature range between Last Glacial Maximum and Late Holocene conditions of approximately 5.9 °C. This result is in close agreement with recent palaeo-proxy-based estimates²⁰. Similar to previous long-term transient climate model simulations conducted with Earth system models of intermediate complexity^{7,48}, the CESM1.2 simulations use an orbital acceleration factor of 5, which means that the 2-million-year orbital history is squeezed into 400,000 model years in CESM. The complete model trajectory is based on 21 individual chunks that were run in parallel, with each covering at least one interglacial–glacial cycle (Supplementary Table 2). Moreover, each chunk overlaps with the next chunk so that the issue of initial conditions and spin-up time can be evaluated properly. The final climate trajectory is obtained by combining the individual chunks and by using sliding linear interpolation in the chunk-overlap periods. The model simulation has been evaluated against numerous palaeo-proxy-based data (Fig. 1 and Extended Data Fig. 1). Unlike other Earth system models^{49,50,51}, the 2Ma simulation conducted with CESM1.2 does not generate strong internal millennial-scale variability such as that shown by Dansgaard–Oeschger cycles. The CESM1.2 data are provided on the climate data server of the Institute for Basic Science (IBS) Center for Climate Physics at <https://climatedata.ibs.re.kr>.

Topographic downscaling

The T31 spectral resolution of the 2Ma CESM1.2 simulation (approximately 3.75° × 3.75° horizontal resolution) is too coarse to properly capture

important topographic barriers, which may have affected the dispersal and distribution of archaic humans. We applied simple downscaling of the simulated monthly surface temperatures $T_s(t)$ onto a $1^\circ \times 1^\circ$ horizontal grid by accounting for the difference in height $\Delta h(t)$ between the ETOPO5 topographic dataset and the orographic forcing of the 2Ma experiment. The lapse rate-corrected temperatures were then calculated as $T_s^*(t) = T_s(t) - g\Delta h(t)$, where g represents a constant average lapse rate of $g = 6^\circ\text{C}$ per 1,000 m. The simulated rainfall $p(t)$ was downscaled onto the high-resolution topography by accounting for temperature-dependent moisture availability through the Clausius–Clapeyron equation as $p^*(t) = p(t)e^{[17.625T_s^*/(T_s^* + 243.04) - 17.625T_s/(T_s + 243.04)]}$.

A posteriori calculation of NPP

2Ma uses fixed plant functional types but a prognostic leaf area index. Therefore, we calculated the NPP a posteriori (Extended Data Figs. 5 and 9) using a simple empirical relationship among temperature, precipitation and tree fraction. The topographically downscaled temperature T_s^* (in degrees Celsius) and precipitation p^* (in millimetres per year) of the 2Ma simulation were used at every grid point to calculate the tree fraction⁵² as $\tau = 0.95\{1 - e^{[-\beta(T_s^* - T_m)]}\}p^{*\alpha}/(p^{*\alpha} + f)$, with the additional term $f = be^{[\gamma(T_s^* - T_m)]}$, and the parameters $\beta = 0.45$, $\alpha = 3$, $b = 2.6 \times 10^6$, $\gamma = 0.155$ and $T_m = -15^\circ\text{C}$; τ is capped between 0 and 1. Subsequently, the downscaled NPP can be calculated from an empirical model⁵³ as $N^* = \{6,116[1 - e^{(-0.0000605p^*)}](1 - \tau) + \tau \min(\text{FP}, \text{FT})\}f(\text{CO}_2)$, where the minimum (min) is taken over the mathematical terms $\text{FP} = 0.551p^{*1.055}/e^{(0.000306p^*)}$ and $\text{FT} = 2,540/[1 + e^{(1.584 - 0.0622T_s^*)}]$. The function $f(\text{CO}_2) = [1 + 0.4\ln(\text{CO}_2/280)/\ln(2)]$ captures the bulk effect of CO_2 fertilization of plants⁵⁴ in the same way as the CLIMBER Earth system of intermediate complexity, and its time evolution is obtained from the transient CO_2 forcing of CESM1.2.

Extended dataset of archaeological and fossil hominin data

The SDM for the *Homo* genus was derived from a recent compilation of archaeological and fossil data²¹. The original data compilation published in 2020 (ref. ²¹) included 2,754 radiometric age estimates for fossil hominin occurrences, each accompanied by the confidence interval around the estimate, the fossil site name and the archaeological layer within the site (where available) from which the dated sample was derived, the geographical coordinates of the site and the possible attribution to one or more than one *Homo* species. Confident attributions to a single species generated a core record, whereas instances with multiple attributions formed an extended record. Six different species were recognized: *H. habilis*, *H. ergaster*, *H. erectus*, *H. heidelbergensis*, *H. neanderthalensis* and *H. sapiens*. The updated record, as shown in Supplementary Table 1, contains 3,245 data entries restricted to the temporal age interval of 2 Ma–30 ka; those from Australia and the Americas were excluded. Further, we combined *H. habilis* and *H. ergaster* into a single African Oldowan toolmaker species. Each occurrence is attributed to a given species depending on the presence of unambiguous anatomical remains, either singly or in connection to a specific lithic tool industry. This helped to guide identification if this was not otherwise possible from the bones and teeth alone (398 entries, 12%), the age limits of the individual species or the stone tool industry. For example, an occurrence in Africa older than the first appearance of *H. sapiens* at Jebel Irhoud⁵⁵ yet younger than the first appearance of *H. heidelbergensis* at Melka Kunture⁵⁶ is attributed to *H. heidelbergensis*. Moreover, French Mousterian stone tools have been unambiguously assigned to *H. neanderthalensis*, whereas Aurignacian tools were attributed to *H. sapiens*. When these criteria were applied, the core record included 94.5% of the attributions, 48.5% of which refer to *H. neanderthalensis* and 37.5% of which refer to *H. sapiens*. Where neither of these criteria was met (in the original compilation, the SDM acknowledges attribution uncertainty), we accounted for this by testing the stability of our results with respect to different versions of the SDM (Extended Data Fig. 10). For example, transitional industries (for example, the Levantine Mousterian or Lincombian–Ranisian–Jerzmanowician industries) received multiple attributions because they fit either *H. sapiens* or *H. neanderthalensis* in terms of toolmaker identity^{57,58}. A detailed explanation of this approach is provided as supplementary material for ref. ²¹ (<https://ars.els-cdn.com/content/image/1-s2.0-S2590332220304760-mmc1.pdf>).

A second source of uncertainty stems from dating. Although approximately 50% of the entries refer to the ^{14}C method (>90% of which are based on accelerator mass spectrometry), other dating methods such as electron spin resonance (14% of the sample), thermoluminescence (12%) and optically stimulated luminescence (12%) are less precise than radiocarbon dating. Nonetheless, multiple datings are present for individual fossil sites, even within a single stratigraphic layer at the site. To account for uncertainties in species attribution and age, we ran our analyses according to the four different approaches described below.

1. 1.

Multiple dates, tier 1. Only the core record, which excludes entries with uncertain species attributions, and all age estimates available for each archaeological layer are used. Multiple age estimates per layer are possible, and the age uncertainty for each is included in our analysis. This subdivision includes 3,060 data entries. Although the main analysis in our study is based on this case, we need to consider possible sampling biases due to the higher weights given to archaeological layers with multiple dates (Figs. [1–4](#) and Extended Data Fig. [10](#)).

2. 2.

Multiple dates, tier 2. The extended record, in which ambiguous species attributions are treated by randomly choosing among the possible candidate species, is used along with multiple age estimates (including uncertainties) per layer. This subdivision includes 3,245 (all) data entries (Extended Data Fig. [10](#)).

3. 3.

Single date, tier 1. Multiple age estimates for a single archaeological layer are combined in this approach to provide a minimum and maximum age for the layer. Each archaeological layer has only one entry, thereby eliminating possible sampling biases in the estimation of our CEM. This subdivision includes 1,567 data entries (Extended Data Fig. [10](#)).

4. 4.

Single date, tier 2. Age estimates for archaeological layers are treated as those in the single date, tier 1 category except that the extended record rather than the core record is used. This subdivision includes 1,652 data entries (Extended Data Fig. 10).

We acknowledge that our species subdivisions may be controversial and that these do not necessarily require constancy of morphology, habitat and behaviour. However, even though some species attributions such as *H. heidelbergensis* could be questioned, we remain confident that the majority of the record presents little challenge considering that 86% of the core data belong to the well-defined, widely accepted *H. neanderthalensis* or *H. sapiens* record and tool-making traditions. Thus, even though some species attributions might be considered invalid, widely accepted constraints are used. Clearly, to the best of current knowledge, 500,000-year-old remains in Africa can belong to either *H. sapiens* nor *H. habilis*⁵⁹, irrespective of whether the name *H. heidelbergensis* is considered appropriate. To further reduce uncertainties, we tested the robustness of our main findings with four alternative scenarios (Extended Data Fig. 10) for species attribution and dating and excluding uncertain and poorly dated species (for example, *Homo floresiensis*, *Homo naledi*, *Homo bodoensis*, *Homo longi* and Denisovans), which are restricted to too few fossil sites for which no climatic variability can possibly be ascertained or do not currently include any other locality or remains in their definition. The final species assignments used in our study should be interpreted here as plausible working hypotheses.

Mahalanobis CEM

To derive the CEM (Extended Data Fig. 5) that best characterizes the habitable conditions for hominins, we chose four key climatic variables: annual mean temperature and precipitation (T^*_{am} and P^*_{am} , respectively), annual minimum precipitation (P^*_{min}) and terrestrial NPP (N^*). Obtained as 1,000-year downscaled averages ($1^\circ \times 1^\circ$ horizontal resolution), these variables, which relate to physiological constraints for hominin survival and the availability of food resources, are combined as a four-dimensional climate environment vector $\mathbf{C}(t) = (T^*_{\text{am}}, P^*_{\text{am}}, P^*_{\text{min}}, N^*)$ with 2,000 values

in time (t) corresponding to 1,000-year (200-year) orbital (model) means from the 2Ma simulation. The fossil and archaeological data entries for the five individual hominin groups are described in the previous section.

Although our main analysis focuses on the multiple date, tier 1 case ([Methods](#) and Supplementary Table 1), the robustness of our results was tested against other ways of treating species and age model uncertainties (Extended Data Fig. 10). The data entries are represented by their longitude $\lambda_{j,i}$ and latitude $\varphi_{j,i}$ coordinates and the respective average age $t_{j,i}$ and age uncertainties $\Delta t_{j,i}$ with $i = 1, \dots, 5$ corresponding to the five hominin groups.

We defined the fossil state vector as $\mathbf{z}^i = (\lambda_{1,i}, \varphi_{1,i}, t_{1,i}, \Delta t_{1,i}, \dots, \lambda_{n,i}, \varphi_{n,i}, t_{n,i}, \Delta t_{n,i})$ with n_i representing the total number of fossils in each group during the past 2 million years. We then built the matrix D from the four-dimensional climate data subsampled at the fossil data sites and the corresponding nearest ages. Age uncertainties were considered through a Monte Carlo sampling method, which expanded the length of the overall data vector. We obtained D^i ($4 \times N_i$ matrix for each $i = 1, \dots, 5$) for each hominin group as $D^i = (T^*_{\text{am}}(\mathbf{z}^i), P^*_{\text{am}}(\mathbf{z}^i), P^*_{\text{min}}(\mathbf{z}^i), N^*(\mathbf{z}^i))$. We then calculated the Mahalanobis squared distance model^{[23](#)} for each group using $\zeta_i^2(D^i) = (D^i - \langle D^i \rangle)^T S^{-1} (D^i - \langle D^i \rangle)$, where $\langle \dots \rangle$ represents the sample mean value and S^{-1} is the inverse co-variance matrix obtained from the data. The Mahalanobis squared distances ζ_i were then transformed into a cumulative chi-squared distribution χ^2_{CDF} in the four-dimensional climate data space C . When using 4 degrees of freedom^{[23](#)}, the corresponding probability $H(C) = 1 - \chi^2_{\text{CDF}}(C)$ represents the likelihood of finding a fossil for a specific quadruplet within the four-dimensional climate data space in the HSM. We interpreted H as a probability, which we refer to as habitat suitability. Given the temporal evolution of C for every grid point of the downsampled $1^\circ \times 1^\circ$ data over the last 2 million years, we were able to calculate the spatiotemporal habitat suitability for each downsampled grid point (x,y,t) in the model as $H(x,y,t) = H(T^*_{\text{am}}(x,y,t), P^*_{\text{am}}(x,y,t), P^*_{\text{min}}(x,y,t), N^*(x,y,t))$. The stability of the HSM was tested by using different dimensionalities and combinations of climate parameters such as annual mean and seasonal range of temperature and precipitation and annual mean and minimum values of temperature and precipitation. The key conclusions of our study remained essentially unchanged. Moreover, we tested the

stability of our results against the omission of hominin sites with ambivalent original species attributions (multiple date, tier 2) and different treatment of archaeological ages (single date, tiers 1 and 2). The calculated $H(x,y,t)$ was qualitatively very similar for the four different cases (Extended Data Fig. 10). Therefore, our main conclusions remain robust with respect to uncertainties in species attribution and archaeological layer dating.

Random climate trajectory

To address the question of whether the actual climate trajectory influenced the distribution of fossil and archaeological data, we developed a CEM and HSM in which fossil data were assigned to randomly chosen climate states from the CESM1.2 simulation under the constraint that the climate range selected must overlap with the total fossil age range of the respective species. We randomized the time variability of the four-dimensional climate data vector (annual mean temperature, annual mean precipitation, minimum precipitation and NPP) while keeping the co-variability among the climate vector components, as well as the mean state, invariant. The original HSM (H), which is based on the real trajectory of CESM1.2, and the model (H_{scr}) that we trained from a scrambled trajectory were then compared. The time-averaged differences between the models for *H. sapiens*, *H. neanderthalensis* and *H. heidelbergensis* were then interpreted as an indication of whether the realistic climate evolution influenced the observed hominin distributions in space and time relative to a system that maintains its orbital climate co-variance and mean state (Extended Data Fig. 8a–c) but does not consider the exact time evolution of glacial–interglacial and orbital cycles. The time-averaged difference between $H(x,y,t)$ in the original HSM and $H_{\text{scr}}(x,y,t)$ in the HSM derived from time-randomized climate data was then tested at each grid point using a paired *t*-test.

Reporting summary

Further information on research design is available in the [Nature Research Reporting Summary](#) linked to this paper.

Data availability

The CESM1.2 data and the calculated hominin habitat suitability data are available on the climate data server at <https://climatedata.ibs.re.kr>. The database of hominin remains and artefacts used here is provided in Supplementary Table 1. The maps in Fig. 1 and Extended Data Figs. 7, 8 and 10 were generated using M_Map: a mapping package for MATLAB, version 1.4m, available at <http://www.eoas.ubc.ca/~rich/map.html>. The map in Fig. 4 was generated using the software Paraview, freely available at <https://www.paraview.org>.

Code availability

The MATLAB codes used to generate Figs. 1–3 will be shared on the climate data server at <https://climatedata.ibs.re.kr>. The CESM1.2 code is available at <https://www.cesm.ucar.edu/models/cesm1.2/>.

References

1. deMenocal, P. B. Climate and human evolution. *Science* **331**, 540–542 (2011).
2. Larrasoana, J. C., Roberts, A. P. & Rohling, E. J. Dynamics of green Sahara periods and their role in hominin evolution. *PLoS ONE* **8**, e76514 (2013).
3. Potts, R. Variability selection in hominid evolution. *Evol. Anthropol.* **7**, 81–96 (1998).
4. Levin, N. E., Quade, J., Simpson, S. W., Semaw, S. & Rogers, M. Isotopic evidence for Plio–Pleistocene environmental change at Gona, Ethiopia. *Earth Planet. Sci. Lett.* **219**, 93–110 (2004).
5. Dart, R. A. *Australopithecus africanus*: the man–ape of South Africa. *Nature* **115**, 195–199 (1925).
6. Dominguez-Rodrigo, M. Is the “savanna hypothesis” a dead concept for explaining the emergence of the earliest hominins? *Curr. Anthropol.* **55**, 59–81 (2014).

7. Timmermann, A. & Friedrich, T. Late Pleistocene climate drivers of early human migration. *Nature* **538**, 92–95 (2016).
8. Breeze, P. S. et al. Palaeohydrological corridors for hominin dispersals in the Middle East similar to 250–70,000 years ago. *Quatern. Sci. Rev.* **144**, 155–185 (2016).
9. Beyer, R. M., Krapp, M., Eriksson, A. & Manica, A. Climatic windows for human migration out of Africa in the past 300,000 years. *Nat. Commun.* **12**, 4889 (2021).
10. Groucutt, H. S. et al. Multiple hominin dispersals into southwest Asia over the past 400,000 years. *Nature* **597**, 376–380 (2021).
11. Pagani, L. et al. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538**, 238–242 (2016).
12. Potts, R. et al. Increased ecological resource variability during a critical transition in hominin evolution. *Sci. Adv.* **6**, eabc8975 (2020).
13. Potts, R. & Faith, J. T. Alternating high and low climate variability: the context of natural selection and speciation in Plio–Pleistocene hominin evolution. *J. Hum. Evol.* **87**, 5–20 (2015).
14. Tigchelaar, M. & Timmermann, A. Mechanisms rectifying the annual mean response of tropical Atlantic rainfall to precessional forcing. *Clim. Dynam.* **47**, 271–293 (2016).
15. Willeit, M., Ganopolski, A., Calov, R. & Brovkin, V. Mid-Pleistocene transition in glacial cycles explained by declining CO₂ and regolith removal. *Sci. Adv.* **5**, eaav7337 (2019).
16. Berger, A. Long-term variations of caloric insolation resulting from the Earth’s orbital elements. *Quatern. Res.* **9**, 139–167 (1978).
17. Lorenz, S. J. & Lohmann, G. Acceleration technique for Milankovitch type forcing in a coupled atmosphere–ocean circulation model: method and application for the Holocene. *Clim. Dynam.* **23**, 727–743 (2004).

18. Friedrich, T., Timmermann, A., Tigchelaar, M., Timm, O. E. & Ganopolski, A. Nonlinear climate sensitivity and its implications for future greenhouse warming. *Sci. Adv.* **2**, e1501923 (2016).
19. Friedrich, T. & Timmermann, A. Using late Pleistocene sea surface temperature reconstructions to constrain future greenhouse warming. *Earth Planet. Sci. Lett.* **530**, 115911 (2020).
20. Tierney, J. E. et al. Glacial cooling and climate sensitivity revisited. *Nature* **584**, 569–573 (2020).
21. Raia, P., Mondanaro, A., Melchionna, M., Di Febbraro, M. & Diniz-Filho, J. A. F. Past extinctions of *Homo* species coincided with increased vulnerability to climatic change. *One Earth* **3**, 1–11 (2020).
22. Mondanaro, A. et al. A major change in rate of climate niche envelope evolution during hominid history. *iScience* **23**, 101693 (2020).
23. Etherington, T. R. Mahalanobis distances and ecological niche modelling: correcting a chi-squared probability error. *PeerJ* **7**, e6678 (2019).
24. Farber, O. & Kadmon, R. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecol. Model.* **160**, 115–130 (2003).
25. Antón, S. Natural history of *Homo erectus*. *Am. J. Phys. Anthropol.* **S37**, 126–170 (2003).
26. Kaboth-Bahr, S. et al. Paleo-ENSO influence on African environments and early modern humans. *Proc. Natl Acad. Sci. USA* **118**, e2018277118 (2021).
27. Santonja, M., Perez-Gonzalez, A., Panera, J., Rubio-Jara, S. & Mendez-Quintas, E. The coexistence of Acheulean and Ancient Middle Palaeolithic techno-complexes in the Middle Pleistocene of the Iberian Peninsula. *Quatern. Int.* **411**, 367–377 (2016).

28. Arsuaga, J. L. et al. Neandertal roots: cranial and chronological evidence from Sima de los Huesos. *Science* **344**, 1358–1363 (2014).
29. Backwell, L. R. et al. New excavations at Border Cave, KwaZulu-Natal, South Africa. *J. Field Archaeol.* **43**, 417–436 (2018).
30. Grun, R. et al. Direct dating of Florisbad hominid. *Nature* **382**, 500–501 (1996).
31. Porat, N. et al. New radiometric ages for the Fauresmith industry from Kathu Pan, southern Africa: implications for the Earlier to Middle Stone Age transition. *J. Archaeol. Sci.* **37**, 269–283 (2010).
32. Chan, E. K. F. et al. Human origins in a southern African palaeo-wetland and first migrations. *Nature* **575**, 185–189 (2019).
33. Schlebusch, C. M. et al. Khoe-San genomes reveal unique variation and confirm the deepest population divergence in *Homo sapiens*. *Mol. Biol. Evol.* **37**, 2944–2954 (2020).
34. Meyer, M. et al. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* **531**, 504–507 (2016).
35. Meyer, M. et al. A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* **505**, 403–406 (2014).
36. Foley, R. A. Mosaic evolution and the pattern of transitions in the hominin lineage. *Philos. Trans. R. Soc. B* **371**, 20150244 (2016).
37. Amos, W. Correlated and geographically predictable Neanderthal and Denisovan legacies are difficult to reconcile with a simple model based on inter-breeding. *R. Soc. Open Sci.* **8**, 201229 (2021).
38. Slon, V. et al. The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* **561**, 113–116 (2018).
39. Jacobs, G. S. et al. Multiple deeply divergent Denisovan ancestries in Papuans. *Cell* **177**, 1010–1021 (2019).

40. Shultz, S. & Maslin, M. Early human speciation, brain expansion and dispersal influenced by African climate pulses. *PLoS ONE* **8**, e76750 (2013).
41. Scerri, E. M. L. et al. Did our species evolve in subdivided populations across Africa, and why does it matter? *Trends Ecol. Evol.* **33**, 582–594 (2018).
42. Holt, R. D. Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. *Proc. Natl Acad. Sci. USA* **106**, 19659–19665 (2009).
43. Hua, X. & Wiens, J. J. How does climate influence speciation? *Am. Nat.* **182**, 1–12 (2013).
44. Prufer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
45. Meehl, G. A. et al. Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Sci. Adv.* **6**, eaba1981 (2020).
46. Masson-Delmotte, V. et al. (eds) IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (Cambridge University Press, in the press).
47. Ganopolski, A., Calov, R. & Claussen, M. Simulation of the last glacial cycle with a coupled climate ice-sheet model of intermediate complexity. *Clim. Past* **6**, 229–244 (2010).
48. Timmermann, A. et al. Modeling obliquity and CO₂ effects on Southern Hemisphere climate during the past 408 ka. *J. Clim.* **27**, 1863–1875 (2014).
49. Ganopolski, A. & Rahmstorf, S. Rapid changes of glacial climate simulated in a coupled climate model. *Nature* **409**, 153–158 (2001).

50. Friedrich, T. et al. The mechanism behind internally generated centennial-to-millennial scale climate variability in an Earth system model of intermediate complexity. *Geosci. Model Dev.* **3**, 377–389 (2010).
51. Vettoretti, G. & Peltier, W. R. Fast physics and slow physics in the nonlinear Dansgaard–Oeschger relaxation oscillation. *J. Clim.* **31**, 3423–3449 (2018).
52. Brovkin, V., Ganopolski, A. & Svirezhev, Y. A continuous climate–vegetation classification for use in climate–biosphere studies. *Ecol. Model.* **101**, 251–261 (1997).
53. Del Grosso, S. et al. Global potential net primary production predicted from vegetation class, precipitation, and temperature. *Ecology* **89**, 2117–2126 (2008).
54. Ainsworth, E. A. & Rogers, A. The response of photosynthesis and stomatal conductance to rising CO₂: mechanisms and environmental interactions. *Plant Cell Environ.* **30**, 258–270 (2007).
55. Hublin, J. et al. New fossils from Jebel Irhoud, Morocco and the pan-African origin of *Homo sapiens*. *Nature* **546**, 289–292 (2017).
56. Profico, A. et al. Filling the gap. Human cranial remains from Gombore II (Melka Kunture, Ethiopia; ca. 850 ka) and the origin of *Homo heidelbergensis*. *J. Anthropol. Sci.* **94**, 1–24 (2016).
57. Meignen, L. in *The Lower and Middle Palaeolithic in the Middle East and Neighbouring Regions* (eds Le Tensorer, J. M. et al.) 85–100 (University of Liège Press, 2011).
58. Krajcarz, M. T. et al. Towards a chronology of the Jerzmanowician—a new series of radiocarbon dates from Nietoperzowa Cave (Poland). *Archaeometry* **60**, 383–401 (2018).
59. de la Torre, I. et al. New excavations in the MNK Skull site, and the last appearance of the Oldowan and *Homo habilis* at Olduvai Gorge,

Tanzania. *J. Anthropol. Archaeol.* **61**, 101255 (2021).

60. Herbert, T. D., Peterson, L. C., Lawrence, K. T. & Liu, Z. H. Tropical ocean temperatures over the past 3.5 million years. *Science* **328**, 1530–1534 (2010).
61. Clemens, S. C., Prell, W. L., Sun, Y. B., Liu, Z. Y. & Chen, G. S. Southern Hemisphere forcing of Pliocene $\delta^{18}\text{O}$ and the evolution of Indo–Asian monsoons. *Paleoceanography* **23**, PA4210 (2008).
62. Lyons, R. P. et al. Continuous 1.3-million-year record of East African hydroclimate, and implications for patterns of evolution and biodiversity. *Proc. Natl Acad. Sci. USA* **112**, 15568–15573 (2015).
63. Jouzel, J. et al. Orbital and millennial Antarctic climate variability over the past 800,000 years. *Science* **317**, 793–796 (2007).
64. Tiedemann, R., Sarnthein, M. & Shackleton, N. J. Astronomic timescale for the Pliocene Atlantic $\delta^{18}\text{O}$ and dust flux records of Ocean Drilling Program Site-659. *Paleoceanography* **9**, 619–638 (1994).
65. DeMenocal, P. B. African climate change and faunal evolution during the Pliocene–Pleistocene. *Earth Planet. Sci. Lett.* **220**, 3–24 (2004).

Acknowledgements

A.T., K.-S.Y., D.L. and E.Z. received funding from IBS under IBS-R028-D1. M.W. acknowledges funding from DFG GA 1202/2-1, BMBF PalMod. C.Z. was supported by the Swiss Platform for Advanced Scientific Computing, grant F-1516 74105-03-01. The 2Ma CESM1.2 simulations were conducted on the ICCP/IBS supercomputer Aleph, a Cray XC50-LC system.

Author information

Affiliations

1. Center for Climate Physics, Institute for Basic Science, Busan, South Korea

Axel Timmermann, Kyung-Sook Yun, Jiaoyang Ruan, Elke Zeller & Danielle Lemmon

2. Pusan National University, Busan, South Korea

Axel Timmermann, Kyung-Sook Yun, Jiaoyang Ruan, Elke Zeller & Danielle Lemmon

3. DiSTAR, Università di Napoli Federico II, Monte Sant'Angelo, Naples, Italy

Pasquale Raia

4. DST, Università degli Studi di Firenze, Florence, Italy

Alessandro Mondanaro

5. Anthropological Institute, University of Zurich, Zurich, Switzerland

Christoph Zollikofer

6. Department of Informatics, University of Zurich, Zurich, Switzerland

Marcia Ponce de León

7. Potsdam Institute for Climate Impact Research, Potsdam, Germany

Matteo Willeit & Andrey Ganopolski

Contributions

A.T. designed the study, developed and coded the HSM, downscaled the CESM1.2 data, prepared the figures and wrote the initial draft of the manuscript. K.-S.Y. conducted the CESM1.2 2Ma simulation. P.R. and A.M. prepared the extended version of the fossil and archaeological dataset. All

other authors contributed to the interpretation of the data and the writing of the manuscript.

Corresponding author

Correspondence to [Axel Timmermann](#).

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Mark Maslin, Michael Petraglia and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

[Extended Data Fig. 1 Model/data comparison.](#)

a) precession index (blue) over last 2 million years and eccentricity timeseries (orange); h) reconstructed⁶⁰ (blue) and simulated (orange, CESM1.2) tropical sea surface temperatures; i) hydroclimate reconstruction from Lake Malawi⁶¹ (blue, southeastern Africa) and simulated precipitation (orange); j) geochemical proxy (Ba/Al ratio) representing variations of the

East Asian Summer Monsoon⁶² (EASM) (blue) and simulated rainfall from CESM1.2 (orange) for the location 116.2°E, 19.3°N (mm/day).

Extended Data Fig. 2 Model/data comparison.

a) simulated global temperatures anomalies (relative to pre-industrial conditions) (blue) and palaeo-climate reconstruction¹⁸; **b)** simulated regional temperature in Antarctica (5°x5° degree average over Dome C location) (blue) with deuterium based temperature reconstruction from EPICA Dome C⁶³; **c)** calculated net primary production for northwestern Sahara (blue) with logarithm of dust flux from marine core ODP659⁶⁴; **d)** calculated net primary production for Arabian Peninsula (blue) with logarithm of terrigenous dust concentration from marine core ODP721/722⁶⁵.

Extended Data Fig. 3 Surface temperature near sites relevant for interpretation of hominin evolution.

from a-l: Lake Turkana (3°N, 36°E), Lake Malawi (12°S, 34°E), Southern Africa (25°S, 25°E), Blombos cave (34°S, 21°E), Denmark (56°N, 9°E), Peçtera cu Oase (45°N, 22°E), Sima de los Huesos (42°N, 5°W), Shkul cave (33°N, 35°E), Sinai (30°N, 34°E), Jebel Irhoud (34°N, 4°W), Lake Challa (3°S, 38°E), Denisova, cave (52°N, 84°E). The timeseries represents a spatial mean of 7°x7° around the center location.

Extended Data Fig. 4 Precipitation near sites relevant for interpretation of hominin evolution.

from a-l: Lake Turkana (3°N, 36°E), Lake Malawi (12°S, 34°E), Southern Africa (25°S, 25°E), Blombos cave (34°S, 21°E), Denmark (56°N, 9°E), Peçtera cu Oase (45°N, 22°E), Sima de los Huesos (42°N, 5°W), Shkul cave (33°N, 35°E), Sinai (30°N, 34°E), Jebel Irhoud (34°N, 4°W), Lake Challa (3°S, 38°E), Denisova, cave (52°N, 84°E). The timeseries represents a spatial mean of 7°x7° around the center location.

Extended Data Fig. 5 Schematic of Climate envelope model (CEM) and Habitat suitability model (HSM) set-up.

The orbital scale transient 2Ma simulation of CESM1.2 is conducted by using CO₂ and ice-sheet forcings from an intermediate complexity model simulation¹⁵ and orbital forcing from astronomical estimates¹⁶. The simulated surface temperatures and precipitation from 2Ma on a ~3.75-degree horizontal grid are then downscaled to a 1x1 degree horizontal grid by including lapse-rate corrected topographic features. Net Primary Production is calculated from empirical parameterizations of the downscaled temperature and precipitation fields (Extended Data Fig. [8d–f](#)). Using an extended fossil and archaeological database (Supplementary Table [1](#)) in combination with the downscaled annual mean temperatures, annual mean rainfall, minimum rainfall and net primary productivity, a statistical Mahalanobis distance-based climate envelope model (CEM) is derived. The model is then forced for every land grid point on a 1x1 degree grid with the temporal evolution of the downscaled climate variables to obtain the temporal evolution of the habitat suitability (HSM) for each of the 5 hominin groups and at every grid point. The impact of resolution on key features in simulated net primary production is further illustrated in Extended Data Fig. [8d–f](#).

Extended Data Fig. 6 Orbital effects on regional habitability in Eurasia.

Timeseries for precession (blue) and eccentricity (a) and simulated regional habitat suitability at selected sites of archaeological interest for *Homo habilis*, *Homo ergaster*, *Homo erectus*, *Homo heidelbergensis*, *Homo neanderthalensis*, and *Homo sapiens*. The centre locations of a 5°x5° averaging area are selected as: **b**) Neanderthal (51°N, 7°E); **c**) Shkul cave (33°N, 35°E); **d**) Denisova Cave (51°N, 85°E).

Extended Data Fig. 7 Potential hominin overlap regions.

a) Shading indicates the relative occurrence frequency (relative to the entire 2-million-year simulation history) of when the habitat suitability for *Homo sapiens* and *Homo heidelbergensis* at a grid point both exceed 0.3 and exhibit a difference of less than 0.1. This constraint is considered here as a measure of habitat similarity during overlap times. Circles show the respective fossil and archaeological sites from the overlap periods. Cyan

boxes highlight areas, for which there are mixed fossils available and high habitat similarity. These areas are potential areas for speciation events or gradual evolutionary transformations. **b**) same as a), but for *Homo neanderthalensis* and *Homo heidelbergensis*; **c**) same as a) but for *Homo erectus* and *Homo heidelbergensis*; **d**) same as a) but for *Homo erectus* and *Homo ergaster/habilis*; **e**) same as a) but for *Homo ergaster/habilis* and *Homo heidelbergensis*. The size of the circles scales with the age of the fossils.

Extended Data Fig. 8 Climate effects on time-mean habitat suitability of *Homo sapiens*, *Homo neanderthalensis*, *Homo heidelbergensis*.

a–c) difference between time-averaged habitat suitability estimated from original climate envelope model H and climate envelope model H_{scr} calculated from vector-randomized climate variables ([Methods](#)) for 3 hominin groups. This type of time randomization ([Methods](#)) maintains the time-mean and climate co-variance among the 4-dimensional climate input vector. Results show the effect of the original orbital-scale climate trajectory (relative to a randomized climate trajectory) on hominin habitat suitability. Only grid points with $p < 0.05$ are shown in colors based on a paired t-test using $H(x,y,t)$ and $H_{scr}(x,y,t)$. **d–f)** Illustration of late-Holocene altitude-corrected downscaling of net primary production: Left: Simulated NPP ($\text{gC/m}^2/\text{year}$) for original T31 atmosphere resolution, obtained from empirical NPP model using 1000-year late Holocene average of total precipitation and surface temperatures simulated by the CESM1.2 2Ma experiment; middle: same as left, but using altitude-corrections for temperature and precipitation as downscaling onto a 1x1 degree target grid, showing the emergence of key topographic features in Africa. This resolution was chosen in the calculations of the climate envelope model; right: for illustration, same as middle but for a 0.25x0.25 degree horizontal target grid. The qualitative gain in terms of regional details from T31 to 1x1 degree resolution outweighs the additional gain going from 1x1 degree resolution to a 0.25x0.25 degree grid.

Extended Data Fig. 9 Net primary production near sites relevant for interpretation of hominin evolution.

from a-l: Lake Turkana (3°N, 36°E), Lake Malawi (12°S, 34°E), Southern Africa (25°S, 25°E), Blombos cave (34°S, 21°E), Denmark (56°N, 9°E), Peçtera cu Oase (45°N, 22°E), Sima de los Huesos (42°N, 5°W), Shkul cave (33°N, 35°E), Sinai (30°N, 34°E), Jebel Irhoud (34°N, 4°W), Lake Challa (3°S, 38°E), Denisova, cave (52°N, 84°E). The timeseries represents a spatial mean of 7°x7° degree around the center location.

Extended Data Fig. 10 Robustness with respect to age and species assignment uncertainties.

Upper panels, same as for Fig. 1c, but for the 4 different cases single date-tier 1,2 and multi date-tier 1,2 ([Methods](#)). Multi-date tier 1 is the standard case used in our analysis. Lower panels, same as Fig. 3b-d, but calculated also for the 4 cases of assessing species and age uncertainties.

Supplementary information

Reporting Summary

Peer Review File

Supplementary Table 1

SDM including raw data (tab 1, including references), data for the multiple date, tier 1 and 2 cases (tabs 2 and 3, respectively) and data for the single date, tier 1 and 2 cases (tabs 4 and 5, respectively). The file presents an extended version of ref. [21](#), focusing on the period from 2 Ma to 30 ka excluding Australia and the Americas. Fossils for *H. floresiensis*, *H. naledi*, *H. longi* and Denisovans were excluded owing to the limited spatiotemporal coverage.

Supplementary Table 2

Simulation chunks of the 2Ma transient simulation. The first column represents the initial condition (in orbital years, including orbital acceleration of factor 5); the second column represents the length of the chunk in unaccelerated model years; and the third column shows the total length of each simulation, including an overlap period into the next chunk.

Supplementary Video 1

Animation showing time evolution of simulated habitat suitability for *H. sapiens* from 315 to 31 ka. The habitat suitability captures a probability range of 0 (habitat unsuitable) to 1 (habitat extremely suitable). Notably, the habitat suitability, which quantifies the potential suitability of climatic conditions needed to support life for a given hominin species, does not necessarily have to match with the locations in which fossils or archaeological artefacts have been found.

Supplementary Video 2

Animation showing time evolution of simulated habitat suitability for *H. neanderthalensis* from 330 to 33 ka. The habitat suitability captures a probability range of 0 (habitat unsuitable) to 1 (habitat extremely suitable). Notably, the habitat suitability, which quantifies the potential suitability of climatic conditions needed to support life for a given hominin species, does not necessarily have to match with the locations in which fossils or archaeological artefacts have been found.

Supplementary Video 3

Animation showing time evolution of simulated habitat suitability for *H. heidelbergensis* from 875 to 225 ka. The habitat suitability captures a probability range of 0 (habitat unsuitable) to 1 (habitat extremely suitable). Notably, the habitat suitability, which quantifies the potential suitability of climatic conditions needed to support life for a given hominin species, does not necessarily have to match with the locations in which fossils or archaeological artefacts have been found.

Supplementary Video 4

Animation showing time evolution of simulated habitat suitability for *H. erectus* from 1,800 to 60 ka. The habitat suitability captures a probability range of 0 (habitat unsuitable) to 1 (habitat extremely suitable). Notably, the habitat suitability, which quantifies the potential suitability of climatic conditions needed to support life for a given hominin species, does not necessarily have to match with the locations in which fossils or archaeological artefacts have been found.

Supplementary Video 5

Animation showing time evolution of simulated habitat suitability for archaic African *Homo* (comprised of *H. ergaster* and *H. habilis*) from 2,000 to 869 ka. The habitat suitability captures a probability range from 0 (habitat unsuitable) to 1 (habitat extremely suitable). Notably, the habitat suitability, which quantifies the potential suitability of climatic conditions needed to support life for a given hominin species, does not necessarily have to match with the locations in which fossils or archaeological artefacts have been found.

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Reprints and Permissions

About this article

Cite this article

Timmermann, A., Yun, KS., Raia, P. *et al.* Climate effects on archaic human habitats and species successions. *Nature* **604**, 495–501 (2022).
<https://doi.org/10.1038/s41586-022-04600-9>

- Received: 04 July 2021
- Accepted: 01 March 2022
- Published: 13 April 2022
- Issue Date: 21 April 2022
- DOI: <https://doi.org/10.1038/s41586-022-04600-9>

Share this article

Anyone you share the following link with will be able to read this content:

[Get shareable link](#)

Sorry, a shareable link is not currently available for this article.

[Copy to clipboard](#)

Provided by the Springer Nature SharedIt content-sharing initiative

Further reading

- [**A lengthy look at climate and its role in hominin evolution**](#)
 - Michael D. Petraglia

Nature (2022)

A lengthy look at climate and its role in hominin evolution

- Michael D. Petraglia

Nature News & Views 13 Apr 2022

This article was downloaded by **calibre** from <https://www.nature.com/articles/s41586-022-04600-9>

| [Section menu](#) | [Main menu](#) |

[Download PDF](#)

- Article
- Open Access
- [Published: 13 April 2022](#)

Somatic mutation rates scale with lifespan across mammals

- [Alex Cagan](#) [ORCID: orcid.org/0000-0002-7857-4771](#)¹,
¹_{na1},
- [Adrian Baez-Ortega](#)¹_{na1},
- [Natalia Brzozowska](#)¹,
- [Federico Abascal](#) [ORCID: orcid.org/0000-0002-6201-1587](#)¹,
- [Tim H. H. Coorens](#) [ORCID: orcid.org/0000-0002-5826-3554](#)¹,
- [Mathijs A. Sanders](#)^{1,2},
- [Andrew R. J. Lawson](#) [ORCID: orcid.org/0000-0003-3592-1005](#)¹,
- [Luke M. R. Harvey](#)¹,
- [Shriram Bhosle](#) [ORCID: orcid.org/0000-0002-7638-2899](#)¹,
- [David Jones](#) [ORCID: orcid.org/0000-0002-0407-0386](#)¹,
- [Raul E. Alcantara](#)¹,
- [Timothy M. Butler](#) [ORCID: orcid.org/0000-0001-5803-1035](#)¹,
- [Yvette Hooks](#)¹,
- [Kirsty Roberts](#)¹,
- [Elizabeth Anderson](#)¹,
- [Sharna Lunn](#)¹,
- [Edmund Flach](#)³,
- [Simon Spiro](#) [ORCID: orcid.org/0000-0002-9621-2192](#)³,
- [Inez Januszczak](#)^{3,4},
- [Ethan Wrigglesworth](#)³,
- [Hannah Jenkins](#)³,
- [Tilly Dallas](#)³,
- [Nic Masters](#)³,

- [Matthew W. Perkins⁵](#),
- [Robert Deaville⁵](#),
- [Megan Druce^{6,7}](#),
- [Ruzhica Bogeska^{6,7}](#),
- [Michael D. Milsom](#) [ORCID: orcid.org/0000-0002-3567-254X^{6,7}](#),
- [Björn Neumann^{8,9}](#),
- [Frank Gorman¹⁰](#),
- [Fernando Constantino-Casas¹⁰](#),
- [Laura Peachey^{10,11}](#),
- [Diana Bochynska](#) [ORCID: orcid.org/0000-0003-0219-8818^{10,12}](#),
- [Ewan St. John Smith](#) [ORCID: orcid.org/0000-0002-2699-1979¹³](#),
- [Moritz Gerstung](#) [ORCID: orcid.org/0000-0001-6709-963X¹⁴](#),
- [Peter J. Campbell](#) [ORCID: orcid.org/0000-0002-3921-0510¹](#),
- [Elizabeth P. Murchison](#) [ORCID: orcid.org/0000-0001-7462-8907¹⁰](#),
- [Michael R. Stratton](#) [ORCID: orcid.org/0000-0001-6035-153X¹](#) &
- [Iñigo Martincorena](#) [ORCID: orcid.org/0000-0003-1122-4416¹](#)

[Nature](#) volume **604**, pages 517–524 (2022)

- 70k Accesses
- 1 Citations
- 1252 Altmetric
- [Metrics details](#)

Subjects

- [Ageing](#)
- [Cancer genomics](#)
- [Comparative genomics](#)
- [Genomics](#)

Abstract

The rates and patterns of somatic mutation in normal tissues are largely unknown outside of humans^{1,2,3,4,5,6,7}. Comparative analyses can shed light on the diversity of mutagenesis across species, and on long-standing hypotheses about the evolution of somatic mutation rates and their role in cancer and ageing. Here we performed whole-genome sequencing of 208 intestinal crypts from 56 individuals to study the landscape of somatic mutation across 16 mammalian species. We found that somatic mutagenesis was dominated by seemingly endogenous mutational processes in all species, including 5-methylcytosine deamination and oxidative damage. With some differences, mutational signatures in other species resembled those described in humans⁸, although the relative contribution of each signature varied across species. Notably, the somatic mutation rate per year varied greatly across species and exhibited a strong inverse relationship with species lifespan, with no other life-history trait studied showing a comparable association. Despite widely different life histories among the species we examined—including variation of around 30-fold in lifespan and around 40,000-fold in body mass—the somatic mutation burden at the end of lifespan varied only by a factor of around 3. These data unveil common mutational processes across mammals, and suggest that somatic mutation rates are evolutionarily constrained and may be a contributing factor in ageing.

Main

Somatic mutations accumulate in healthy cells throughout life. They underpin the development of cancer⁹ and, for decades, have been speculated to contribute to ageing^{10,11,12}. Directly studying somatic mutations in normal tissues has been challenging owing to the difficulty of detecting mutations present in single cells or small clones in a tissue. Only recent technological developments, such as in vitro expansion of single cells into colonies^{13,14}, microdissection of histological units^{8,15}, single-cell sequencing^{16,17} or single-molecule sequencing¹⁸, are beginning to enable the study of somatic mutation in normal tissues.

Over the last few years, studies in humans have started to provide a detailed understanding of somatic mutation rates and the contribution of endogenous

and exogenous mutational processes across normal tissues^{8,13,14,19,20}. These studies are also revealing how, as we age, some human tissues are colonized by mutant cells that contain cancer-driving mutations, and how this clonal composition changes with age and disease. With the exception of some initial studies, far less is known about somatic mutation in other species^{1,2,3,4,5,6,7}. Yet, comparative analyses of somatic mutagenesis would shed light on the diversity of mutagenic processes across species, and on long-standing questions regarding the evolution of somatic mutation rates and their role in cancer and ageing.

A decades-long hypothesis on the evolution of somatic mutation rates pertains to the relationship between body mass and cancer risk. Some models predict that the risk of cancer should increase proportionally to the number of cells at risk of transformation. However, there appears to be no correlation between body mass and cancer risk across species^{21,22}. This observation, known as Peto's paradox, suggests that the evolution of larger body sizes is likely to require the evolution of stronger cancer suppression mechanisms^{23,24}. Whether evolutionary reduction of cancer risk across species is partly achieved by a reduction of somatic mutation rates remains unknown.

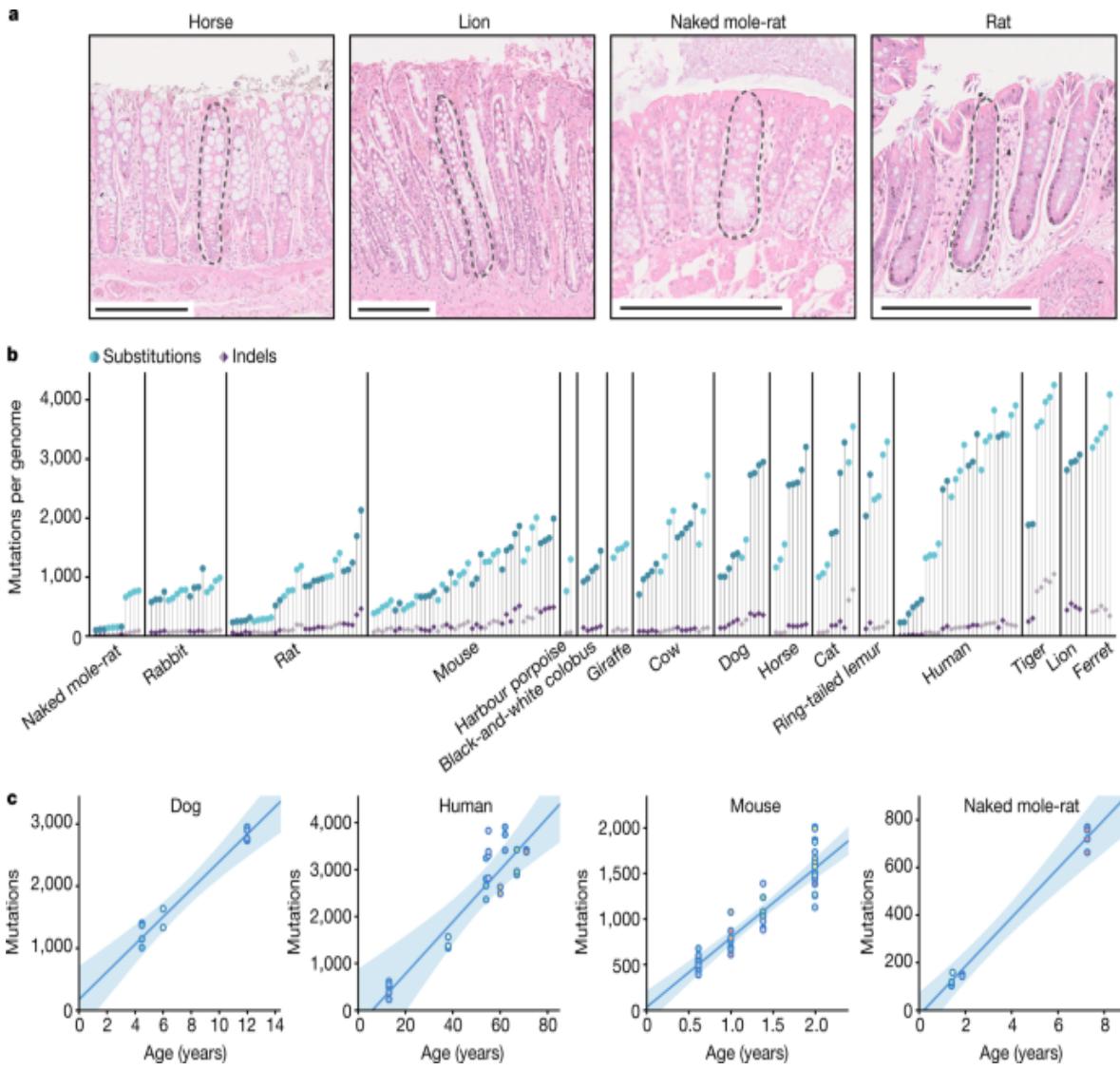
A second long-standing hypothesis on the evolution of somatic mutation rates relates to the proposed role of somatic mutations in ageing. Multiple forms of molecular damage, including somatic mutations, telomere attrition, epigenetic drift and loss of proteostasis, have been proposed to contribute to ageing, but their causal roles and relative contributions remain debated^{25,26}. Evolutionary theory predicts that species will evolve protection or repair mechanisms against life-threatening damage to minimize death from intrinsic causes, but that selection is too weak to delay ageing far beyond the typical life expectancy of an organism in the wild (Supplementary Note 1). If somatic mutations contribute to ageing, theory predicts that somatic mutation rates may inversely correlate with lifespan across species^{27,28}. This prediction has remained largely untested owing to the difficulty of measuring somatic mutation rates across species.

Detection of somatic mutations across species

The study of somatic mutations with standard whole-genome sequencing requires isolating clonal groups of cells recently derived from a single cell^{8,13,14}. To study somatic mutations across a diverse set of mammals, we isolated 208 individual intestinal crypts from 56 individuals across 16 species with a wide range of lifespans and body sizes: black-and-white colobus monkey, cat, cow, dog, ferret, giraffe, harbour porpoise, horse, human, lion, mouse, naked mole-rat, rabbit, rat, ring-tailed lemur and tiger (Supplementary Table 1). We chose intestinal crypts for several reasons. First, they are histologically identifiable units that line the epithelium of the colon and small intestine and are amenable to laser microdissection. Second, human studies have confirmed that individual crypts become clonally derived from a single stem cell and show a linear accumulation of mutations with age, which enables the estimation of somatic mutation rates through genome sequencing of single crypts⁸. Third, in most human crypts, most somatic mutations are caused by endogenous mutational processes common to other tissues, rather than by environmental mutagens^{8,18}.

A colon sample was collected from each individual, with the exception of a ferret from which only a small intestine sample was available. This sample was included because results in humans have shown that the mutation rates of colorectal and small intestine epithelial stem cells are similar^{14,20} (Extended Data Fig. 1). We then used laser microdissection on histological sections to isolate individual crypts for whole-genome sequencing with a low-input library preparation method²⁹ (Fig. 1a, Extended Data Fig. 2, Supplementary Table 2), with the exception of human crypts, for which sequencing data were obtained from a previous study⁸. A bioinformatic pipeline was developed to call somatic mutations robustly in all these species despite the variable quality of their genome assemblies (Methods). The distribution of variant allele fractions of the mutations detected in each crypt confirmed that crypts are clonal units in all species, enabling the study of somatic mutation rates and signatures (Extended Data Fig. 3).

Fig. 1: Somatic mutation burden in mammalian colorectal crypts.



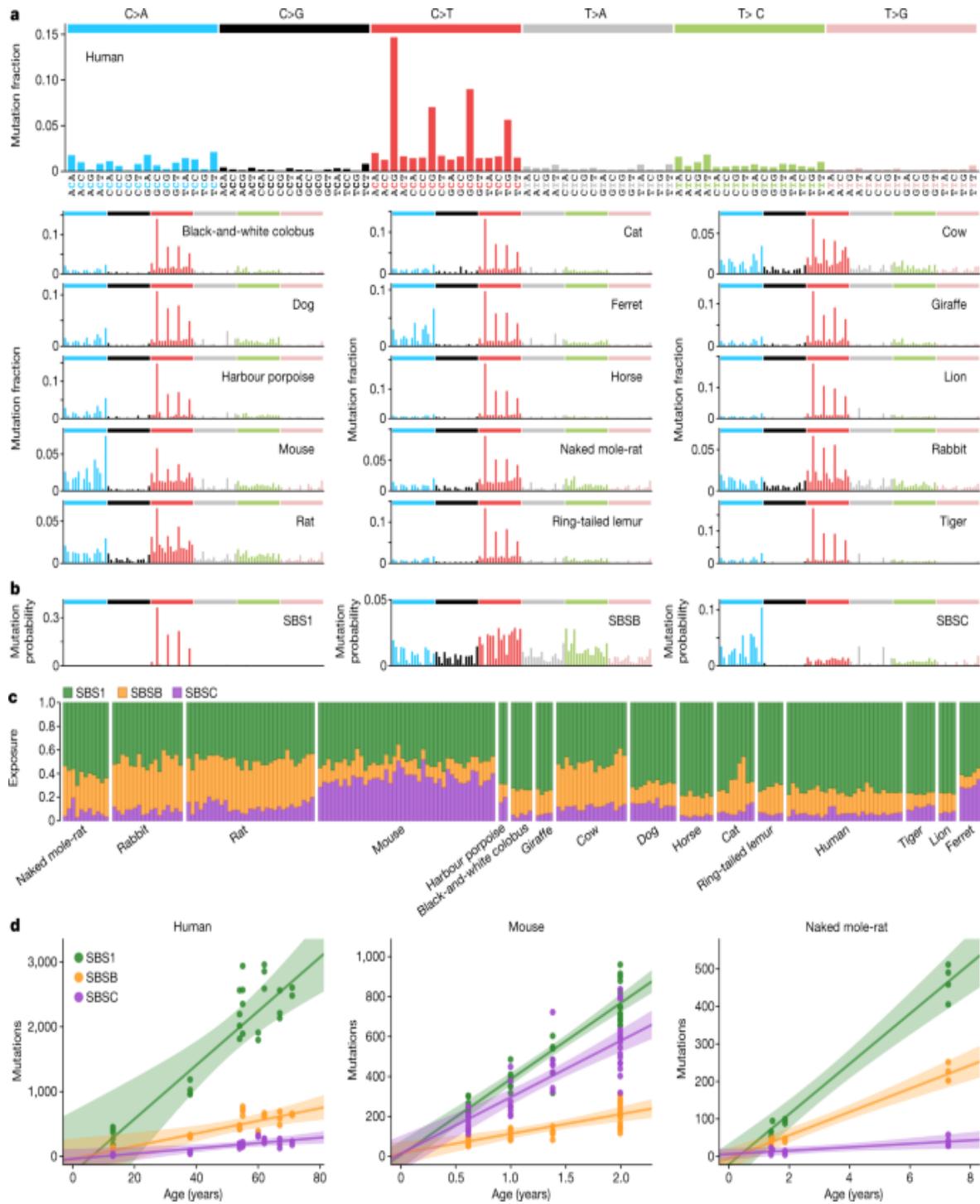
a, Histology images of colon samples from horse, lion, naked mole-rat and rat, with one colorectal crypt marked in each. Scale bars, 250 μm . **b**, Burden of somatic substitutions and indels per diploid genome in each colorectal crypt sample (corrected for the size of the analysable genome). Samples are grouped by individual, with samples from the same individual coloured in the same shade. Species, and individuals within each species, are sorted by mean mutation burden. **c**, Linear regression of somatic substitution burden (corrected for analysable genome size) on individual age for dog, human, mouse and naked mole-rat samples. Samples from the same individual are shown in the same colour. Regression was performed using mean mutation burdens per individual. Shaded areas indicate 95% confidence intervals of the regression line.

We found substantial variation in the number of somatic single-base substitutions across species and across individuals within each species (Fig. 1b). For five species with samples from multiple individuals (dog, human, mouse, naked mole-rat and rat), linear regression confirmed a clear accumulation of somatic mutations with age (Fig. 1c, Extended Data Fig. 4, Supplementary Table 3). All linear regressions were also consistent with a non-significant intercept. This resembles observations in humans²⁰ and suggests that the time required for a single stem cell to drift to fixation within a crypt is a small fraction of the lifespan of a species. This facilitates the estimation of somatic mutation rates across species by dividing the number of mutations in a crypt by the age of the individual (Supplementary Table 4). The number of somatic insertions and deletions (indels) was consistently lower than that of substitutions in all crypts (Fig. 1b), in agreement with previous findings in humans⁸.

Mutational signatures across mammals

Somatic mutations can be caused by multiple mutational processes, involving different forms of DNA damage and repair. Different processes cause characteristic frequencies of base substitution types and indels at different sequence contexts, often referred to as mutational signatures, which can be inferred from mutation data³⁰. Across species, the mutational spectra showed clear similarities, with a dominance of cytosine-to-thymine (C>T) substitutions at CpG sites, as observed in human colon, but with considerable variation in the frequency of other substitution types (Fig. 2a). To quantify the contribution of different mutational processes to the observed spectra, we applied mutational signature decomposition^{8,30}. We used a Bayesian model to infer mutational signatures de novo, while accounting for differences in genome sequence composition across species, and using the COSMIC human signature SBS1 (C>T substitutions at CpG sites) as a fixed prior to ensure its complete deconvolution³¹ (Methods). This approach identified two signatures beyond SBS1, labelled SBSB and SBSC, which resemble COSMIC human signatures SBS5 and SBS18, respectively (cosine similarities 0.93 and 0.91) (Fig. 2b).

Fig. 2: Mutational processes in the mammalian colon.



a, Mutational spectra of somatic substitutions in each species. The *x* axis shows 96 mutation types on a trinucleotide context, coloured by base substitution type. **b**, Mutational signatures inferred from (SBSB, SBSC) or fitted to (SBS1) the species mutational spectra shown in **a**, and normalized to the human genome trinucleotide frequencies. The *y* axis shows mutation

probability. **c**, Estimated contribution of each signature to each sample. Samples are arranged horizontally as in Fig. 1b. **d**, Linear regression of signature-specific mutation burdens (corrected for analysable genome size) on individual age for human, mouse and naked mole-rat samples. Regression was performed using mean mutation burdens per individual. Shaded areas indicate 95% confidence intervals of the regression line.

This analysis suggests that the same three signatures that dominate somatic mutagenesis in the human colon are dominant in other mammals: SBS1, which is believed to result from the spontaneous deamination of 5-methylcytosine^{8,32}; SSB (SBS5), a common signature across human tissues that may result from endogenous damage and repair^{18,33}; and SBSC (SBS18), which is dominated by C>A substitutions and attributed to oxidative DNA damage³⁰. Signature SBSC contains a minor component of T>A substitutions (resembling COSMIC SBS34), which appear to be the result of DNA polymerase slippage at the boundaries between adjacent adenine and thymine homopolymer tracts, but could also reflect assembly errors at those sites³³. Although all of the species that we examined shared the three mutational signatures, their contributions varied substantially across species (Fig. 2c). SBSC was particularly prominent in mouse and ferret, and the ratio of SBS1 to SBSB/5 varied from approximately 1.2 in rat or rabbit to 6.4 in tiger. In several species with data from multiple individuals, separate linear regressions for each signature confirmed that mutations from all three signatures accumulate with age (Fig. 2d, Extended Data Fig. 5).

Although signature deconvolution identified three signatures that are active across species, we noticed some differences in the mutational profile of signature SSB among species. To investigate this further, we inferred independent versions of SSB from each species, while accounting for differences in genome sequence composition (Methods). This revealed inter-species variability in the mutational profile of this signature, particularly in the C>T component (Extended Data Fig. 6). Species-specific versions of SSB showed different similarities to the related human signatures SBS5 and SBS40. For example, SSB inferred from the human data showed a stronger similarity with the reference human signature SBS5 (cosine similarities with SBS5 and SBS40: 0.93 and 0.84), whereas SSB from

rabbit more closely resembled the reference human signature SBS40 (0.87 and 0.91). These observations are consistent with the hypothesis that SBS5 and SBS40 result from a combination of correlated mutational processes, with some variation across human tissues^{18,33} and across species.

Analysis of the indel mutational spectra revealed a dominance of the human indel signatures ID1 and ID2, which are characterized by single-nucleotide indels at A/T homopolymers, and probably caused by strand slippage during DNA replication³⁰ (Extended Data Fig. 7a). The ratio of insertions (ID1) to deletions (ID2) appears to vary across species, possibly reflecting a differential propensity for slippage of the template and nascent DNA strands³⁰. In addition, the indel spectra suggest a potential contribution of signature ID9 (the aetiology of which remains unknown) to human, colobus, cow, giraffe and rabbit. Analysis of indels longer than one base pair also suggested the presence of a signature of four-base-pair insertions at tetrameric repeats, which was particularly prevalent in mouse and tiger; a pattern of insertions of five or more base pairs at repeats in colobus; and a pattern of deletions of five or more base pairs at repeats, which was prominent in rabbit and resembles ID8 (a signature possibly caused by double-strand break repair through non-homologous end joining³⁰) (Extended Data Fig. 7a).

Other mutational processes and selection

The apparent lack of additional mutational signatures is noteworthy. A previous study of 445 colorectal crypts from 42 human donors found that many crypts were affected by a signature that was later attributed to colibactin, a genotoxin produced by *pks*⁺ strains of *Escherichia coli*^{8,34,35}. Analysing the original human data and our non-human data with the same methodology, we found evidence of colibactin mutagenesis in 21% of human crypts, but only uncertain evidence of colibactin in one non-human crypt (0.6%) (Extended Data Fig. 7b, Methods). This revealed a significant depletion of colibactin mutagenesis in the non-human crypts studied (Fisher's exact test, $P = 7 \times 10^{-14}$). The apparent difference in colibactin mutagenesis observed between species, or between the cohorts studied, might result from a different prevalence of *pks*⁺ *E. coli* strains³⁶ or a

different expression of colibactin by *pks⁺* *E. coli* across species³⁷. Finally, we also searched for evidence of APOBEC signatures (SBS2 and SBS13), which have been reported in a small number of human crypts and are believed to be caused by APOBEC DNA-editing cytidine deaminases. We detected APOBEC signatures in 2% ($n = 9$) of human crypts and found only uncertain evidence in one non-human crypt ($P = 0.30$).

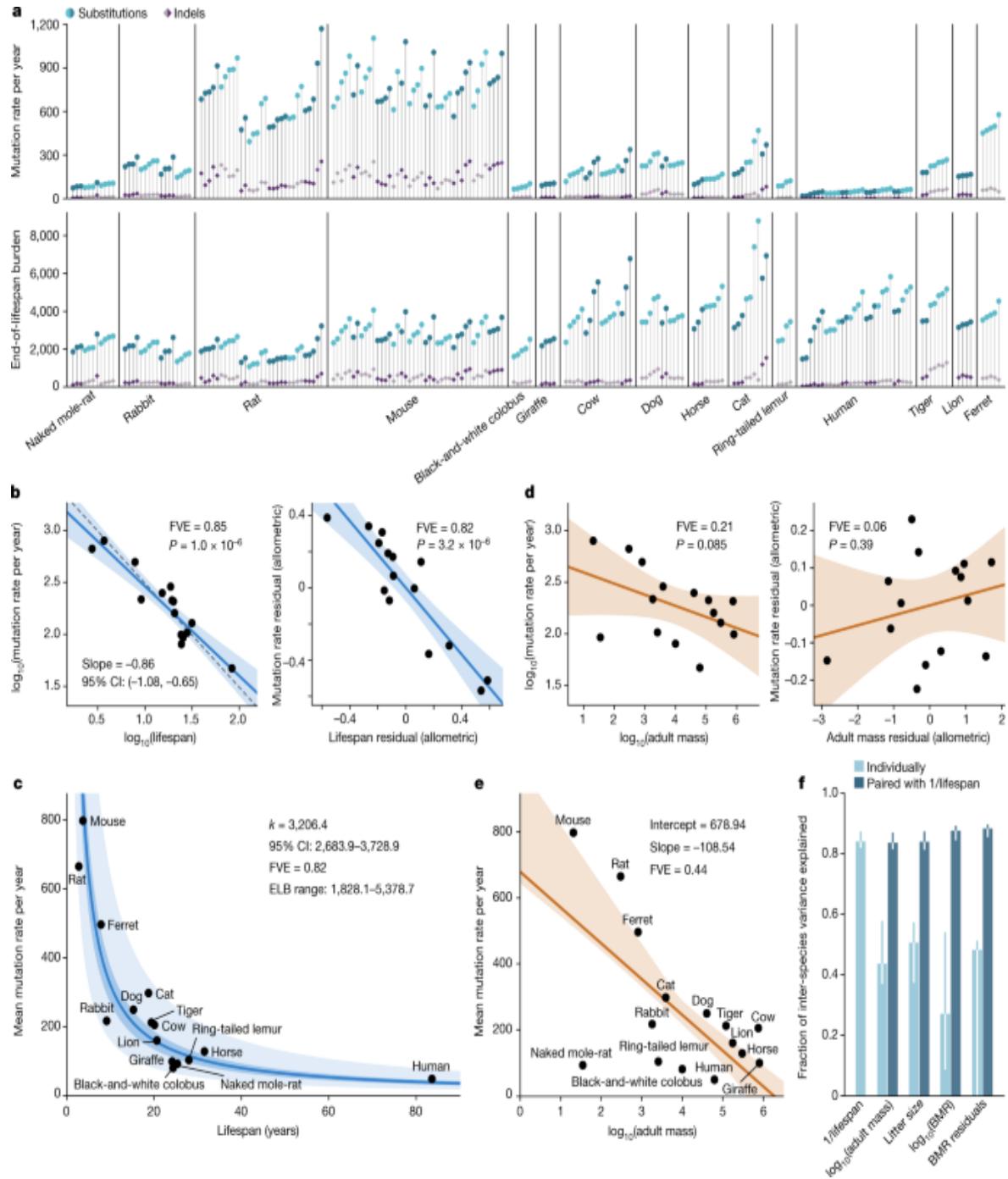
Beyond substitutions and indels, crypts from the eight species with chromosome-level genome assemblies were inspected for large-scale copy number changes (at least 1 Mb) (Methods). Studies in humans have found that large-scale copy number changes are relatively rare in normal tissues, including colorectal epithelium⁸. Consistent with these results, we only identified 4 large copy number changes across the 162 crypts included in this analysis: 2 megabase-scale deletions in 2 crypts from the same cow; the loss of an X chromosome in a female mouse crypt; and a 52-Mb segment with copy-neutral loss of heterozygosity in a human crypt (Extended Data Fig. 8, Methods). These results suggest that large-scale somatic copy number changes in normal tissues are also rare in other mammalian species.

Previous analyses in humans have shown that most somatic mutations in colorectal crypts accumulate neutrally, without clear evidence of negative selection against non-synonymous mutations and with a low frequency of positively selected cancer-driver mutations⁸. To study somatic selection in our data, we calculated the exome-wide ratio of non-synonymous to synonymous substitution rates (dN/dS) in each of the 12 species with available genome annotation. To do so and to detect genes under positive selection, while accounting for the effects of trinucleotide sequence context and mutation rate variation across genes, we used the dNdScv model³⁸ (Methods). Although the limited number of coding somatic mutations observed in most species precluded an in-depth analysis of selection, exome-wide dN/dS ratios for somatic substitutions were not significantly different from unity in any species, in line with previous findings in humans⁸ (Extended Data Fig. 9). Gene-level analysis did not find genes under significant positive selection in any species, although larger studies are likely to identify rare cancer-driver mutations⁸.

Correlation with life-history traits

Whereas similar mutational processes operate across the species surveyed, the mutation rate per genome per year varied widely. Across the 15 species with age information, we found that substitution rates per genome ranged from 47 substitutions per year in humans to 796 substitutions per year in mice, and indel rates from 2.5 to 158 indels per year, respectively (Fig. [3a](#), Supplementary Table [4](#), Methods).

Fig. 3: Associations between somatic mutation rates and life-history traits.



a, Somatic mutation rate per year and expected end-of-lifespan mutation burden (ELB) per crypt. Samples are arranged horizontally as in Fig. 1b; harbour porpoise samples were excluded owing to the age of the sampled individual being unknown. **b**, Left, allometric regression of somatic mutation rate on lifespan. Right, regression of body-mass-adjusted residuals for somatic mutation rate and lifespan (partial correlation; Methods).

Regression was performed using mean mutation rates per species. Shaded areas represent 95% confidence intervals (CI) of regression lines. FVE and P values (by F -test) are indicated (note that, for simple linear regression, FVE = R^2). The dashed line denotes a reference slope of -1 . **c**, Zero-intercept LME regression of somatic mutation rate on inverse lifespan ($1/\text{lifespan}$), presented on the scale of untransformed lifespan (x axis). For simplicity, the y axis shows mean mutation rates per species, although rates per crypt were used in the regression. The darker shaded area indicates 95% CI of the regression line, and the lighter shaded area marks a twofold deviation from the line. Point estimate and 95% CI of the regression slope (k), FVE and range of end-of-lifespan burden are indicated. **d**, Allometric regression and linear regression of lifespan-adjusted residuals, for somatic mutation rate and body mass (elements as described in **b**). **e**, Free-intercept LME regression of somatic mutation rate on log-transformed body mass. The y axis shows mean mutation rates per species, although rates per crypt were used in the regression. Shaded area indicates 95% bootstrap interval of the regression line ($n = 10,000$ replicates). Point estimates of the regression intercept and slope, and FVE, are indicated. **f**, FVE values for free-intercept LME models using $1/\text{lifespan}$ or other life-history variables (alone or combined with $1/\text{lifespan}$) as explanatory variables. Error bars indicate 95% bootstrap intervals ($n = 10,000$).

To investigate the relationship between somatic mutation rates, lifespan and other life-history traits, we first estimated the lifespan of each species using survival curves. We used a large collection of mortality data from animals in zoos to minimize the effect of extrinsic mortality (Extended Data Fig. [10](#)). We defined lifespan as the age at which 80% of individuals reaching adulthood have died, to reduce the effects of outliers and variable cohort sizes that affect maximum lifespan estimates^{[39](#)} (Methods). Notably, we found a tight anticorrelation between somatic mutation rates per year and lifespan across species (Fig. [3b](#)). A log-log allometric regression yielded a strong linear anticorrelation between mutation rate per year and lifespan (fraction of inter-species variance explained (FVE) = 0.85 , $P = 1 \times 10^{-6}$), with a slope close to and not significantly different from -1 . This supports a simple model in which somatic mutation rates per year are inversely proportional to the lifespan of a species (rate $\propto 1/\text{lifespan}$), such

that the number of somatic mutations per cell at the end of the lifespan (the end-of-lifespan burden; ELB) is similar in all species.

To further study the relationship between somatic mutation rates and life-history variables, we used linear mixed-effects (LME) regression models. These models account for the hierarchical structure of the data (with multiple crypts per individual and multiple individuals per species), as well as the heteroscedasticity of somatic mutation rate estimates across species (Methods). Using these models, we estimated that the inverse of lifespan explained 82% of the inter-species variance in somatic substitution rates (rate = $k/\text{lifespan}$) ($P = 2.9 \times 10^{-9}$; Fig. 3c), with the slope of this regression (k) representing the mean estimated ELB across species (3,206.4 substitutions per genome per crypt, 95% confidence interval 2,683.9–3,728.9). Of note, despite uncertainty in the estimates of both somatic mutation rates and lifespans, and despite the diverse life histories of the species surveyed—including around 30-fold variation in lifespan and around 40,000-fold variation in body mass—the estimated mutation load per cell at the end of lifespan varied by only around threefold across species (Table 1). Analogous results were obtained when repeating the analysis with estimates of the protein-coding mutation rate, which may be a better proxy for the functional effect of somatic mutations (85% of variance explained; ELB: 31 coding substitutions per crypt) (Extended Data Fig. 11, Methods).

Table 1 Variation in adult body mass, lifespan, somatic mutation rate and end-of-lifespan mutation burden across the 16 mammalian species surveyed

We next examined the association between somatic mutation rates and adult body mass, which is known to be a common confounder in correlations that involve lifespan^{40,41}. An anticorrelation between somatic mutation rates and body mass may be expected if the modulation of cancer risk across species of vastly different sizes has been a major factor in the evolution of somatic mutation rates. We observed that log-transformed adult body mass was less strongly associated with somatic substitution rates than the inverse of lifespan (allometric regression FVE = 0.21, Fig. 3d; LME regression FVE = 0.44, Fig. 3e). Given that lifespan is correlated with body mass, we performed two tests to assess whether body mass explained any variation in somatic mutation rates that was not explained by lifespan. First, including

both the inverse of lifespan and log-transformed adult body mass in the regression model suggested that body mass does not explain a significant amount of variance in somatic mutation rates across species after accounting for the effect of lifespan (likelihood ratio tests: $P = 0.16$ for body mass on a model with lifespan; $P < 10^{-4}$ for lifespan on a model with body mass; Fig. 3f, Methods). Second, partial correlation analyses using allometric regressions further confirmed that the association between somatic mutation rates and lifespan is unlikely to be mediated by the effect of body mass on both variables (lifespan residuals: $P = 3.2 \times 10^{-6}$, FVE = 0.82, Fig. 3b; body mass residuals: $P = 0.39$, FVE = 0.06, Fig. 3d; Methods).

The fact that the variation in somatic mutation rates across species appears to be dominated by lifespan rather than body size is also apparent when looking at particularly informative species. Giraffe and naked mole-rat, for instance, have similar somatic mutation rates (99 and 93 substitutions per year, respectively), in line with their similar lifespans (80th percentiles: 24 and 25 years, respectively), despite a difference of around 23,000-fold in adult body mass (Fig. 3c,e). Similarly, cows, giraffes and horses weigh much more than an average human, and yet have somatic mutation rates that are several fold higher, in line with expectation from their lifespan but not their body mass. Altogether, the weak correlation between body mass and somatic mutation rates after correction for lifespan suggests that the evolution of larger body sizes may have relied on alternative or additional strategies to limit cancer risk, as has been speculated^{24,42} (Supplementary Note 2). Of note, the low somatic mutation rate of naked mole-rats, which is unusual for their body mass but in line with their long lifespan (Fig. 3c,e), might contribute to the exceptionally low incidence rates of cancer in this species⁴³.

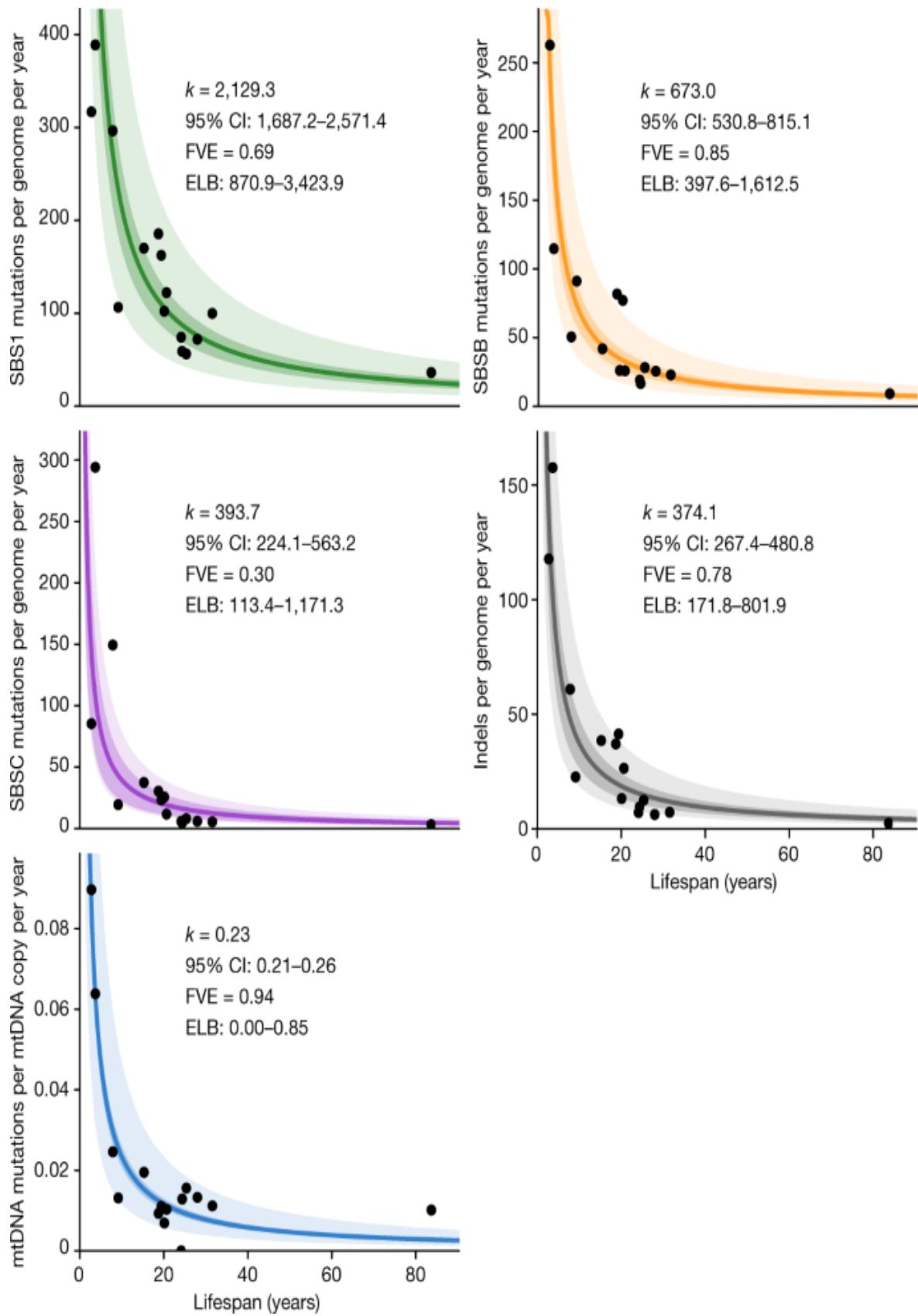
We found similar results for other life-history variables that have been proposed to correlate with lifespan, namely basal metabolic rate (BMR) and litter size⁴⁴ (Fig. 3f). With the caveat that estimates for these variables vary in quality, they showed weaker correlations with the somatic mutation rate as single predictors, and small non-significant increases in explanatory power when considered together with lifespan (likelihood ratio tests: $P = 0.92$ for litter size; $P = 0.083$ for log-BMR; $P = 0.79$ for allometric BMR residuals; Fig. 3f, Methods). We note that the results above are robust to the

use of alternative measures of the somatic mutation rate, including the rate per exome or mutations per Mb (Extended Data Fig. 11, Methods); alternative estimates of lifespan, including maximum lifespan (Extended Data Fig. 12, Methods); alternative regression models, including a Bayesian hierarchical model and a phylogenetic generalised least-squares regression, which accounts for the effect of phylogenetic relationships (Extended Data Fig. 13a, b, Methods); and bootstrapping analyses at the level of individuals or species (Extended Data Fig. 13c, Methods).

Mutational processes and lifespan

To investigate whether a single biological process could drive the association between somatic mutation rates and lifespan, we analysed each mutational signature separately. SBS1, SBSB/5 and SBSC/18 are believed to result from different forms of DNA damage and are expected to be subject to different DNA repair pathways^{18,33}. They also appear to differ in their association with the rate of cell division in humans, with SBS1 being more common in fast-proliferating tissues, such as colon and embryonic or foetal tissues, and SBS5 dominating in post-mitotic cells in the absence of cell division^{14,18,20}. Overall, we found clear anticorrelations between mutation rates per year and lifespan for the three substitution signatures and for indels, suggesting that a single biological process or DNA repair pathway is unlikely to be responsible for this association (Fig. 4). The total mutation burden also appears to show a closer fit with lifespan than individual mutational processes, as measured by the range of end-of-lifespan burden for each process across species (Fig. 4). This might be expected if the observed anticorrelation were the result of evolutionary pressure on somatic mutation rates.

Fig. 4: Association between mutation rate subtypes and species lifespan.



Zero-intercept LME regression of somatic rates of signature-specific substitutions, indels and mtDNA mutations on inverse lifespan (1/lifespan), presented on the scale of untransformed lifespan (x axis). For simplicity, y axes present mean mutation rates per species, although mutation rates per crypt were used in the regressions. The darker shaded areas indicate 95% confidence intervals (CI) of the regression lines, and the lighter shaded areas mark a twofold deviation from the regression lines. Point estimates and 95% CI of the regression slope (k), fraction of inter-species variance explained by for each model (FVE) and ranges of end-of-lifespan burden (ELB) are indicated.

DNA damage and somatic mutations in the mitochondrial genome have also attracted considerable interest in the ageing field⁴⁵. Our whole-genome sequencing of individual crypts provided high coverage of the mitochondrial genome, ranging from 2,188- to 29,691-fold. Normalized against the nuclear coverage, these data suggest that colorectal crypts contain on the order of around 100–2,000 mitochondrial genomes per cell (Extended Data Fig. 14a). Using a mutation-calling algorithm that is sensitive to low-frequency variants, we found a total of 261 mitochondrial mutations across 199 crypts (Extended Data Fig. 14a, Methods). The mutational spectra across species appeared broadly consistent with that observed in humans, with a dominance of C>T and A>G substitutions that are believed to result from mitochondrial DNA (mtDNA) replication errors rather than DNA damage⁴⁶ (Extended Data Fig. 14b). Although the low number of mitochondrial mutations detected per species precludes a detailed analysis, the estimated number of somatic mutations per copy of mtDNA also appears to show an anticorrelation with lifespan. Across species, we obtained an average of 0.23 detectable mutations per copy of the mitochondrial genome by the end of lifespan (Fig. 4, Methods)—a considerable burden given the coding-sequence density and the functional relevance of the mitochondrial genome.

Discussion

Using whole-genome sequencing of 208 colorectal crypts from 56 individuals, we provide insights into the somatic mutational landscape of 16 mammalian species. Despite their different diets and life histories, we found considerable similarities in their mutational spectra. Three main mutational

signatures explain the spectra across species, albeit with varying contributions and subtle variations in the profile of signature SBSB. These results suggest that, at least in the colorectal epithelium, a conserved set of mutational processes dominate somatic mutagenesis across mammals.

The most notable finding of this study is the inverse scaling of somatic mutation rates with lifespan—a long-standing prediction of the somatic mutation theory of ageing^{11,27}. Considering evolutionary and mechanistic models of ageing together provides a framework for discussing the possible implications of these results for ageing (see Supplementary Note 1). Jointly, these models predict ageing to be multifactorial, with multiple forms of molecular and cellular damage contributing to organismal ageing owing to evolutionary limits to selection acting on the rates of these processes. The inverse scaling of somatic mutation rates and lifespan is consistent with somatic mutations contributing to ageing and with somatic mutation rates being evolutionarily constrained, although we discuss alternative explanations below. This interpretation is also supported by studies reporting more efficient DNA repair in longer-lived species^{47,48}. Somatic mutations could contribute to ageing in different ways. Traditionally, they have been proposed to contribute to ageing through deleterious effects on cellular fitness^{11,49}, but recent findings question this assumption (Supplementary Note 1). Instead, the discovery of widespread clonal expansions in ageing human tissues^{19,50,51,52} raises the possibility that some somatic mutations contribute to ageing by driving clonal expansions of functionally altered cells at a cost to the organism^{49,53,54}. Examples include the possible links between clonal haematopoiesis and cardiovascular disease⁵⁴; between mutations in liver disease and insulin resistance⁵⁵; and between driver mutations in cavernomas and brain haemorrhages^{49,53,56}. Detailed studies on the extent and effect of somatic mutations and clonal expansions on age-related diseases and ageing phenotypes may help to clarify the precise role—if any—of somatic mutations in ageing. Even if clear causal links between somatic mutations and ageing are established, ageing is likely to be multifactorial. Other forms of molecular damage involved in ageing could be expected to show similar anticorrelations with lifespan and, indeed, such anticorrelations have been reported for telomere shortening and protein turnover^{57,58}.

Alternative non-causal explanations for the observed anticorrelation between somatic mutation rates and lifespan need to be considered. One alternative explanation is that cell division rates could scale with lifespan and explain the observed somatic mutation rates. Available estimates of cell division rates, although imperfect and limited to a few species, do not readily support this argument (Methods). More importantly, studies in humans have shown that cell division rates are not a major determinant of somatic mutation rates across human tissues^{14,18}. Another alternative explanation for the observed anticorrelation might be that selection acts to reduce germline mutation rates in species with longer reproductive spans, which in turn causes an anticorrelation of somatic mutation rates and lifespan. Although selective pressure on germline mutation rates could influence somatic mutation rates, it is unlikely that germline mutation rates tightly determine somatic mutation rates: somatic mutation rates in humans are 10–20 times higher than germline mutation rates, show variability across cell types and are influenced by additional mutational processes^{18,20}. Overall, the strong scaling of somatic mutation rates with lifespan across mammals, despite the different rates between germline and soma and the variable contributions of different mutational processes across species, suggests that somatic mutation rates themselves have been evolutionarily constrained, possibly through selection on multiple DNA repair pathways. Alternative explanations need to be able to explain the strength of the scaling despite these differences.

Altogether, this study provides a detailed description of somatic mutation across mammals, identifying common and variable features and shedding light on long-standing hypotheses. Scaled across the tree of life and across tissues, in species with markedly different physiologies, life histories, genome compositions and mutagenic exposures, similar studies promise to transform our understanding of somatic mutation and its effects on evolution, ageing and disease.

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to

allocation during experiments and outcome assessment.

Ethics statement

All animal samples were obtained with the approval of the local ethical review committee (AWERB) at the Wellcome Sanger Institute and those at the holding institutions.

Sample collection

We obtained colorectal epithelium and skin samples from a range of sources (Supplementary Table 1). For comparability across species an approximately 1-cm biopsy of the colorectal epithelium was taken from the terminal colon during necropsy. All necropsies occurred as soon as possible post-mortem to minimize tissue and DNA degradation. Tissue samples taken later than 24 h post-mortem typically showed extensive degradation of the colorectal epithelium, making the identification of colorectal crypts challenging. These samples were also associated with poor DNA yields and so were not included in the study. Sampled tissue was fixed in PAXgene FIX (PreAnalytiX, Switzerland), a commercially available fixative, during the necropsy. After 24 h in the fixative at room temperature, samples were transferred into the PAXgene STABILIZER and stored at –20 °C until further processing.

Sample processing

Samples were processed using a workflow designed for detection of somatic mutations in solid tissues by laser-capture microdissection (LCM) using low-input DNA sequencing. For a more detailed description see the paraffin workflow described in another study²⁹. In brief, PAXgene-fixed tissue samples of the colorectal epithelium were paraffin-embedded using a Sakura Tissue-Tek VIP tissue processor. Sections of 16 µm were cut using a microtome, mounted on PEN-membrane slides and stained with Gill's haematoxylin and eosin by sequential immersion in the following: xylene (two minutes, twice), ethanol (100%, 1 min, twice), deionized water (1 min, once), Gill's haematoxylin (10 s, once), tap water (20 s, twice), eosin (5 s,

once), tap water (20 s, once), ethanol (70%, 20 s, twice) and xylene or Neo-Clear, a xylene substitute (20 s, twice).

High-resolution scans were obtained from representative sections of each species. Example images are shown in Fig. 1a, Extended Data Fig. 2. Individual colorectal crypts were isolated from sections on polyethylene naphthalate (PEN) membrane slides by LCM with a Leica LMD7 microscope. Haematoxylin and eosin histology images were reviewed by a veterinary pathologist. For some samples we also cut a section of muscle tissue from below the colorectal epithelium of the section to use as a germline control for variant calling (Supplementary Table 2). Pre- and post-microdissection images of the tissue were recorded for each crypt and muscle sample taken. Each microdissection was collected in a separate well of a 96-well plate.

Crypts were lysed using the Arcturus PicoPure Kit (Applied Biosystems) as previously described^{8,29}. Each crypt then underwent DNA library preparation, without a quantification step to avoid loss of DNA, following a protocol described previously²⁹. For some animals, a PAXgene fixed bulk skin biopsy was used as the germline control. For these skin samples, DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen).

Library preparation and sequencing

Libraries from microdissected samples were prepared using enzymatic fragmentation, adapter ligation and whole-genome sequencing following a method described previously²⁹. Libraries from skin samples were prepared using standard Illumina whole-genome library preparation. Samples were multiplexed and sequenced using Illumina X TEN and Novaseq 6000 machines to generate 150 base pair (bp) paired-end reads. Samples were sequenced to around 30× depth (Supplementary Table 2).

Sequence read alignment

For each species, sequences were aligned to a reference assembly (Supplementary Table 2) using the BWA-MEM algorithm⁵⁹ as implemented in BWA v.0.7.17-r1188, with options ‘-T 30 -Y -p -t 8’. The aligned reads

were sorted using the bamsort tool from the biobambam2 package (v.2.0.86; gitlab.com/german.tischler/biobambam2), with options ‘fixmates=1 level=1 calmdnm=1 calmdnmrecompindetonly=1 calmdnmreference=<reference_fasta> outputthreads=7 sortthreads=7’. Duplicate reads were marked using the bammarkduplicates2 tool from biobambam2, with option ‘level = 0’.

Variant calling

Identification of somatic substitutions and short indels was divided into two steps: variant calling, and variant filtering to remove spurious calls (see ‘Variant filtering’ below). For human colorectal crypts, we obtained previously sequenced and mapped reads from a study in which colorectal crypts were isolated by LCM⁸, and processed them using the sample variant calling and filtering process that was applied to the non-human samples.

Substitutions were identified using the cancer variants through expectation maximization (CaVEMan) algorithm⁶⁰ (v.1.13.15). CaVEMan uses a naive Bayesian classifier to perform a comparative analysis of the sequence data from a target and control sample from the same individual to derive a probabilistic estimate for putative somatic substitutions at each site. The copy number options were set to ‘major copy number = 5’ and ‘minor copy number = 2’, as in our experience this maximizes the sensitivity to detect substitutions in normal tissues. CaVEMan identifies and excludes germline variants shared in the target (colorectal crypt) and matched normal (skin or muscle tissue) samples, and produces a list of putative somatic mutations that are present only in the target sample. CaVEMan was run separately for each colorectal crypt, using either bulk skin or muscle microdissected from the sample colorectal biopsy as the matched normal control (Supplementary Table 2). For two human donors for whom an alternative tissue was not available, a colonic crypt not included as a target sample was used as the matched normal control.

Indels were identified using the Pindel algorithm⁶¹ (v.3.3.0), using a second sample from the same individual as a matched control. The indel calls produced by Pindel were subsequently re-genotyped using the vafCorrect tool (<https://github.com/cancerit/vafCorrect>), which performs a local

sequence assembly to address alignment errors for indels located at the end of sequence reads, and produces corrected counts of sequence reads supporting the indel and corrected estimates of variant allele fraction (VAF; the fraction of reads supporting the alternate allele at the variant site).

Variant filtering

A number of post-processing filters were applied to the variant calls to remove false positives (Supplementary Fig. [1a,b](#)).

Quality flag filter

CaVEMan and Pindel annotate variant calls using a series of quality flags, with the ‘PASS’ flag denoting no quality issues affecting the call^{[60,61](#)}. Variant calls presenting any flag other than ‘PASS’ were discarded.

Alignment quality filter

Variants were excluded if more than half of the supporting reads were clipped. The library preparation methods create short insert size libraries that can result in reads overlapping. To avoid the risk of double counting mutant reads we used fragment-based statistics. Variants without at least four high-quality fragments (alignment score ≥ 40 and base Phred quality score ≥ 30) were excluded. Variants were excluded if reads supporting the variant had a secondary alignment score that was greater than the primary alignment score. This filter was not applied to indel calls.

Hairpin filter

To remove variants introduced by erroneous processing of cruciform DNA during the enzymatic digestion, we applied a custom filter to remove variants in inverted repeats^{[29](#)}. This filter was not applied to indel calls.

Chromosome and contig filter

For species with chromosome-level assemblies, we discarded variants located in non-chromosomal contigs, including the mitochondrial genome (calling of mitochondrial variants is described in the section ‘Mitochondrial variant calling and filtering’). For males, variants on the Y chromosome were excluded for species in which the Y chromosome was annotated in the assembly.

N-tract and contig-end filter

To reduce artefactual calls due to read misalignment, we discarded variants located within 1 kb of a tract of 50 or more consecutive N bases in the reference assembly, as well as variants within 1 kb of the start or end of a contig (this implies discarding all variants in contigs shorter than 2 kb).

Sequencing coverage filter

A sample-specific read depth filter was designed to exclude sites with coverage above the 99th coverage percentile in the sample or its matched normal control, as well as sites with coverage of less than 10 \times in the sample or its matched normal control.

Allelic strand bias filter

We discarded variants without any supporting reads on either the forward or the reverse strand.

Indel proximity filter

We discarded variants for which the total number of reads supporting the presence of an indel within 10 bp of the variant was more than three times larger than the number of reads supporting the presence of the variant. This filter was not applied to indel calls.

Spatial clustering filter

Visual assessment of variant calls and mutational spectra showed spatially clustered variants to be highly enriched for artefacts. Therefore, we discarded groups of two or more variants located within 1 kb of each other.

Beta-binomial filter

For each crypt, an artefact filter based on the beta-binomial distribution was applied, which exploits read count information in other crypts from the same individual. More specifically, for each sample, we fitted a beta-binomial distribution to the variant allele counts and sequencing depths of somatic variants across samples from the same individual. The beta-binomial distribution was used to determine whether read support for a mutation varies across samples from an individual, as expected for genuine somatic mutations but not for artefacts. Artefacts tend to be randomly distributed across samples and can be modelled as drawn from a binomial or a lowly overdispersed beta-binomial distribution. True somatic variants will be present at a high VAF in some samples, but absent in others, and are hence best captured by a highly overdispersed beta-binomial. For each variant site, the maximum likelihood estimate of the overdispersion factor (ρ) was calculated using a grid-based method, with values ranging between 10^{-6} and $10^{-0.05}$. Variants with $\rho > 0.3$ were considered to be artefactual and discarded. The code for this filter is based on the Shearwater variant caller⁶². We found this to be one of the most effective filters against spurious calls (Supplementary Fig. 1b).

Minimum VAF filter

For each sample, we discarded variants for which the VAF was less than half the median VAF of variants passing the beta-binomial filter (see above) in that sample.

Maximum indel VAF filter

For each sample, we discarded indels that presented a VAF of greater than 0.9, as such indels were found to be highly enriched in spurious calls in some species. This filter was not applied to substitution calls.

To validate our variant calling strategy, we used LCM to microdissect two sections from the same mouse colorectal crypt. We expected to detect a high fraction of shared somatic variants in these two sections, as their cells should be derived from the same ancestral epithelial stem cell. Both sections were submitted for independent library preparation, genome sequencing, variant calling and filtering using our pipeline. The majority of substitution variant calls (2,742 of 2,933, 93.5%) were shared between both sections (Supplementary Fig. [1c](#)). By contrast, when comparing five separate crypts from a mouse, a maximum of two variants were shared between two crypts, and no variants were shared by three or more crypts (Supplementary Fig. [1d](#)).

Sample filtering

Our method for estimation of mutation rates assumes monoclonality of colorectal crypt samples. This assumption can be violated owing to several causes, including contamination from other colorectal crypts during microdissection or library preparation, contamination with non-epithelial cells located in or near the crypt, insufficient time for a stem cell to drift to clonality within the crypt, or the possibility that in some species, unlike in humans⁸, polyclonal crypts are the norm. Therefore, a truncated binomial mixture model was applied so as to remove crypts that showed evidence of polyclonality, or for which the possibility of polyclonality could not be excluded. An expectation–maximization (EM) algorithm was used to determine the optimal number of VAF clusters within each crypt sample, as well as each cluster’s location and relative contribution to the overall VAF distribution. The algorithm considered a range of numbers of clusters (1–5), with the optimal number being that which minimized the Bayesian information criterion (BIC). As the minimum number of supporting reads to call a variant was four, the binomial probability distribution was truncated to incorporate this minimum requirement for the number of successes, and subsequently re-normalized. The EM algorithm returned the inferred optimal number of clusters, the mean VAF (location) and mixing proportion (contribution) of each clone, and an assignment of each input variant to the most likely cluster. After applying this model to the somatic substitutions identified in each sample, sample filtering was performed on the basis of the following three criteria.

Low mutation burden

We discarded samples that presented fewer than 50 somatic variants, which was indicative of low DNA quality or sequencing issues.

High mutation burden

We discarded samples with a number of somatic variants greater than 3 times the median burden of samples from the same individual (excluding samples with fewer than 50 variants). This served to exclude a small minority of samples that presented evident sequencing quality problems (such as low sequencing coverage), but which did not fulfil the low-VAF criterion for exclusion (see below).

Low VAF

We discarded samples in which less than 70% of the somatic variants were assigned to clusters with $\text{VAF} \geq 0.3$. However, this rule was not applied to those cases in which all the samples from the same individual had primary clusters with mean $\text{VAF} < 0.3$; this was done to prevent the removal of samples from individuals presenting high fractions of non-epithelial cells, but whose crypts were nonetheless dominated by a single clone.

These criteria led to the exclusion of 41 out of 249 samples. On the basis of visual assessment of sequencing coverage and VAF distributions, we decided to preserve three samples (ND0003c_lo0004, ND0003c_lo0011, TIGRD0001b_lo0010) that we considered to be clonal, but which would have been discarded on the basis of the criteria above.

Mitochondrial variant calling and filtering

For six species whose reference genome assemblies did not include the mitochondrial sequence, mitochondrial reference sequences were obtained from the GenBank database (Supplementary Table 5). For each species, alignment to the reference genome was performed using BWA (v.0.7.17-r1188), as described above (see ‘Sequence read alignment’). Pileup files were generated for mtDNA genomes using the ‘bam2R’ function in the

deepSNV (v.1.32.0) R package^{[62,63](#)}. The mapping quality cut-off was set to 0, taking advantage of the fact that the mitochondrial genome coverage for most samples was more than 100-fold higher than the nuclear genome coverage, and hence most reads with poor mapping scores should be of mitochondrial origin. Mitochondrial variants were called using the Shearwater algorithm^{[62](#)} (deepSNV package v.1.32.0). Multiple rounds of filtering were applied to identify and remove false positives. The first set of filters removed germline polymorphisms, applied a maximum false discovery rate (FDR) threshold of $q > 0.01$, required that mismatches should be supported by at least one read on both the forward and reverse strands, and merged consecutive indel calls. Further filtering steps were as follows.

Minimum VAF filter

Only variants with $\text{VAF} > 0.01$ were considered for analysis, based on the quality of the mutational spectra.

Sequencing coverage filter

Owing to species-specific mtDNA regions of poor mappability, we discarded sites with a read coverage of less than $500\times$.

D-loop filter

Analysis of the distribution of mutations along the mitochondrial genome revealed clusters of mutations within the hypervariable region of mtDNA known as the D-loop. To obtain estimates of the mutation burden in mtDNA unaffected by hypermutation of the D-loop, mutations in the D-loop region (coordinates MT:1–576 and MT:16,024–16,569 in human) were excluded from this analysis.

High mutation burden

We discarded samples that had a number of somatic mtDNA variants greater than four times the mean mtDNA burden across all samples. This served to exclude a small minority of samples that were suspected of enrichment in

false positive calls. Visual inspection of these samples in a genome browser confirmed the presence of high numbers of variants found on sequence reads with identical start positions and/or multiple base mismatches, suggestive of library preparation or sequencing artefacts.

We examined the mutational spectra of somatic mtDNA substitutions on a trinucleotide sequence context (Extended Data Fig. 14b). The specificity of the filtered variant calls was supported by the observation that the mutational spectra across species were broadly consistent with those previously observed in studies of human tissues⁴⁶, with a dominance of C>T and T>C transversions and a strong replication strand bias.

Mitochondrial copy number analysis

Sequence reads from each sample were separately mapped to the species-specific mtDNA reference sequence to estimate average mtDNA sequencing coverage. Excluding nuclear reference sequences from the alignment enabled even coverage to be obtained across the mitochondrial genome by preventing the mismatching of sequence reads to inherited nuclear insertions of mitochondrial DNA (known as NuMTs). Next, coverage information from individual mtDNA and whole-genome alignment (BAM) files was obtained using the genomecov tool in the bedtools suite (v.2.17.0)⁶⁴. Mitochondrial copy number was calculated according to the formula

$$\text{Copy Number} = \frac{\text{depth}_{\text{mtDNA}}}{\text{depth}_{\text{gDNA}}} \times 2$$

where $\text{depth}_{\text{mtDNA}}$ and $\text{depth}_{\text{gDNA}}$ are the mean coverage values for mtDNA and the nuclear genome, respectively, and ploidy = 2 (assuming normal somatic cells to be diploid). For simplicity, the sex chromosomes were excluded from the calculation of the mean nuclear genome coverage.

Calculation of analysable genome size

To estimate the somatic mutation rate, it was first necessary to establish the size of the analysable nuclear genome (that is, the portion of the genome in which variant calling could be performed reliably) for each sample

(Supplementary Table 4). For both substitutions and indels, the analysable genome of a sample was defined as the complement of the union of the following genomic regions: regions reported as ‘not analysed’ by the CaVEMan variant caller; regions failing the ‘chromosome and contig’ filter; regions failing the ‘N-tract and contig-end’ filter; and regions failing the ‘sequencing coverage’ filter (see ‘Variant filtering’). For the analysis of mitochondrial variants, the analysable genome of a sample was defined as the portion of mtDNA that satisfied the ‘sequencing coverage’ filter (see ‘Mitochondrial variant calling and filtering’), after subtracting the hypervariable region (D-loop).

Life-history data

Obtaining accurate lifespan estimates is challenging; although point estimates of maximum lifespan are available for many species, their veracity is often difficult to assess and estimates can vary widely for the same species (Supplementary Table 6). There can be many causes for this variation, including errors in recording and real variation in longevity between populations (that is, captive versus wild). As we were interested in whether the somatic mutation burden has an association with lifespan in the absence of extrinsic mortality, we sought to obtain estimates of longevity from individuals under human care, to minimize the effect of external factors such as predation or infection.

Mortality records for 14 species were obtained from the Species360 database, authorized by Species360 research data use agreement no. 60633 (Species360 Zoological Information Management System (ZIMS) (2020), <https://zims.species360.org>). This database contains lifespan data of zoo animals from international zoo records. Using records from 1980 to the present, we excluded animals for which the date of birth or death was unknown or uncertain. To avoid infant mortality influencing the longevity estimates for each species, we removed animals that died before the age of female sexual maturity, as defined by the AnAge database⁶⁵. This resulted in a mean of 2,681 animal lifespan records per species for the species in the study (minimum 309, maximum 8,403; Supplementary Table 6). For the domestic dog, we combined records for domestic dogs (*Canis lupus familiaris*) and wolves (*Canis lupus*), because of the paucity of records for

domestic dogs in Species360. Although the data are curated, they are still vulnerable to the presence of inaccurate records, which can bias the lifespan estimates. To reduce the effect of these outliers, for each species lifespan was estimated as the age at which 80% of the adults from that species had died⁶⁶ (Supplementary Table 6).

Human longevity estimates were obtained using census birth and death record data from Denmark, (1900–2020), Finland (1900–2019) and France (1900–2018), retrieved from the Human Mortality Database (University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany); <https://www.mortality.org>, <https://www.humanmortality.de>). We selected these countries because they had census records going back at least 100 years. To remove the effect of infant mortality, we excluded individuals who died before the age of 13. For each country, we selected the cohort born in 1900 and calculated the age at which 80% of the individuals had died (Denmark, 87 years; Finland, 83 years; France, 81 years). We then used the mean of the three countries as our estimate of the human 80% lifespan (83.7 years) (Supplementary Table 6).

To test the effects of different estimates of lifespan on our results, we also obtained maximum longevity estimates for each species from a range of databases⁶⁷ and a survey of the literature (Supplementary Table 6). Other life-history metrics were obtained from the AnAge database⁶⁵ (Supplementary Table 6).

Mutational signature analysis

Mutational signatures of substitutions on a trinucleotide sequence context were inferred from sets of somatic mutation counts using the sigfit (v.2.1.0) R package³¹. Initially, signature extraction was performed de novo for a range of numbers of signatures ($N=2,\dots,10$), using counts of mutations grouped per sample, per individual and per species. To account for differences in sequence composition across samples, and especially across species, mutational opportunities per sample, per individual and per species were calculated from the reference trinucleotide frequencies across the analysable genome of each sample (see ‘Calculation of analysable genome size’), and supplied to the ‘extract_signatures’ function in sigfit. The

‘convert_signatures’ function in sigfit was subsequently used to transform the extracted signatures to a human-relative representation (Fig. 2b), by scaling the mutation probability values using the corresponding human genome trinucleotide frequencies. The best-supported number of signatures, on the basis of overall goodness-of-fit³¹ and consistency with known COSMIC signatures (<https://cancer.sanger.ac.uk/signatures/>), was found to be $N=3$. The cleanest deconvolution of the three signatures was achieved when using the mutation counts grouped by species, rather than by sample or individual. The three extracted signatures (labelled SBSA, SBSB and SBSC) were found to be highly similar to COSMIC signatures SBS1 (cosine similarity 0.96), SBS5 (0.89) and SBS18 (0.91), respectively. These signatures were independently validated using the MutationalPatterns (v.1.12.0) R package⁶⁸, which produced comparable signatures (respective cosine similarities 0.999, 0.98 and 0.89).

This de novo signature extraction approach, however, failed to deconvolute signatures SBSA and SBSB entirely from each other, resulting in a general overestimation of the exposure to SBSA (Extended Data Fig. 15). To obtain more accurate estimates of signature exposure, the deconvolution was repeated using an alternative approach that combines signature fitting and extraction in a single inference process³¹. More specifically, the ‘fit_extract_signatures’ function in sigfit was used to fit COSMIC signature SBS1 (retrieved from the COSMIC v.3.0 signature catalogue; <https://cancer.sanger.ac.uk/signatures/>) to the mutation counts grouped by species (with species-specific mutational opportunities), while simultaneously extracting two additional signatures de novo (SBSB and SBSC). Before this operation, COSMIC SBS1 was transformed from its human-relative representation to a genome-independent representation using the ‘convert_signatures’ function in sigfit. By completely deconvoluting SBS1 and SBSB, this approach yielded a version of SBSB that was more similar to COSMIC SBS5 (cosine similarity 0.93); the similarity of SBSC to COSMIC SBS18 was the same under both approaches (0.91).

Finally, the inferred signatures were re-fitted to the mutational spectra of mutations in each sample (using the ‘fit_signatures’ function in sigfit with sample-specific mutational opportunities) to estimate the exposure of each sample to each signature. The fitting of the three signatures yielded spectrum

reconstruction similarity values (measured as the cosine similarity between the observed mutational spectrum and a spectrum reconstructed from the inferred signatures and exposures) with median 0.98 and interquartile range 0.96–0.99. Although the purely de novo extraction approach and the ‘fitting and extraction’ approach yielded comparable versions of signatures SBSB and SBSC, the fixing of COSMIC SBS1 in the latter approach resulted in lower SBS1 exposures and higher SBSB exposures in most samples, owing to the cleaner deconvolution of these two signatures (Fig. 2, Extended Data Fig. 15).

To examine potential variation in the spectrum of signature SBS5 across species, the following procedure was conducted for each species: individual-specific mutation counts and mutational opportunities were calculated for each individual in the species, and the ‘fit_extract_signatures’ function was used to fit COSMIC signatures SBS1, SBS18 and SBS34 (transformed to a genome-independent representation using the ‘convert_signatures’ function) to the mutational spectra of each individual, while simultaneously inferring one additional signature (corresponding to signature SBS5 as manifested in that species; Extended Data Fig. 6).

To assess the presence in non-human colorectal crypts of mutational signatures caused by APOBEC or colibactin, which have been previously observed in human crypts⁸, we used an expectation–maximization algorithm for signature fitting, in combination with likelihood ratio tests (LRTs). More specifically, for each non-human sample, we tested for exposure to colibactin (signature SBS88, COSMIC v.3.2) by comparing the log-likelihoods of (i) a model fitting COSMIC signatures SBS1, SBS5, SBS18, SBS34 and SBS88, and (ii) a reduced model fitting only the first four signatures. Benjamini–Hochberg multiple-testing correction was applied to the *P* values that resulted from the LRTs, and colibactin exposure was considered significant in a sample if the corresponding corrected *q*-value was less than 0.05. We followed the same approach to assess exposure to APOBEC (SBS2 and SBS13), using two separate sets of LRTs for models including either SBS2 or SBS13, in addition to SBS1, SBS5, SBS18 and SBS34. APOBEC exposure was considered significant in a sample if its *q*-values for the models including SBS2 and SBS13 were both less than 0.05. This analysis identified 1/180 crypts with significant exposure to each of colibactin and APOBEC (although the evidence for the presence of the

relevant signatures in these two crypts was not conclusive). To test for depletion of colibactin or APOBEC exposure in non-human crypts relative to human crypts, we first applied the LRT-based method described above to a published set of 445 human colorectal crypts⁸, identifying 92 colibactin-positive and 9 APOBEC-positive crypts. We then compared the numbers of colibactin- and APOBEC-positive crypts in the human and non-human sets using two separate Fisher's exact tests ('fisher.test' function in R). This revealed the difference in colibactin exposure to be highly significant ($P = 7 \times 10^{-14}$), unlike the difference in APOBEC exposure ($P = 0.30$).

Mutational spectra of somatic indels identified in each species were generated using the 'indel.spectrum' function in the Indelwald tool for R (24/09/2021 version; <https://github.com/MaximilianStammnitz/Indelwald>).

Selection analysis

Evidence of selection was assessed using the ratio of nonsynonymous to synonymous substitution rates (dN/dS) in the somatic mutations called in each species. The dNdScv (v.0.0.1.0) R package³⁸ was used to estimate dN/dS ratios for missense and truncating substitutions in each species separately. Reference CDS databases for the dNdScv package were built for those species with available genome annotation in Ensembl (<https://www.ensembl.org>; Supplementary Table 2), using the 'buildref' function in dNdScv. For each species, the 'dndscv' function was applied to the list of somatic substitutions called in samples of that species, after de-duplicating any substitutions that were shared between samples from the same individual to avoid counting shared somatic mutations multiple times. In addition, the analysis was restricted to genes that were fully contained in the analysable genomes of all samples from the species (a condition satisfied by the vast majority of protein-coding genes). Genome-wide and gene-specific dN/dS ratios were obtained for missense and truncating substitutions in each species; no genes with statistically significant dN/dS $\neq 1$ were observed.

Copy number analysis

For species with chromosome-level assemblies (cat, cow, dog, horse, human, mouse, rabbit and rat), the total and the allele-specific copy number (CN) was assessed in each sample, adapting a likelihood model that was previously applied to the detection of subclonal CN changes in healthy human skin¹⁹. This method exploits two sources of evidence: relative sequencing coverage and B-allele fraction (BAF; the fraction of reads covering a heterozygous single-nucleotide polymorphism (SNP) that support one of the alleles). Human samples PD36813x15 and PD36813x16 were excluded from this analysis owing to the poor quality of their SNP data.

For each sample, sequencing coverage was measured in non-overlapping 100-kb bins along the reference genome of the species, using the coverageBed tool in the bedtools suite (v.2.17.0)⁶⁴. For each bin, the coverage per base pair was calculated by dividing the number of reads mapping to the bin by the bin length, and multiplying the result by the read length (150 bp). A normalized sample–normal coverage ratio was then calculated for each bin by dividing the bin coverage in the sample by the corresponding coverage in its matched normal control (see ‘Sample processing’). Heterozygous SNPs were isolated for each sample by selecting germline SNPs with a BAF between 0.4 and 0.6 in the matched normal sample, and a coverage of at least 15 reads in both the target sample and its matched normal sample. After assigning each SNP to its corresponding 100-kb genome bin, the bins in each sample were divided into two sets: (i) bins with coverage ≥ 10 in both the target sample and its matched normal, and at least one heterozygous SNP; and (ii) bins with coverage ≥ 10 in both the target sample and its matched normal, and no heterozygous SNPs. For the first set, estimates of total and allele-specific CN were inferred by maximizing the joint likelihood of a beta-binomial model for BAF and a negative binomial model for relative coverage, as previously described¹⁹. The most likely combination of allele CN values was obtained for each bin by conducting an exhaustive search of CN values between 0 and 4, and selecting the combination maximizing the joint likelihood (calculated on the basis of expected BAF and relative coverage values). A penalty matrix was used to penalize more complex solutions over simpler ones, as previously described¹⁹. For the second set of bins (bins without SNPs), only estimates of total CN were inferred, by maximizing the likelihood of a negative binomial model for relative coverage. The most substantial differences

between these methods and the one previously published are: (i) SNPs were obtained from the variant calling output, instead of from a public database; (ii) relative coverage was calculated per 100-kb bin, rather than per SNP; (iii) SNPs were not phased within each gene, but within each bin; (iv) no reference bias was assumed (that is, the underlying BAF of heterozygous SNPs was assumed to be 0.5); (v) the minimum sample purity was raised to 0.85; (vi) putative CN changes were not subjected to significance testing, but selected according to their likelihood, and subsequently filtered by means of a segmentation algorithm (see below).

Estimates of total and allele-specific CN per bin were merged into CN segments, which were defined as contiguous segments composed of five or more bins with identical CN states. Segmentation was performed separately for total and allele-specific CN estimates in each sample. After this process, any pair of adjacent segments with the same CN assignment, and separated by a distance shorter than five bins, was merged into a single segment. Finally, within each species, segments presenting CN values other than 2 (or 1/1 for allele-specific CN), and being either shorter than 10 bins (1 Mb), or shared among two or more samples, were discarded, resulting in the removal of nearly all spurious CN changes.

Estimation of mutation rate

For each sample, the somatic mutation density (mutations per bp) was calculated by dividing the somatic mutation burden (total number of mutations called) by the analysable genome size for the sample (see ‘Calculation of analysable genome size’). The adjusted somatic mutation burden (number of mutations per whole genome) was then calculated by multiplying the mutation density by the total genome size of the species (see below). The somatic mutation rate per year (mutations per genome per year) was obtained by dividing this adjusted mutation burden by the age of the individual, expressed in years (Supplementary Table 2). The expected ELB for each sample was calculated by multiplying the somatic mutation rate by the estimated lifespan of the species (see ‘Life-history data’).

The total genome size of a species was estimated as the total size of its reference genome assembly. Across species, the mean genome size was 2.67 Gb, ranging between 2.41 Gb and 3.15 Gb and with a standard deviation of

221 Mb (Supplementary Table 4). This suggests that inter-species variation in genome size should not have a substantial influence on the somatic mutation rate estimates. For an assessment of alternative measures of mutation rate, see ‘Association of mutation rate and end-of-lifespan burden with lifespan’.

Association of mutation rate with life-history traits

The association of the somatic mutation rate with different life-history traits was assessed using LME models. In particular, associations with the following traits were examined: lifespan (in years), adult mass (or adult weight, in grams), BMR (in watts), and litter size (see ‘Life-history data’). Associations for lifespan, adult mass and BMR were assessed using the following transformed variables: $1/\text{lifespan}$, $\log_{10}(\text{adult mass})$ and $\log_{10}(\text{BMR})$. To account for the potentially confounding effect of the correlation between metabolic rate and body mass, the residuals of a fitted allometric regression model of BMR on adult mass (equivalent to a simple linear regression of $\log_{10}(\text{BMR})$ on $\log_{10}(\text{adult mass})$) were used as a mass-adjusted measure of metabolic rate, referred to as ‘BMR residuals’.

For each variable, an LME model was implemented for the regression of somatic mutation rates per sample on the variable of interest, using the ‘lme’ function in the nlme R package (v.3.1-137; <https://cran.r-project.org/web/packages/nlme>). To account for non-independence of the samples, both at the individual level and at the species level, the model included fixed effects (intercept and slope parameters) for the variable of interest, and random effects (slope parameters) at the individual and species levels. In addition, to account for the heteroscedasticity of mutation rate estimates across species, the usual assumption of constant response variance was replaced with explicit species-specific variances, to be estimated within the model.

To determine the fraction of inter-species variance in mutation rate explained by each life-history variable individually, the LME model described above was used to produce predictions of the mean mutation rate per species; only fixed effects were used when obtaining these predictions, random effects being ignored. The variance of these predictions was then

compared to the variance in observed mean mutation rates; the latter were calculated for each species as the mean of the observed mean rates per individual, to avoid individuals with larger numbers of samples exerting a stronger influence on the species mean. The fraction of inter-species variance explained by the model was calculated using the standard formula for the coefficient of determination,

$$\$ \$ \{R\}^2 = \{\text{rm}\{\text{ESS}\}\} / (\{\text{rm}\{\text{ESS}\}\} + \{\text{rm}\{\text{RSS}\}\}), \$ \$$$

where ESS is the explained sum of squares, and RSS is the residual sum of squares:

$$\$ \$ \{\text{rm}\{\text{ESS}\}\} = \{\sum\}_{i=1}^n \{(\{\hat{y}_i\} - \bar{y})\}^2, \\ \{\text{rm}\{\text{RSS}\}\} = \{\sum\}_{i=1}^n \{(\{y_i\} - \{\hat{y}_i\})\}^2. \$ \$$$

In this formulation, $\{y_i\}$ and $\{\hat{y}_i\}$ denote the observed and predicted mutation rates for species i , respectively, and $\{\bar{y}\}$ is the overall mean rate. This definition of R^2 coincides with the fraction of variance explained (FVE), defined as 1 minus the fraction of variance unexplained (FVU):

$$\$ \$ \{\text{rm}\{\text{FVE}\}\} = 1 - \{\text{rm}\{\text{mbox}\{\dots\}\}\} \{\text{rm}\{\text{FVU}\}\} = 1 - \{\text{rm}\{\text{mbox}\{\dots\}\}\} [\{\text{rm}\{\text{RSS}\}\} / (\{\text{rm}\{\text{ESS}\}\} + \{\text{rm}\{\text{RSS}\}\})] = \{\text{rm}\{\text{ESS}\}\} / (\{\text{rm}\{\text{ESS}\}\} + \{\text{rm}\{\text{RSS}\}\}) = \{R\}^2. \$ \$$$

As the predicted and observed values correspond to mean mutation rates per species, rather than mutation rates per sample, FVE provides a measure of the fraction of inter-species variance explained by the fixed effects of the LME model. Among the variables considered, 1/lifespan was found to have the greatest explanatory power (FVE = 0.84, using a free-intercept model).

To compare the explanatory power of variables other than 1/lifespan when considered either individually or in combination with 1/lifespan, the method described above was also applied to two-variable combinations of 1/lifespan and each of the remaining variables, using an LME model with fixed effects for both variables and random effects for 1/lifespan only. The R^2 formula above was used to measure the fraction of inter-species variance explained by each model. In addition, to test whether the inclusion of a second

explanatory variable was justified by the increase in model fit, LRTs between each two-variable LME model and a reduced LME model including only 1/lifespan were performed using the ‘anova’ function in the nlme R package.

To further assess the potential effects of body mass and lifespan on each other’s association with the somatic mutation rate, allometric regression models (equivalent to simple linear models under logarithmic transformation of both variables) were fitted to the mean somatic mutation rate per species, using either adult mass or lifespan as the explanatory variable. In addition, the ‘allometric residuals’ of mutation rate, adult mass and lifespan (that is, the residuals of pairwise allometric regressions among these three variables) were used to examine the associations between somatic mutation rate and either body mass or lifespan, after accounting for the effect of the third variable (partial correlation analysis). For example, to account for the potential influence of body mass on the relationship between somatic mutation rate and lifespan, the residuals of an allometric regression between mutation rate and adult mass, and the residuals of an allometric regression between lifespan and adult mass, were analysed using simple linear regression. This analysis supported a strong association between somatic mutation rate and lifespan (independently of the effect of mass; $FVE = 0.82$, $P = 3.2 \times 10^{-6}$; Fig. 3c), and a non-significant association between somatic mutation rate and body mass (independently of the effect of lifespan). Therefore, the relationship between somatic mutation rate and lifespan does not appear to be mediated by the effect of body mass on both variables. Of note, this result remains after excluding naked mole-rat: after removing this species, partial correlation analysis still reveals a strong association between somatic mutation rate and lifespan ($FVE = 0.77$, $P = 4.1 \times 10^{-5}$), and a non-significant association between somatic mutation rate and body mass ($P = 0.84$). This demonstrates that the observed relationships are not dependent on the presence of naked mole-rat in the study.

To assess the robustness of the LME regression analyses described above, we performed bootstrap analysis on each LME model, at the level of both individuals and species. More specifically, for each level we used each of the LME models to perform regression on 10,000 bootstrap replicates, produced by resampling either species or individuals with replacement. We then assessed the distributions of FVE across bootstrap replicates (Extended Data

Fig. [13c](#)). In addition, we performed a similar bootstrap analysis using a collection of maximum longevity estimates obtained from the literature (see ‘Life-history data’). We applied the zero-intercept LME model described above (for regressing mutation rate on inverse lifespan) on a set of 5,000 bootstrap replicates, each of which used a set of species lifespan estimates randomly sampled from the collection of literature-derived estimates (Extended Data Fig. [12](#)).

The results obtained with the LME models were additionally validated using an independent hierarchical Bayesian model, in which the mean somatic mutation burden of each individual was modelled as following a normal distribution with mean defined as a linear predictor containing a species-specific slope parameter and a multiplicative offset (corresponding to the individual’s age; inclusion of this offset minimizes the heteroscedasticity of the mutation rate across species, which results from dividing mutation burdens by age). Species-specific slope parameters were in turn modelled as normally distributed around a global slope parameter, equivalent to the fixed-effect slope estimated by the LME model. This hierarchical model produced very similar results to those of the LME model for all life-history variables (Extended Data Fig. [13a](#)).

We note that samples CATD0002b_lo0003 and MD6267ab_lo0003 were excluded from all regression analyses, owing to the fact that each shared the most of its somatic variants with another sample from the same individual (indicating, in each case, that both samples were closely related), hence violating the assumption of independence among samples. The inclusion of these two samples, however, had no effect on the outcome of the analyses.

Association of mutation rate and end-of-lifespan burden with lifespan

The relationship between somatic mutation rate and species lifespan was further explored by adapting the LME model described in the previous section to perform constrained (zero-intercept) regression of the adjusted mutation rate per year on the inverse of lifespan, $1/\text{lifespan}$ (see ‘Life-history data’, ‘Estimation of mutation rate’ and ‘Association of mutation rate with life-history traits’). The use of zero-intercept regression was motivated by

the prediction that, if somatic mutation is a determinant of maximum lifespan, then it would be expected for all species to end their lifespans with a similar somatic mutation burden. Indeed, this was confirmed by simple linear regression of the species mean end-of-lifespan mutation burden against species lifespan (slope $P = 0.39$). Thus, if m is the mutation rate per year, and L is the species' lifespan, the expected relationship is of the form.

$$m \approx k/L$$

where k is a constant representing the typical end-of-lifespan mutation burden across species. According to this relationship, the mutation rate per year is linearly related to the inverse of lifespan,

$$m \approx k(1/L)$$

Therefore, the cross-species average end-of-lifespan burden (k), can be estimated as the slope parameter of a zero-intercept linear regression model with the mutation rate per year (m) as the dependent variable, and the inverse of lifespan ($1/L$) as the explanatory variable. To this purpose, the LME model described in the previous section was altered by removing the fixed-effect intercept parameter, thus considering only fixed- and random-effect slope parameters for $1/L$ ifespan.

The zero-intercept LME model estimated a value of $k = 3,210.52$ (95% confidence interval 2,686.89–3,734.15). The fraction of inter-species variance explained by the zero-intercept model (FVE) was 0.82, whereas the LME model described in the previous section (which estimated $k = 2,869.98$, and an intercept of 14.76) achieved FVE = 0.84 (see ‘Association of mutation rate with life-history traits’). To test whether the increase in model fit justifies the inclusion of an intercept, both models were compared using a LRT (as implemented by the ‘anova’ function in the nlme R package (v.3.1-137)). This yielded $P = 0.23$, indicating that the free-intercept model does not achieve a significantly better fit than the zero-intercept model. Similarly, the zero-intercept model yielded lower values for both the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). Notably, equivalent analyses using somatic mutation rates per megabase and per protein-coding exome (instead of per whole genome) yielded comparable results (Extended Data Fig. 11).

To investigate the possibility of phylogenetic relationships between the species sampled confounding the analysis, a phylogenetic generalized linear model was used to regress the mean mutation rate of each species on the inverse of its lifespan ($1/L$), while accounting for the phylogenetic relationships among species. A phylogenetic tree of the 15 species examined was obtained from the TimeTree resource⁶⁹, and the phylogenetic linear model was fitted using the ‘pgls’ function in the caper R package (v.1.0.1; <https://cran.r-project.org/web/packages/caper>). The estimates produced by zero-intercept regression of mean mutation rates per species on $1/\text{lifespan}$ were compared between this phylogenetic generalized linear model and a simple linear model ('lm' function in R). The use of this simple model, as well as the use of mean mutation rates per species (rather than mutation rates per sample), was necessary owing to the impossibility of replicating the heteroscedastic mixed-effects structure of the LME model used for the main association analyses (see ‘Association of mutation rate with life-history traits’) within the phylogenetic linear model. Both the phylogenetic linear model and the simple linear model produced similar estimates (Extended Data Fig. 13b), suggesting that phylogenetic non-independence of the samples does not have a substantial effect on the association analyses.

Cell division analysis

To investigate the extent to which differences in cell division rate could explain differences in mutation rate and burden across species, we obtained estimates of intestinal crypt cell division rates from mouse⁷⁰, rat⁷¹ and human^{72,73} (Supplementary Table 7). Using these cell division rates, our lifespan estimates and the observed substitution rates, we calculated the number of cell divisions at the end of lifespan and the corresponding number of mutations per cell division expected under a simple model assuming that all mutations occur during cell division (Supplementary Table 7).

We investigated whether differences in the number of cell divisions among species could explain the observed differences in mutation burden.

Although colorectal cell division rate estimates are lacking for most species, existing estimates from mouse, rat and human indicate that the total number of stem cell divisions per crypt in a lifetime varies greatly across species—for example, there are around 6- to 31-fold more divisions per intestinal stem cell in a human than in a rat over their respective lifetimes, depending on the estimate of cell division rate used (Supplementary Table 7). Mouse intestinal stem cells are estimated to divide once every 24 h (ref. 70), whereas estimates of the human intestinal stem cell division rate vary from once every 48 h (ref. 72) to once every 264 h (ref. 73). Thus, mouse intestinal stem cells divide 2–11 times faster than human intestinal stem cells. By the end of lifespan, an intestinal stem cell is predicted to have divided around 1,351 times in a mouse, around 486 times in a rat and 2,774–15,257 times in a human (depending on the estimate of cell division rate used). Applying our somatic mutation burden and lifespan data, this implies that the somatic mutation rate per cell division in a mouse is around 1.5- to 8.4-fold higher than in a human. However, the observed fold difference in somatic mutation rate between these two species is 16.9 (Table 1). Therefore, differences in cell division rate appear unable to fully account for the observed differences in mutation rate across species. Nevertheless, we note that accurate cell division rate estimates for basal intestinal stem cells are lacking for most species.

Reporting summary

Further information on research design is available in the [Nature Research Reporting Summary](#) linked to this paper.

Data availability

DNA sequence data have been deposited in the European Genome-Phenome Archive (EGA; <https://ega-archive.org>) under overarching accession [EGAD00001008032](#). Human DNA sequence data from a previous study⁸ are deposited in the EGA (accession [EGAD00001004192](#)). Processed mutation calls and other data used in the analyses have been deposited in Zenodo (<https://doi.org/10.5281/zenodo.5554777>). Raw mortality data used to estimate lifespan (Species360 Data Use Approval Number 60633) cannot be publicly shared, as Species360 is the custodian (not the owner) of their members' data. Raw data are accessible through Research Request applications to Species360. Once Species360 grants access to data, they are intended only for and restricted to use in the project they were approved for and for a single publication. Any email communications should be directed to support@species360.org.

Code availability

The computer code used in the analyses has been deposited in Zenodo (<https://doi.org/10.5281/zenodo.5554801>) and GitHub (<https://github.com/baezortega/CrossSpecies2021>).

References

1. Garcia, A. M. et al. Age- and temperature-dependent somatic mutation accumulation in *Drosophila melanogaster*. *PLoS Genet.* **6**, e1000950 (2010).
2. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).

3. Milholland, B. et al. Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.* **8**, 15183 (2017).
4. Schmid-Siegert, E. et al. Low number of fixed somatic mutations in a long-lived oak tree. *Nat. Plants* **3**, 926–929 (2017).
5. Jager, M. et al. Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. *Genome Res.* **29**, 1067–1077 (2019).
6. Orr, A. J. et al. A phylogenomic approach reveals a low somatic mutation rate in a long-lived plant. *Proc. R. Soc. B* **287**, 20192364 (2020).
7. López, E. H. & Palumbi, S. R. Somatic mutations and genome stability maintenance in clonal coral colonies. *Mol. Biol. Evol.* **37**, 828–838 (2020).
8. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
9. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
10. Szilard, L. On the nature of the aging process. *Proc. Natl Acad. Sci. USA* **45**, 30–45 (1959).
11. Morley, A. A. The somatic mutation theory of ageing. *Mutat. Res.* **338**, 19–23 (1995).
12. Vijg, J. & Dong, X. Pathogenic mechanisms of somatic mutation and genome mosaicism in aging. *Cell* **182**, 12–23 (2020).
13. Welch, J. S. et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
14. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).

15. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
16. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
17. Zhang, L. et al. Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc. Natl Acad. Sci. USA* **116**, 9014–9019 (2019).
18. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
19. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
20. Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature* **597**, 381–386 (2021).
21. Peto, R., Roe, F. J., Lee, P. N., Levy, L. & Clack, J. Cancer and ageing in mice and men. *Br. J. Cancer* **32**, 411–426 (1975).
22. Vincze, O. et al. Cancer risk across mammals. *Nature* **601**, 263–267 (2022).
23. Peto, R. Epidemiology, multistage models, and short-term mutagenicity tests. *Int. J. Epidemiol.* **45**, 621–637 (2016).
24. Tollis, M., Boddy, A. M. & Maley, C. C. Peto’s paradox: how has evolution solved the problem of cancer prevention? *BMC Biol.* **15**, 60 (2017).
25. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
26. Schumacher, B., Pothof, J., Vijg, J. & Hoeijmakers, J. H. J. The central role of DNA damage in the ageing process. *Nature* **592**, 695–703

(2021).

27. Burnet, M. *Intrinsic Mutagenesis: a Genetic Approach to Ageing* (Springer, 1974).
28. Kirkwood, T. B. & Holliday, R. The evolution of ageing and longevity. *Proc. R. Soc. B* **205**, 531–546 (1979).
29. Ellis, P. et al. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**, 841–871 (2021).
30. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
31. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. Preprint at *bioRxiv* <https://doi.org/10.1101/372896> (2020).
32. Lindahl, T. & Nyberg, B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**, 3405–3410 (1974).
33. Zou, X. et al. A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nat. Cancer* **2**, 643–657 (2021).
34. Wilson, M. R. et al. The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**, eaar7785 (2019).
35. Pleguezuelos-Manzano, C. et al. Mutational signature in colorectal cancer caused by genotoxic *pks⁺* *E. coli*. *Nature* **580**, 269–273 (2020).
36. Smati, M. et al. Quantitative analysis of commensal *Escherichia coli* populations reveals host-specific enterotypes at the intra-species level. *MicrobiologyOpen* **4**, 604–615 (2015).

37. Oliero, M. et al. Oligosaccharides increase the genotoxic effect of colibactin produced by *pks⁺* *Escherichia coli* strains. *BMC Cancer* **21**, 172 (2021).
38. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
39. Moorad, J. A., Promislow, D. E. L., Flesness, N. & Miller, R. A. A comparative assessment of univariate longevity measures using zoological animal records. *Aging Cell* **11**, 940–948 (2012).
40. Speakman, J. R. Correlations between physiology and lifespan—two widely ignored problems with comparative studies. *Aging Cell* **4**, 167–175 (2005).
41. de Magalhães, J. P., Costa, J. & Church, G. M. An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. *J. Gerontol. A* **62**, 149–160 (2007).
42. Vazquez, J. M. & Lynch, V. J. Pervasive duplication of tumor suppressors in Afrotherians during the evolution of large bodies and reduced cancer risk. *eLife* **10**, e65041 (2021).
43. Smith, E. S. J., Schuhmacher, L.-N. & Husson, Z. The naked mole-rat as an animal model in biomedical research: current perspectives. *Open Access Anim. Physiol.* **7**, 137–148 (2015).
44. Millar, J. S. & Zammuto, R. M. Life histories of mammals: an analysis of life tables. *Ecology* **64**, 631–635 (1983).
45. Kauppila, T. E. S., Kauppila, J. H. K. & Larsson, N.-G. Mammalian mitochondria and aging: an update. *Cell Metab.* **25**, 57–71 (2017).
46. Ju, Y. S. et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *eLife* **3**, e02935 (2014).

47. Hall, K. Y., Hart, R. W., Benirschke, A. K. & Walford, R. L. Correlation between ultraviolet-induced DNA repair in primate lymphocytes and fibroblasts and species maximum achievable life span. *Mech. Ageing Dev.* **24**, 163–173 (1984).
48. Zhang, L. et al. Maintenance of genome sequence integrity in long- and short-lived rodent species. *Sci. Adv.* **7**, eabj3284 (2021).
49. Smith, J. M. Review lectures on senescence—I. The causes of ageing. *Proc. R. Soc. B* **157**, 115–127 (1962).
50. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
51. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
52. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
53. Ren, A. A. et al. *PIK3CA* and CCM mutations fuel cavernomas through a cancer-like mechanism. *Nature* **594**, 271–276 (2021).
54. Jaiswal, S. & Libby, P. Clonal haematopoiesis: connecting ageing and inflammation in cardiovascular disease. *Nat. Rev. Cardiol.* **17**, 137–144 (2020).
55. Ng, S. W. K. et al. Convergent somatic mutations in metabolism genes in chronic liver disease. *Nature* **598**, 473–478 (2021).
56. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* **366**, eaan4673 (2019).
57. Swovick, K. et al. Interspecies differences in proteome turnover kinetics are correlated with life spans and energetic demands. *Mol. Cell. Proteomics* **20**, 100041 (2021).
58. Whittemore, K., Vera, E., Martínez-Nevado, E., Sanpera, C. & Blasco, M. A. Telomere shortening rate predicts species life span. *Proc. Natl*

Acad. Sci. USA **116**, 15122–15127 (2019).

59. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
60. Jones, D. et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinformatics* **56**, 15.10.1–15.10.18 (2016).
61. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.7.1–15.7.12 (2015).
62. Gerstung, M., Papaemmanuil, E. & Campbell, P. J. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* **30**, 1198–1204 (2014).
63. Gerstung, M. et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3**, 811 (2012).
64. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
65. Tacutu, R. et al. Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res.* **46**, D1083–D1090 (2018).
66. Tidière, M. et al. Comparative analyses of longevity and senescence reveal variable survival benefits of living in zoos across mammals. *Sci. Rep.* **6**, 36361 (2016).
67. Conde, D. A. et al. Data gaps and opportunities for comparative and conservation biology. *Proc. Natl Acad. Sci. USA* **116**, 9658–9664 (2019).
68. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).

69. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
70. Snippert, H. J. et al. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).
71. Rijke, R. P., Plaisier, H. M. & Langendoen, N. J. Epithelial cell kinetics in the descending colon of the rat. *Virchows Arch. B Cell Pathol. Incl. Mol. Pathol.* **30**, 85–94 (1979).
72. Potten, C. S., Kellett, M., Rew, D. A. & Roberts, S. A. Proliferation in human gastrointestinal epithelium using bromodeoxyuridine in vivo: data for different sites, proximity to a tumour, and polyposis coli. *Gut* **33**, 524–529 (1992).
73. Bach, S. P., Renehan, A. G. & Potten, C. S. Stem cells: the intestinal stem cell as a paradigm. *Carcinogenesis* **21**, 469–476 (2000).

Acknowledgements

We thank the staff of the Wellcome Sanger Institute for help in this project, including the Scientific Operations, Informatics and CASM support teams. This research was made possible by the worldwide information network of zoos and aquariums that are members of Species360, and is authorized by Species360 research data use and grant agreement 60633. We thank D. Conde and J. Staerk for help accessing Species360 data; and J. P. de Magalhaes, J. DeGregori, J. Vijg and M. Lynch for their comments on the manuscript. This research was funded by Wellcome (grant number 206194), the Dunhill Medical Trust (RPGF2002\188) and the Deutsche José Carreras Leukämie-Stiftung. I.M. is funded by Cancer Research UK (C57387/A21777) and the Wellcome Trust. P.J.C. is a Wellcome Trust Senior Clinical Fellow.

Author information

Author notes

1. These authors contributed equally: Alex Cagan, Adrian Baez-Ortega

Affiliations

1. Cancer, Ageing and Somatic Mutation (CASM), Wellcome Sanger Institute, Hinxton, UK

Alex Cagan, Adrian Baez-Ortega, Natalia Brzozowska, Federico Abascal, Tim H. H. Coorens, Mathijs A. Sanders, Andrew R. J. Lawson, Luke M. R. Harvey, Shriram Bhosle, David Jones, Raul E. Alcantara, Timothy M. Butler, Yvette Hooks, Kirsty Roberts, Elizabeth Anderson, Sharna Lunn, Peter J. Campbell, Michael R. Stratton & Iñigo Martincorena

2. Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, the Netherlands

Mathijs A. Sanders

3. Wildlife Health Services, Zoological Society of London, London, UK

Edmund Flach, Simon Spiro, Inez Januszczak, Ethan Wrigglesworth, Hannah Jenkins, Tilly Dallas & Nic Masters

4. The Natural History Museum, London, UK

Inez Januszczak

5. Institute of Zoology, Zoological Society of London, London, UK

Matthew W. Perkins & Robert Deaville

6. Division of Experimental Hematology, German Cancer Research Center (DKFZ), Heidelberg, Germany

Megan Druce, Ruzhica Bogeska & Michael D. Milsom

7. Heidelberg Institute for Stem Cell Technology and Experimental Medicine GmbH (HI-STEM), Heidelberg, Germany

Megan Druce, Ruzhica Bogeska & Michael D. Milsom

8. Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK

Björn Neumann

9. Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK

Björn Neumann

10. Department of Veterinary Medicine, University of Cambridge, Cambridge, UK

Frank Gorman, Fernando Constantino-Casas, Laura Peachey, Diana Bochynska & Elizabeth P. Murchison

11. Bristol Veterinary School, Faculty of Health Sciences, University of Bristol, Langford, UK

Laura Peachey

12. Department of Pathology, Faculty of Veterinary Medicine, Universitatea de Stiinte Agricole si Medicina Veterinara, Cluj-Napoca, Romania

Diana Bochynska

13. Department of Pharmacology, University of Cambridge, Cambridge, UK

Ewan St. John Smith

14. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK

Moritz Gerstung

Contributions

A.C., E.P.M., M.R.S. and I.M. conceived the project. I.M., E.P.M. and M.R.S. supervised the project. E.F., S.S., I.J., E.W., N.M., R.D., M.W.P., M.D., R.B., M.D.M., F.G., F.C.-C., L.P., D.B., E.St.J.S., B.N. and E.P.M. performed and facilitated sample collection. A.C. performed the laser capture microdissection. A.C., L.M.R.H., A.R.J.L., Y.H., K.R., E.A. and S.L. processed the samples. A.C., A.B.-O., F.A., N.B., T.H.H.C., M.A.S., D.J., R.E.A. and S.B. processed the data. A.C., A.B.-O. and N.B. led the analysis with help from F.A., T.H.H.C., M.A.S., A.R.J.L., T.M.B., T.D., H.J., E.P.M. and I.M. The manuscript was written by A.C., A.B.-O., N.B. and I.M., with input from all of the authors.

Corresponding authors

Correspondence to [Alex Cagan](#) or [Iñigo Martincorena](#).

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Kamila Naxerova and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. [Peer reviewer reports](#) are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

[Extended Data Fig. 1 Somatic mutational spectra of human colon and small intestine.](#)

Trinucleotide-context mutational spectra of somatic substitutions from human adult stem cells in colon (top) and small intestine, using mutation calls obtained from a previous study¹⁴.

[Extended Data Fig. 2 Histology images of intestinal crypts across species.](#)

Histological images of the colorectal or intestinal (ferret) epithelium for each non-human species. Scale bars are provided at the bottom of each image.

[Extended Data Fig. 3 Somatic VAF distributions per species.](#)

Distributions of variant allele fraction (VAF) for somatic substitutions in each crypt for each species. Each distribution refers to the variants in a single sequenced crypt.

[Extended Data Fig. 4 Somatic mutation accumulation across species.](#)

Each panel presents somatic substitution burdens per genome (corrected for analysable genome size) for a given species. Each dot represents a crypt sample, with samples from the same individual sharing the same colour. For species with two or more individuals, the estimated regression line from a simple linear regression model on individual mean burdens is shown. For species with three or more individuals, the shaded region indicates 95%

confidence intervals of the regression line. Harbour porpoise samples were excluded owing to unknown age of the sampled individual.

Extended Data Fig. 5 Signature-specific mutation accumulation across species.

Each panel presents somatic substitution burdens per genome for mutational signatures SBS1 (green), SBSB (yellow) and SBSC (purple) in a given species. For species with two or more individuals, the estimated regression lines from simple linear regression models on individual mean burdens per signature are shown. For species with three or more individuals, shaded regions indicate 95% confidence intervals of the regression lines. Harbour porpoise samples were excluded owing to unknown age of the sampled individual.

Extended Data Fig. 6 Profiles of signature SBSB as inferred from different species.

Trinucleotide-context mutational spectra of signature SBSB, as inferred independently from somatic mutations in crypts from four representative species (top to bottom): human, naked mole-rat, rat and rabbit (Methods). Signatures are shown in a human-genome-relative representation. Cosine similarities between each signature and the COSMIC human signatures SBS5 and SBS40 are provided.

Extended Data Fig. 7 Somatic indels and colibactin exposure.

a, Mutational spectra of somatic indels in each species. The *x* axis shows 83 types of insertion or deletion, coloured by type and length³⁰. **b**, Colibactin exposure in non-human and human colorectal crypts. Exposures to mutational signatures SBS1, SBS5, SBS18, SBS34 and SBS88, as inferred by expectation–maximization, for 180 non-human crypts in this study (top) and 445 human crypts sequenced in a previous study⁸. Asterisks indicate samples with statistically significant colibactin (SBS88) exposure, based on a LRT (Methods). BW, black-and-white; H, harbour; N, naked; RT, ring-tailed.

Extended Data Fig. 8 Identified copy number changes.

a–d, Somatic copy number changes in cow (**a, b**), mouse (**c**) and human (**d**) colorectal crypts. In each case, chromosomes are presented along the *x* axis, with each point representing a 100-kb genomic bin. The top panel presents the ratio between observed and expected sequencing coverage per bin; the middle panel shows the median BAF of heterozygous germline SNPs per bin; and the bottom panel presents the inferred segments of total copy number (green) and allele-specific copy number (red/blue). Regions of copy number change are highlighted in pink. The sparsity of BAF and allele-specific copy number values in the mouse crypt (**c**) are related to the fact that mouse samples generally had very low numbers of germline SNPs.

Extended Data Fig. 9 Somatic dN/dS.

Estimates of dN/dS for missense and truncating somatic mutations in each of the species with available genome annotation. Dots and error bars represent maximum likelihood estimates and 95% confidence intervals, respectively ($n = 27, 2, 32, 2, 136, 12, 118, 9, 39, 7, 102, 10, 440, 34, 231, 22, 25, 3, 30, 2, 110, 10, 75$ and 6 mutations, left to right). Note the logarithmic scale of the *y* axis.

Extended Data Fig. 10 Kaplan–Meier curves of longevity in captivity.

Kaplan–Meier survival curves for each species, calculated using captive lifespan data from Species360 for non-human species and census record data for humans (Methods). The shaded areas represent 95% confidence intervals of the survival curves. A horizontal grey bar indicates the age at which 80% of individuals had already died (80th percentile), which was adopted as a robust estimate of species lifespan.

Extended Data Fig. 11 Associations between life-history variables and alternative measures of somatic mutation rate.

a, b, Same analyses as Fig. 3c, f, but using somatic mutation rates per megabase (**a**), or per protein-coding exome (**b**), rather than per genome (Methods). Leftmost panels show zero-intercept LME regressions of somatic mutation rates on inverse lifespan (1/lifespan), presented on the scale of untransformed lifespan (x axis). The y axes present mean mutation rates per species, although mutation rates per crypt were used in the regressions. Darker shaded areas indicate 95% confidence intervals (CI) of the regression lines; lighter shaded areas mark a two-fold deviation from the regression lines. Point estimate and 95% CI of the regression slope (k), fraction of inter-species variance explained (FVE), and range of ELB are provided. Rightmost panels show comparisons of FVE values achieved by free-intercept LME models using inverse lifespan and other life-history variables (alone or in combination with inverse lifespan) as explanatory variables. BW, black-and-white; N, naked; RT, ring-tailed.

Extended Data Fig. 12 Bootstrapped regression of somatic mutation rates on published lifespan estimates.

a, Bootstrapped regression of somatic substitution rates on the inverse of lifespan (1/lifespan), using a zero-intercept LME model (Methods). For each of 5,000 bootstrap samples (replicates), lifespan values per species were randomly chosen from a set of published maximum longevity estimates (Supplementary Table 6). The blue line indicates the median regression slope (k) across bootstrap samples, and the shaded area depicts the range of estimates of k across bootstrap samples. Black dots and error bars indicate the mean and range, respectively, of published longevity estimates for each species. The median and range of both k and the fraction of inter-species variance explained (FVE) are provided. **b**, Histogram of FVE values across the 5,000 bootstrap samples.

Extended Data Fig. 13 Comparison of regression models for somatic mutation rates.

a, Zero-intercept regression of somatic substitution rates on inverse lifespan (1/lifespan), using a LME model applied to mutation rates per crypt (left) and a Bayesian hierarchical normal regression model applied to mean

mutation rates per individual (Methods). For simplicity, black dots present mean mutation rates per species. Darker shaded areas indicate 95% confidence/credible intervals (CI) of the regression lines; lighter shaded areas mark a two-fold deviation from the regression lines. Point estimates and 95% CI of the regression slopes (k) and fraction of inter-species variance explained (FVE) are provided. **b**, Comparison of regression lines for the regression of somatic substitution rates on 1/lifespan (left; zero intercept) and log-transformed adult body mass (right; free intercept), using simple linear models (dark and light blue), phylogenetic generalized least-squares models (orange and yellow), Bayesian hierarchical normal models (green) and LME models (red) (Methods). Point estimates of the regression coefficients for each model are provided. **c**, Distributions of regression FVE under individual- and species-level bootstrapping. For the LME models regressing somatic mutation rates on inverse lifespan (zero intercept; left) and log-transformed mass (free intercept), the curves present distributions of FVE from 10,000 bootstrap replicates, obtained through random resampling of either individuals (blue) or species (orange) (Methods). Vertical lines indicate the FVE values obtained using the entire dataset.

Extended Data Fig. 14 mtDNA mutation burden and spectrum.

a, Total somatic mtDNA mutations called (substitutions and indels; top), somatic mutation burden per mtDNA copy (middle), and estimated mtDNA copy number in each crypt sample. Samples are arranged horizontally as in Fig. 1b, with samples from the same individual coloured in the same shade of grey. **b**, Mutational spectra of mtDNA substitutions in each species. The x axis shows 96 mutation types on a trinucleotide context, coloured by base substitution type; the y axis shows mutation counts. Mutations on the upper and lower halves of the spectrum represent substitutions with the pyrimidine base located on the heavy and light strands of mtDNA, respectively.

Extended Data Fig. 15 Mutational signatures and exposures as inferred de novo.

a, Mutational signatures inferred de novo from the species mutational spectra shown in Fig. 2a. Signatures are shown in a human-genome-relative

representation. SBSA is the de novo equivalent of COSMIC signature SBS1 (Fig. 2b). **b**, Exposure of each sample to each of the mutational signatures shown in **a**. Samples are arranged horizontally as in Fig. 1b. **c**, Regression of signature-specific mutation burdens on individual age for human, mouse and naked mole-rat samples. Regression was performed using mean mutation burden per individual. Shaded areas indicate 95% confidence intervals of the regression lines. BW, black-and-white; H, harbour; N, naked; RT, ring-tailed.

Supplementary information

Supplementary Information

This file contains Supplementary Note 1 (Theoretical considerations on ageing), Supplementary Note 2 (Cancer risk modulation in the Armitage–Doll model), Supplementary References, Supplementary Figures 1–2, and legends for Supplementary Tables 1–7.

Reporting Summary

Peer Review File

Supplementary Table 1

Species information. For each of the (sub)species in the study, the provides: common name, scientific name, number of individuals in the study, number of colorectal crypts sequenced, range of individual ages, and source institution.

Supplementary Table 2

Sample information. For each colorectal crypt sample in the study, the table provides: sample ID, individual ID, species name, matched normal sample ID, matched normal sample type, and median sequencing depth.

Supplementary Table 3

Mutation rate and burden regression coefficients per species. For each species in the study (except harbour porpoise), the table provides the mean observed values of the rate of somatic substitutions per genome per year, and point estimates and 95% confidence intervals for simple linear regression of mean substitution burdens per individual on individual ages. The estimated regression slopes correspond to the estimated mutation rate per year for each species. Estimates are provided for constrained-intercept linear models applied to all species, and for free-intercept linear models applied to the eight species with at least three individuals.

Supplementary Table 4

Somatic mutation burdens, rates and signature exposures. For each colorectal crypt sample in the study, the table provides: sample ID, individual ID, species name, individual age, total genome size, coding genome size, analysable genome size, analysable mtDNA size, mutational signature exposures (SBS1, SBSB, SBSC); somatic mutation burdens per genome for substitutions, indels, signature-specific substitutions (SBS1, SBSB, SBSC), and mtDNA mutations; and somatic mutation rates per genome per year for substitutions, indels, signature-specific substitutions (SBS1, SBSB, SBSC), and mtDNA mutations.

Supplementary Table 5

Reference genome information. For each species in the study, the table provides: reference genome version used, reference mtDNA sequence used, Ensembl genome annotation version used (where applicable), reference genome file source, reference mtDNA file source, reference genome file URL, reference mtDNA file URL.

Supplementary Table 6

Life-history data. For each species in the study (except harbour porpoise), the table provides: adult mass (g), BMR (W), litter/clutch size, maximum longevity (years); and maximum likelihood estimate and 95% confidence limits for the estimated 80% lifespan, together with the corresponding sample size. The source of each estimate is given in brackets.

Supplementary Table 7

Number of cell divisions per lifespan. For mouse, rat and human, the table provides: estimated rate of colorectal cell division (hours), estimated lifespan (years), estimated number of cell division at the end of lifespan, estimated number of mutations per cell division (obtained using our mutation rate estimates), and the reference for the cell division rate. Two estimates of cell division rate are included for human.

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Reprints and Permissions

About this article

Cite this article

Cagan, A., Baez-Ortega, A., Brzozowska, N. *et al.* Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517–524 (2022).
<https://doi.org/10.1038/s41586-022-04618-z>

- Received: 17 August 2021

- Accepted: 07 March 2022
- Published: 13 April 2022
- Issue Date: 21 April 2022
- DOI: <https://doi.org/10.1038/s41586-022-04618-z>

Share this article

Anyone you share the following link with will be able to read this content:

[Get shareable link](#)

Sorry, a shareable link is not currently available for this article.

[Copy to clipboard](#)

Provided by the Springer Nature SharedIt content-sharing initiative

Further reading

- [**Mutational clocks tick differently across species**](#)
 - Alexander N. Gorelick
 - Kamila Naxerova

Nature (2022)

[**Mutational clocks tick differently across species**](#)

- Alexander N. Gorelick
- Kamila Naxerova

Nature News & Views 13 Apr 2022

| [Section menu](#) | [Main menu](#) |

[Download PDF](#)

- Article
- Open Access
- [Published: 06 April 2022](#)

Brain charts for the human lifespan

- [R. A. I. Bethlehem](#)^{1,2 na1},
- [J. Seidlitz](#)^{3,4,5 na1},
- [S. R. White](#)^{6,7 na1},
- [J. W. Vogel](#)^{3,8},
- [K. M. Anderson](#)⁹,
- [C. Adamson](#)^{10,11},
- [S. Adler](#)¹²,
- [G. S. Alexopoulos](#)¹³,
- [E. Anagnostou](#)^{14,15},
- [A. Areces-Gonzalez](#)^{16,17},
- [D. E. Astle](#)¹⁸,
- [B. Auyeung](#)^{1,19},
- [M. Ayub](#)^{20,21},
- [J. Bae](#)²²,
- [G. Ball](#)^{10,23},
- [S. Baron-Cohen](#)^{1,24},
- [R. Beare](#)^{10,11},
- [S. A. Bedford](#)¹,
- [V. Benegal](#)²⁵,
- [F. Beyer](#)²⁶,
- [J. Blangero](#)²⁷,
- [M. Blesa Cábez](#)²⁸,
- [J. P. Boardman](#)²⁸,
- [M. Borzage](#)²⁹,
- [J. F. Bosch-Bayard](#)^{30,31},

- N. Bourke^{32,33},
- V. D. Calhoun³⁴,
- M. M. Chakravarty^{31,35},
- C. Chen³⁶,
- C. Chertavian⁵,
- G. Chetelat³⁷,
- Y. S. Chong^{38,39},
- J. H. Cole^{40,41},
- A. Corvin⁴²,
- M. Costantino^{43,44},
- E. Courchesne^{45,46},
- F. Crivello⁴⁷,
- V. L. Cropley⁴⁸,
- J. Crosbie⁴⁹,
- N. Crossley^{50,51,52},
- M. Delarue³⁷,
- R. Delorme^{53,54},
- S. Desrivieres⁵⁵,
- G. A. Devenyi^{56,57},
- M. A. Di Biase^{48,58},
- R. Dolan^{59,60},
- K. A. Donald^{61,62},
- G. Donohoe⁶³,
- K. Dunlop⁶⁴,
- A. D. Edwards^{65,66,67},
- J. T. Elison⁶⁸,
- C. T. Ellis^{9,69},
- J. A. Elman⁷⁰,
- L. Eyler^{71,72},
- D. A. Fair⁶⁸,
- E. Feczko⁶⁸,
- P. C. Fletcher^{73,74},
- P. Fonagy^{75,76},
- C. E. Franz⁷⁰,

- L. Galan-Garcia⁷⁷,
- A. Gholipour⁷⁸,
- J. Giedd^{79,80},
- J. H. Gilmore⁸¹,
- D. C. Glahn^{82,83},
- I. M. Goodyer⁶,
- P. E. Grant⁸⁴,
- N. A. Groenewold^{62,85},
- F. M. Gunning⁸⁶,
- R. E. Gur^{3,5},
- R. C. Gur^{3,5},
- C. F. Hammill^{49,87},
- O. Hansson^{88,89},
- T. Hedden^{90,91},
- A. Heinz⁹²,
- R. N. Henson^{6,18},
- K. Heuer^{93,94},
- J. Hoare⁹⁵,
- B. Holla^{96,97},
- A. J. Holmes⁹⁸,
- R. Holt¹,
- H. Huang^{99,100},
- K. Im⁸²,
- J. Ipser¹⁰¹,
- C. R. Jack Jr¹⁰²,
- A. P. Jackowski^{103,104},
- T. Jia^{105,106,107},
- K. A. Johnson^{83,108,109,110},
- P. B. Jones^{6,74},
- D. T. Jones^{102,111},
- R. S. Kahn¹¹²,
- H. Karlsson^{113,114},
- L. Karlsson^{113,114},
- R. Kawashima¹¹⁵,

- E. A. Kelley¹¹⁶,
- S. Kern^{117,118},
- K. W. Kim^{119,120,121,122},
- M. G. Kitzbichler^{2,6},
- W. S. Kremen⁷⁰,
- F. Lalonde¹²³,
- B. Landeau³⁷,
- S. Lee¹²⁴,
- J. Lerch^{125,126},
- J. D. Lewis¹²⁷,
- J. Li¹²⁸,
- W. Liao¹²⁸,
- C. Liston¹²⁹,
- M. V. Lombardo^{1,130},
- J. Lv^{48,131},
- C. Lynch⁶⁴,
- T. T. Mallard¹³²,
- M. Marcelis^{133,134},
- R. D. Markello¹³⁵,
- S. R. Mathias⁸²,
- B. Mazoyer^{47,136},
- P. McGuire⁵¹,
- M. J. Meaney^{136,137},
- A. Mechelli¹³⁸,
- N. Medic⁶,
- B. Misic¹³⁵,
- S. E. Morgan^{6,139,140},
- D. Mothersill^{141,142,143},
- J. Nigg¹⁴⁴,
- M. Q. W. Ong¹⁴⁵,
- C. Ortinau¹⁴⁶,
- R. Ossenkoppele^{147,148},
- M. Ouyang⁹⁹,
- L. Palaniyappan¹⁴⁹,

- L. Paly³⁷,
- P. M. Pan^{150,151},
- C. Pantelis^{152,153,154},
- M. M. Park¹⁵⁵,
- T. Paus^{156,157},
- Z. Pausova^{49,158},
- D. Paz-Linares^{16,159},
- A. Pichet Binette^{160,161},
- K. Pierce⁴⁵,
- X. Qian¹⁴⁵,
- J. Qiu¹⁶²,
- A. Qiu¹⁶³,
- A. Raznahan¹²³,
- T. Rittman¹⁶⁴,
- A. Rodrigue⁸²,
- C. K. Rollins^{165,166},
- R. Romero-Garcia^{6,167},
- L. Ronan⁶,
- M. D. Rosenberg¹⁶⁸,
- D. H. Rowitch¹⁶⁹,
- G. A. Salum^{170,171},
- T. D. Satterthwaite^{3,8},
- H. L. Schaare^{172,173},
- R. J. Schachar⁴⁹,
- A. P. Schultz^{83,108,174},
- G. Schumann^{175,176},
- M. Schöll^{177,178,179},
- D. Sharp^{32,180},
- R. T. Shinohara^{36,181},
- I. Skoog^{117,118},
- C. D. Smyser¹⁸²,
- R. A. Sperling^{83,108,109},
- D. J. Stein¹⁸³,
- A. Stolicyn¹⁸⁴,

- J. Suckling^{6,74},
- G. Sullivan²⁸,
- Y. Taki¹¹⁵,
- B. Thyreau¹¹⁵,
- R. Toro^{94,185},
- N. Traut^{185,186},
- K. A. Tsvetanov^{164,187},
- N. B. Turk-Browne^{9,188},
- J. J. Tuulari^{113,189,190},
- C. Tzourio¹⁹¹,
- É. Vachon-Presseau¹⁹²,
- M. J. Valdes-Sosa⁷⁷,
- P. A. Valdes-Sosa^{128,193},
- S. L. Valk^{194,195},
- T. van Amelsvoort¹⁹⁶,
- S. N. Vandekar¹⁹⁷,
- L. Vasung¹³⁵,
- L. W. Victoria⁸⁶,
- S. Villeneuve^{160,161,198},
- A. Villringer^{26,199},
- P. E. Vértes^{6,140},
- K. Wagstyl⁶⁰,
- Y. S. Wang^{200,201,202,203},
- S. K. Warfield⁷⁸,
- V. Warrier⁶,
- E. Westman²⁰⁴,
- M. L. Westwater⁶,
- H. C. Whalley¹⁸⁴,
- A. V. Witte^{26,199,205},
- N. Yang^{200,201,202,203},
- B. Yeo^{206,207,208},
- H. Yun²⁰⁹,
- A. Zalesky⁴⁸,
- H. J. Zar^{85,210},

- [A. Zettergren¹¹⁷](#),
- [J. H. Zhou^{145,206,211}](#),
- [H. Ziauddeen^{6,74,212}](#),
- [A. Zugman^{151,213,214}](#),
- [X. N. Zuo^{199,200,201,202,215}](#),
- [3R-BRAIN](#),
- [AIBL](#),
- [Alzheimer's Disease Neuroimaging Initiative](#),
- [Alzheimer's Disease Repository Without Borders Investigators](#),
- [CALM Team](#),
- [Cam-CAN](#),
- [CCNP](#),
- [COBRE](#),
- [cVEDA](#),
- [ENIGMA Developmental Brain Age Working Group](#),
- [Developing Human Connectome Project](#),
- [FinnBrain](#),
- [Harvard Aging Brain Study](#),
- [IMAGEN](#),
- [KNE96](#),
- [The Mayo Clinic Study of Aging](#),
- [NSPN](#),
- [POND](#),
- [The PREVENT-AD Research Group](#),
- [VETSA](#),
- [E. T. Bullmore^{6 na2}](#) &
- [A. F. Alexander-Bloch^{3,4,5 na2}](#)

[**Nature**](#) volume **604**, pages 525–533 (2022)

- 89k Accesses
- 1 Citations
- 1603 Altmetric
- [Metrics details](#)

Subjects

- [Cognitive neuroscience](#)
- [Development of the nervous system](#)
- [Diseases of the nervous system](#)
- [Neural ageing](#)

Abstract

Over the past few decades, neuroimaging has become a ubiquitous tool in basic research and clinical studies of the human brain. However, no reference standards currently exist to quantify individual differences in neuroimaging metrics over time, in contrast to growth charts for anthropometric traits such as height and weight¹. Here we assemble an interactive open resource to benchmark brain morphology derived from any current or future sample of MRI data (<http://www.brainchart.io/>). With the goal of basing these reference charts on the largest and most inclusive dataset available, acknowledging limitations due to known biases of MRI studies relative to the diversity of the global population, we aggregated 123,984 MRI scans, across more than 100 primary studies, from 101,457 human participants between 115 days post-conception to 100 years of age. MRI metrics were quantified by centile scores, relative to non-linear trajectories² of brain structural changes, and rates of change, over the lifespan. Brain charts identified previously unreported neurodevelopmental milestones³, showed high stability of individuals across longitudinal assessments, and demonstrated robustness to technical and methodological differences between primary studies. Centile scores showed increased heritability compared with non-centiled MRI phenotypes, and provided a standardized measure of atypical brain structure that revealed patterns of neuroanatomical variation across neurological and psychiatric disorders. In summary, brain charts are an essential step towards robust quantification of individual variation benchmarked to normative trajectories in multiple, commonly used neuroimaging phenotypes.

Main

The simple framework of growth charts to quantify age-related change was first published in the late eighteenth century¹ and remains a cornerstone of paediatric healthcare—an enduring example of the utility of standardized norms to benchmark individual trajectories of development. However, growth charts are currently available only for a small set of anthropometric variables, such as height, weight and head circumference, and only for the first decade of life. There are no analogous charts available for quantification of age-related changes in the human brain, although it is known to go through a prolonged and complex maturational program from pregnancy to the third decade⁴, followed by progressive senescence from approximately the sixth decade⁵. The lack of tools for standardized assessment of brain development and ageing is particularly relevant to research studies of psychiatric disorders, which are increasingly recognized as a consequence of atypical brain development⁶, and neurodegenerative diseases that cause pathological brain changes in the context of normative senescence⁷. Preterm birth and neurogenetic disorders are also associated with marked abnormalities of brain structure^{8,9} that persist into adult life^{9,10} and are associated with learning disabilities and mental health disorders. Mental illness and dementia collectively represent the single biggest global health burden¹¹, highlighting the urgent need for normative brain charts as an anchor point for standardized quantification of brain structure over the lifespan¹².

Such standards for human brain measurement have not yet materialized from decades of neuroimaging research, probably owing to the challenges of integrating MRI data across multiple, methodologically diverse studies targeting distinct developmental epochs and clinical conditions¹³. For example, the perinatal period is rarely incorporated in analysis of age-related brain changes, despite evidence that early biophysical and molecular processes powerfully influence life-long neurodevelopmental trajectories^{14,15} and vulnerability to psychiatric disorders³. Primary case-control studies are usually focused on a single disorder despite evidence of trans-diagnostically shared risk factors and pathogenic mechanisms, especially in psychiatry^{16,17}. Harmonization of MRI data across primary studies to address these and other deficiencies in the extant literature is challenged by methodological and technical heterogeneity. Compared with relatively simple anthropometric measurements such as height or weight,

brain morphometrics are known to be highly sensitive to variation in scanner platforms and sequences, data quality control, pre-processing and statistical analysis¹⁸, thus severely limiting the generalizability of trajectories estimated from any individual study¹⁹. Collaborative initiatives spurring collection of large-scale datasets^{20,21}, recent advances in neuroimaging data processing^{22,23} and proven statistical frameworks for modelling biological growth curves^{24,25} provide the building blocks for a more comprehensive and generalizable approach to age-normed quantification of MRI phenotypes over the entire lifespan (see Supplementary Information 1 for details and consideration of previous work focused on the related but distinct objective of inferring brain age from MRI data). Here, we demonstrate that these convergent advances now enable the generation of brain charts that (1) robustly define normative processes of sex-stratified, age-related change in multiple MRI-derived phenotypes; (2) identify previously unreported brain growth milestones; (3) increase sensitivity to detect genetic and early life environmental effects on brain structure; and (4) provide standardized effect sizes to quantify neuroanatomical atypicality of brain scans collected across multiple clinical disorders. We do not claim to have yet reached the ultimate goal of quantitatively precise diagnosis of MRI scans from individual patients in clinical practice. However, the present work proves the principle that building normative charts to benchmark individual differences in brain structure is already achievable at global scale and over the entire life-course; and provides a suite of open science resources for the neuroimaging research community to accelerate further progress in the direction of standardized quantitative assessment of MRI data.

Mapping normative brain growth

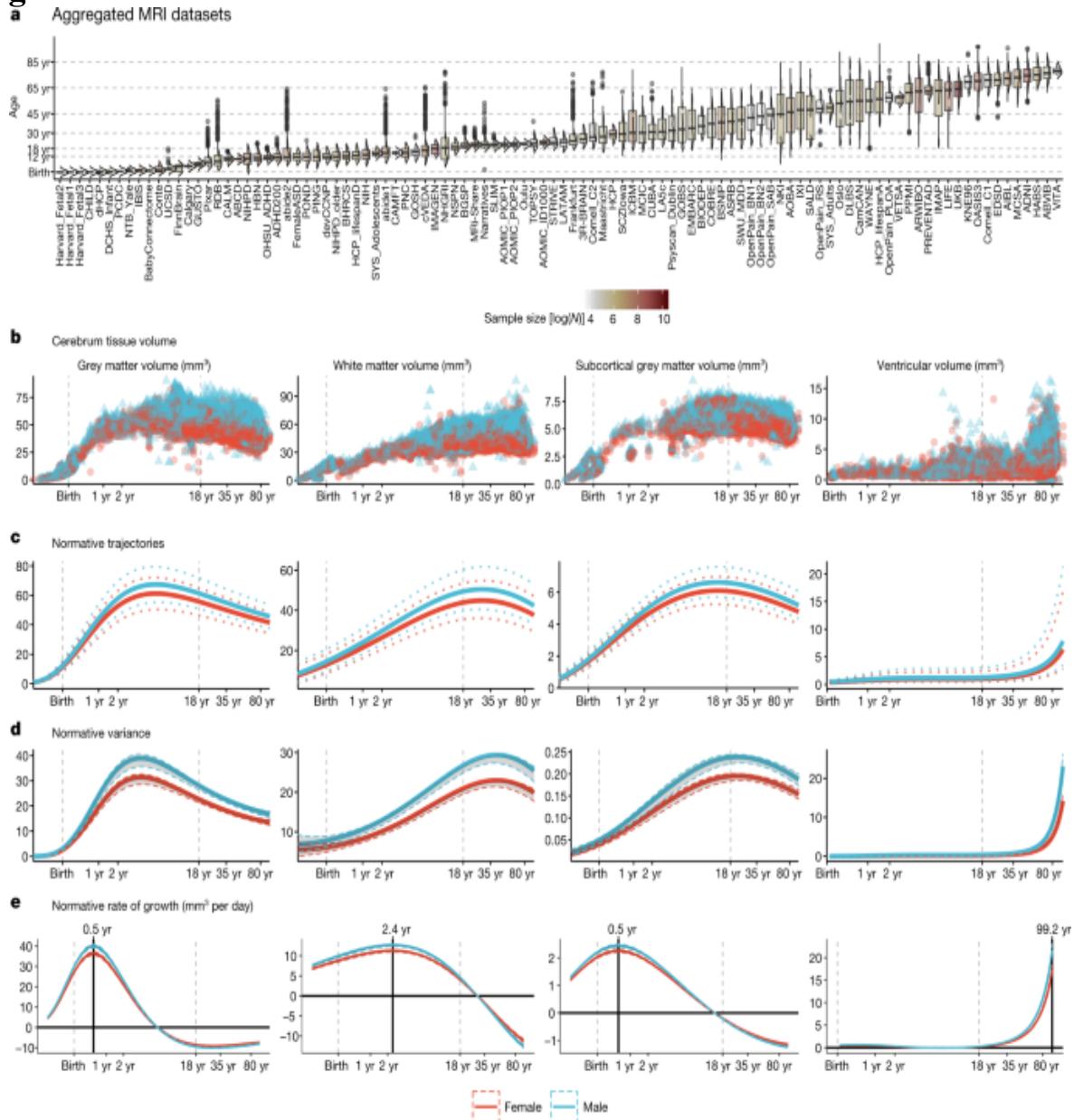
We created brain charts for the human lifespan using generalized additive models for location, scale and shape^{2,24} (GAMLSS), a robust and flexible framework for modelling non-linear growth trajectories recommended by the World Health Organization²⁴. GAMLSS and related statistical frameworks have previously been applied to developmental modelling of brain structural and functional MRI phenotypes in open datasets^{19,26,27,28,29,30,31}. Our approach to GAMLSS modelling leveraged the greater scale of data available to optimize model selection empirically, to

estimate non-linear age-related trends (in median and variance) stratified by sex over the entire lifespan, and to account for site- or study-specific ‘batch effects’ on MRI phenotypes in terms of multiple random effect parameters. Specifically, GAMLSS models were fitted to structural MRI data from control subjects for the four main tissue volumes of the cerebrum (total cortical grey matter volume (GMV), total white matter volume (WMV), total subcortical grey matter volume (sGMV) and total ventricular cerebrospinal fluid volume (ventricles or CSF)). Supplementary Tables 1.1–1.8 present details on acquisition, processing and demographics of the dataset; see [Methods](#), ‘Model generation and specification’ and [Supplementary Information 1](#) for further details regarding GAMLSS model specification and estimation; image quality control, which used a combination of expert visual curation and automated metrics of image quality ([Supplementary Information 2](#)); model stability and robustness ([Supplementary Information 3, 4](#)); phenotypic validation against non-imaging metrics ([Supplementary Information 3 and 5.2](#)); inter-study harmonization ([Supplementary Information 5](#)); and assessment of cohort effects ([Supplementary Information 6](#)). See [Supplementary Information 19](#) for details on all primary studies contributing to the reference dataset, including multiple publicly available open MRI datasets [32,33,34,35,36,37,38,39,40,41,42](#).

Lifespan curves (Fig. 1, [Supplementary Table 2.1](#)) showed an initial strong increase in GMV from mid-gestation onwards, peaking at 5.9 years (95% bootstrap confidence interval (CI) 5.8–6.1), followed by a near-linear decrease. This peak was observed 2 to 3 years later than previous reports relying on smaller, more age-restricted samples [43,44](#). WMV also increased rapidly from mid-gestation to early childhood, peaking at 28.7 years (95% bootstrap CI 28.1–29.2), with subsequent accelerated decline in WMV after 50 years. Subcortical GMV showed an intermediate growth pattern compared with GMV and WMV, peaking in adolescence at 14.4 years (95% bootstrap CI 14.0–14.7). Both the WMV and sGMV peaks are consistent with previous neuroimaging and postmortem reports [45,46](#). By contrast, CSF showed an increase until age 2, followed by a plateau until age 30, and then a slow linear increase that became exponential in the sixth decade of life. Age-related variance (Fig. 1d), explicitly estimated by GAMLSS, formally quantifies developmental changes in between-subject variability. There was

an early developmental increase in GMV variability that peaked at 4 years, whereas subcortical volume variability peaked in late adolescence. WMV variability peaked during the fourth decade of life, and CSF was maximally variable at the end of the human lifespan.

Fig. 1: Human brain charts.



a, MRI data were aggregated from over 100 primary studies comprising 123,984 scans that collectively spanned the age range from mid-gestation to 100 postnatal years. Box–violin plots show the age distribution for each study coloured by its relative sample size (log-scaled using the natural

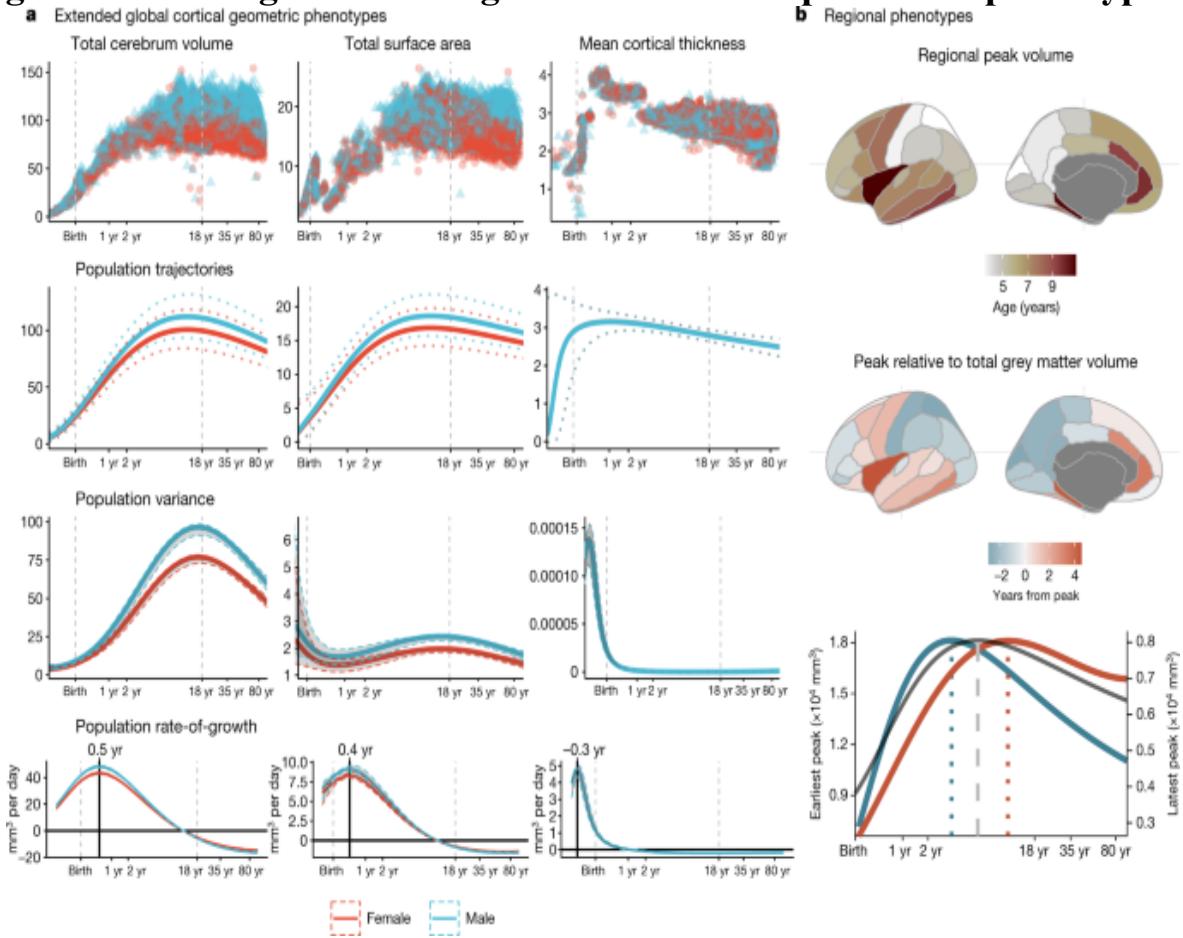
logarithm for visualization purposes). **b**, Non-centiled, ‘raw’ bilateral cerebrum tissue volumes for grey matter, white matter, subcortical grey matter and ventricles are plotted for each cross-sectional control scan as a function of age (log-scaled); points are coloured by sex. **c**, Normative brain-volume trajectories were estimated using GAMLSS, accounting for site- and study-specific batch effects, and stratified by sex (female, red; male, blue). All four cerebrum tissue volumes demonstrated distinct, non-linear trajectories of their medians (with 2.5% and 97.5% centiles denoted as dotted lines) as a function of age over the lifespan. Demographics for each cross-sectional sample of healthy controls included in the reference dataset for normative GAMLSS modelling of each MRI phenotype are detailed in Supplementary Table 1.2–1.8. **d**, Trajectories of median between-subject variability and 95% confidence intervals for four cerebrum tissue volumes were estimated by sex-stratified bootstrapping (see Supplementary Information 3 for details). **e**, Rates of volumetric change across the lifespan for each tissue volume, stratified by sex, were estimated by the first derivatives of the median volumetric trajectories. For solid (parenchymal) tissue volumes, the horizontal line ($y = 0$) indicates when the volume at which each tissue stops growing and starts shrinking and the solid vertical line indicates the age of maximum growth of each tissue. See Supplementary Table 2.1 for all neurodevelopmental milestones and their confidence intervals. Note that y axes in **b–e** are scaled in units of 10,000 mm³ (10 ml).

Extended neuroimaging phenotypes

To extend the scope of brain charts beyond the four cerebrum tissue volumes, we generalized the same GAMLSS modelling approach to estimate normative trajectories for additional MRI phenotypes including other morphometric properties at a global scale (mean cortical thickness and total surface area) and regional volume at each of 34 cortical areas⁴⁷ (Fig. 2, Supplementary Information 7–9, Supplementary Tables 1, 2). We found, as expected, that total surface area closely tracked the development of total cerebrum volume (TCV) across the lifespan (Fig. 2a), with both metrics peaking at approximately 11–12 years of age (surface area peak at 10.97 years (95% bootstrap CI 10.42–11.51); TCV peak at 12.5 years (95% bootstrap CI 12.14–12.89)). By contrast, cortical thickness peaked distinctively early at 1.7 years (95% bootstrap CI 1.3–2.1), which reconciles

previous observations that cortical thickness increases during the perinatal period⁴⁸ and declines during later development⁴⁹ (Supplementary Information 7).

Fig. 2: Extended global and regional cortical morphometric phenotypes.



a, Trajectories for total cerebrum volume (TCV), total surface area and mean cortical thickness. For each global cortical MRI phenotype, the following sex-stratified results are shown as a function of age over the lifespan. From top to bottom: raw, non-centiled data; population trajectories of the median (with 2.5% and 97.5% centiles (dotted lines)); between-subject variance (with 95% confidence intervals); and rate of growth (the first derivatives of the median trajectory and 95% confidence intervals). All trajectories are plotted as a function of log-scaled age (x axis) and y axes are scaled in units of the corresponding MRI metrics ($10,000 \text{ mm}^3$ for TCV, $10,000 \text{ mm}^2$ for surface area and mm for cortical thickness). **b**, Regional variability of cortical volume trajectories for 34 bilateral brain regions, as defined by the

Desikan–Killiany parcellation⁴⁷, averaged across sex (see Supplementary Information [7,8](#) for details). Since models were generated from bilateral averages of each cortical region, the cortical maps are plotted on the left hemisphere purely for visualization purposes. Top, a cortical map of age at peak regional volume (range 2–10 years). Middle, a cortical map of age at peak regional volume relative to age at peak GMV (5.9 years), highlighting regions that peak earlier (blue) or later (red) than GMV. Bottom, illustrative trajectories for the earliest peaking region (superior parietal lobe, blue line) and the latest peaking region (insula, red line), showing the range of regional variability relative to the GMV trajectory (grey line). Regional volume peaks are denoted as dotted vertical lines either side of the global peak, denoted as a dashed vertical line, in the bottom panel. The left *y* axis on the bottom panel refers to the earliest peak (blue line); the right *y* axis refers to the latest peak (red line).

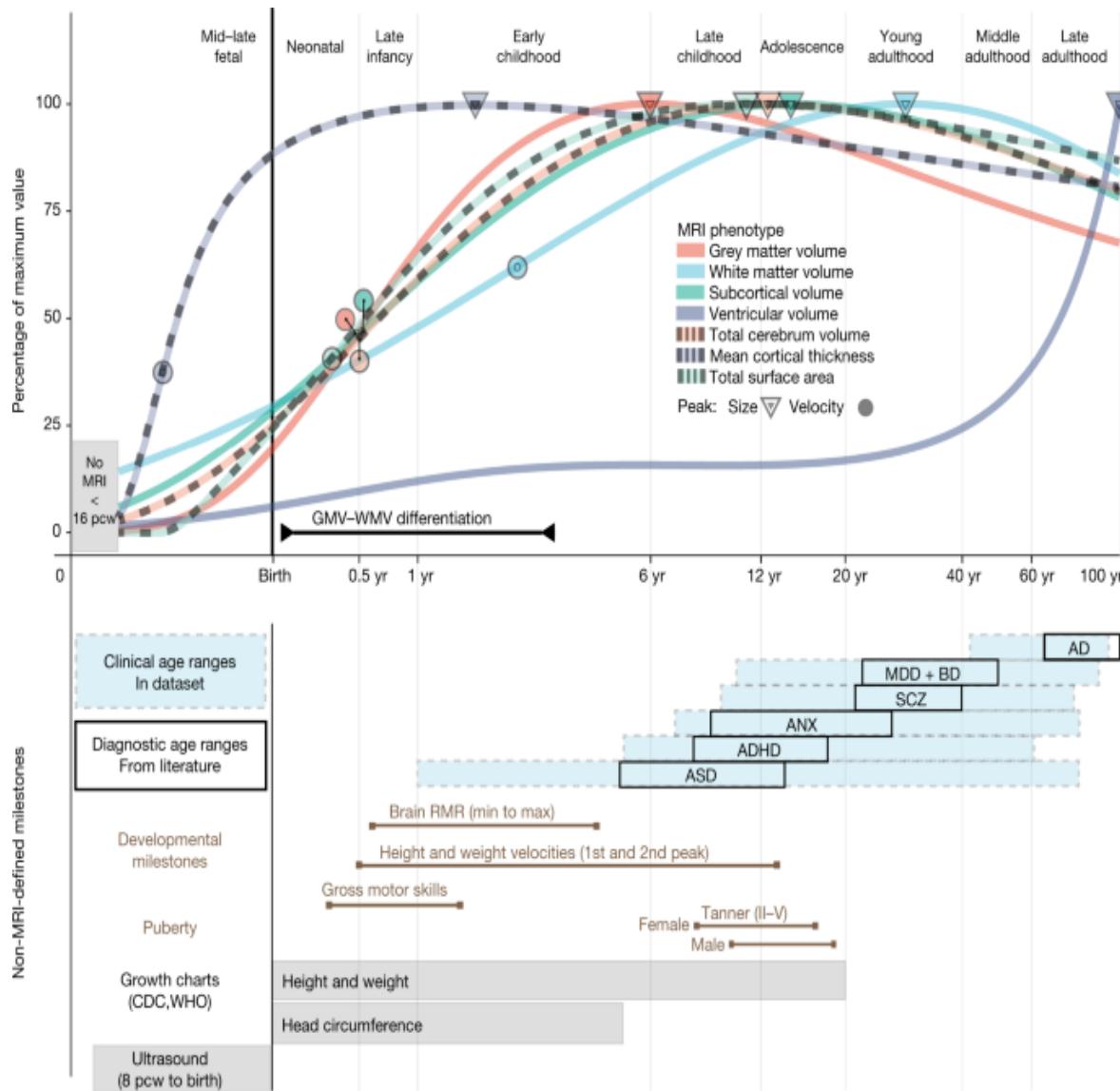
We also found evidence for regional variability in volumetric neurodevelopmental trajectories. Compared with peak GMV at 5.9 years, the age of peak regional grey matter volume varied considerably—from approximately 2 to 10 years—across 34 cortical areas. Primary sensory regions reached peak volume earliest and showed faster post-peak declines, whereas fronto-temporal association cortical areas peaked later and showed slower post-peak declines (Fig. [2b](#), Supplementary Information [8.2](#)). Notably, this spatial pattern recapitulated a gradient from sensory-to-association cortex that has been previously associated with multiple aspects of brain structure and function⁵⁰.

Developmental milestones

Neuroimaging milestones are defined by inflection points of the tissue-specific volumetric trajectories (Fig. [3](#), Methods, ‘Defining developmental milestones’). Among the total tissue volumes, only GMV peaked before the typical age at onset of puberty⁵¹, with sGMV peaking mid-puberty and WMV peaking in young adulthood (Fig. [3](#)). The rate of growth (velocity) peaked in infancy and early childhood for GMV (5.08 months (95% bootstrap CI 4.85–5.22)), sGMV (5.65 months (95% bootstrap CI 5.75–5.83)) and WMV (2.4 years (95% bootstrap CI 2.2–2.6)). TCV velocity peaked between the maximum velocity for GMV and WMV at

approximately 7 months. Two major milestones of TCV and sGMV (peak velocity and size) (Fig. 3) coincided with the early neonatal and adolescent peaks of height and weight velocity^{52,53}. The velocity of mean cortical thickness peaked even earlier, in the prenatal period at -0.38 years (95% bootstrap CI -0.4 to -0.34) (relative to birth), corresponding approximately to mid-gestation. This early peak in cortical thickness velocity has not been reported previously—to our knowledge—in part owing to challenges in acquiring adequate and consistent signal from typical MRI sequences in the perinatal period⁵⁴. Similarly, normative trajectories revealed an early period of GMV:WMV differentiation, beginning in the first month after birth with the switch from WMV to GMV as the proportionally dominant tissue compartment, and ending when the absolute difference of GMV and WMV peaked around 3 years (Supplementary Information 9). This epoch of GMV:WMV differentiation, which may reflect underlying changes in myelination and synaptic proliferation^{4,55,56,57,58}, has not been demarcated in previous studies^{45,59}. It was probably identified in this study owing to the substantial amount of early developmental MRI data available for analysis in the aggregated dataset (in total across all primary studies, $N=2,571$ and $N=1,484$ participants aged less than 2 years were available for analysis of cerebrum tissue volumes and extended global MRI phenotypes, respectively). The period of GMV:WMV differentiation encompasses dynamic changes in brain metabolites⁶⁰ (0–3 months), resting metabolic rate⁶¹ (RMR) (minimum = 7 months, maximum = 4.2 years), the typical period of acquisition of motor capabilities and other early paediatric milestones⁶², and the most rapid change in TCV (Fig. 3).

Fig. 3: Neurodevelopmental milestones.



Top, a graphical summary of the normative trajectories of the median (50th centile) for each global MRI phenotype, and key developmental milestones, as a function of age (log-scaled). Circles depict the peak rate of growth milestones for each phenotype (defined by the maxima of the first derivatives of the median trajectories (Fig. 1e)). Triangles depict the peak volume of each phenotype (defined by the maxima of the median trajectories); the definition of GMV:WMV differentiation is detailed in Supplementary Information 9.1. Bottom, a graphical summary of additional MRI and non-MRI developmental stages and milestones. From top to bottom: blue shaded boxes denote the age range of incidence for each of the major clinical disorders represented in the MRI dataset; black boxes denote the age at which these conditions are generally diagnosed as derived from

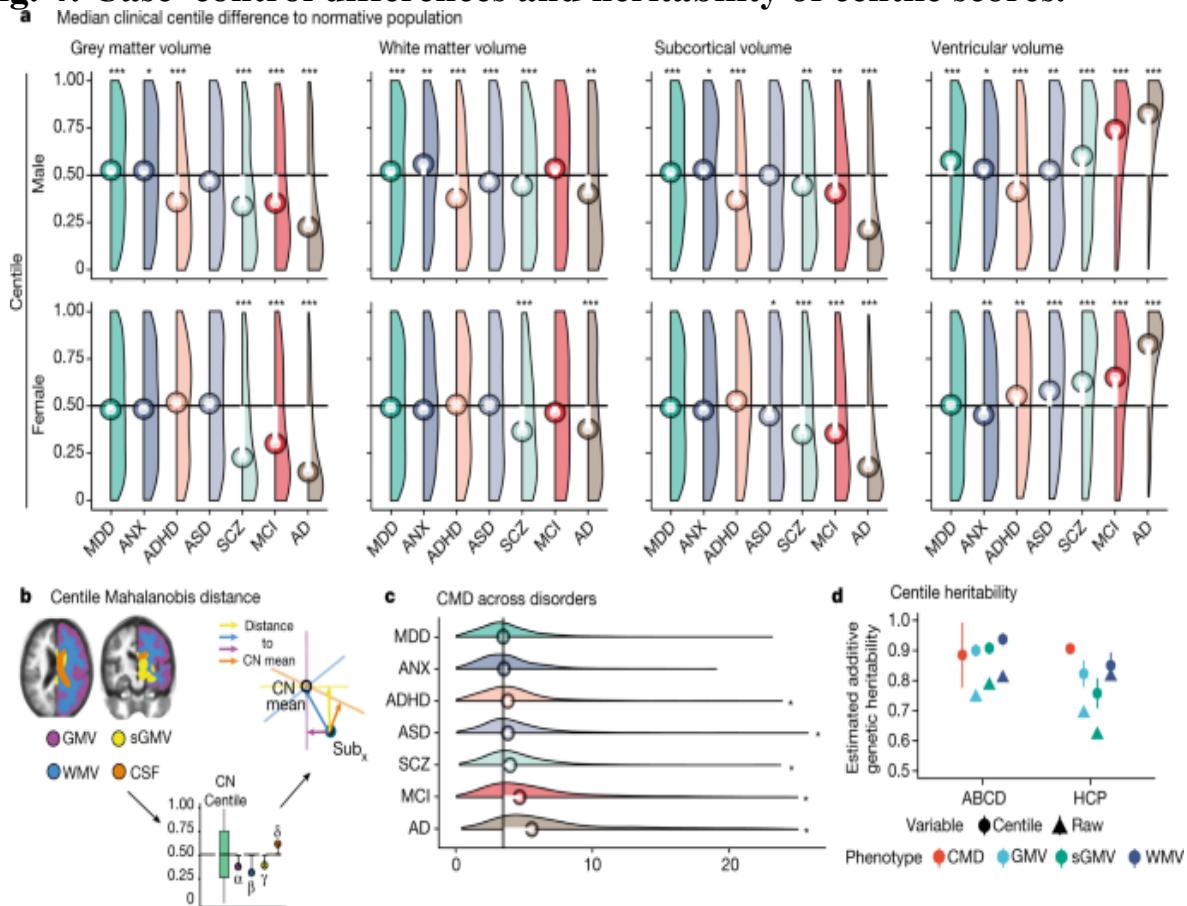
literature⁷³ ([Methods](#)); brown lines represent the normative intervals for developmental milestones derived from non-MRI data, based on previous literature and averaged across males and females ([Methods](#)); grey bars depict age ranges for existing (World Health Organization (WHO) and Centers for Disease Control and Prevention (CDC)) growth charts of anthropometric and ultrasonographic variables²⁴. Across both panels, light grey vertical lines delimit lifespan epochs (labelled above the top panel) previously defined by neurobiological criteria⁶³. Tanner refers to the Tanner scale of physical development. AD, Alzheimer's disease; ADHD, attention deficit hyperactivity disorder; ASD, autism spectrum disorder (including high-risk individuals with confirmed diagnosis at a later age); ANX, anxiety or phobic disorders; BD, bipolar disorder; MDD, major depressive disorder; RMR, resting metabolic rate; SCZ, schizophrenia.

Individualized centile scores

We computed individualized centile scores that benchmarked each individual scan in the context of normative age-related trends (Methods, 'Centile scores and case-control differences' and Supplementary Information [1–6](#) for further details). This approach is conceptually similar to quantile rank mapping, as previously reported^{26,28,29}, where the typicality or atypicality of each phenotype in each scan is quantified by its score on the distribution of phenotypic parameters in the normative or reference sample of scans, with more atypical phenotypes having more extreme centile (or quantile) scores. The clinical diversity of the aggregated dataset enabled us to comprehensively investigate case-control differences in individually specific centile scores across a range of conditions. Relative to the control group (CN), there were highly significant differences in centile scores across large ($N > 500$) groups of cases diagnosed with multiple disorders (Fig. [4a](#), Supplementary Information [10](#)), with effect sizes ranging from medium ($0.2 < \text{Cohen's } d < 0.8$) to large ($\text{Cohen's } d > 0.8$) (see Supplementary Tables [3, 4](#) for all false discovery rate (FDR)-corrected P values and effect sizes). Clinical case-control differences in cortical thickness and surface area generally followed the same trend as volume differences (Supplementary Information [10](#)). Alzheimer's disease showed the greatest overall difference, with a maximum difference localized to grey matter volume in biologically female patients (median centile score = 14%, 36 percentage points difference

from CN median, corresponding to Cohen's $d = 0.88$; Fig. 4a). In addition, we generated a cumulative deviation metric, the centile Mahalanobis distance (CMD), to summarize a comparative assessment of brain morphology across all global MRI phenotypes relative to the CN group (Fig. 4b, Supplementary Information 1.6). Notably, schizophrenia ranked third overall behind Alzheimer's disease and mild cognitive impairment (MCI) on the basis of CMD (Fig. 4c). Assessment across diagnostic groups, based on profiles of the multiple centile scores for each MRI phenotype and for CMD, highlighted shared and distinct patterns across clinical conditions (Supplementary Information 10, 11). However, when examining cross-disorder similarity of multivariate centile scores, hierarchical clustering yielded three clusters broadly comprising neurodegenerative, mood and anxiety, and neurodevelopmental disorders (Supplementary Information 11).

Fig. 4: Case-control differences and heritability of centile scores.



a, Centile score distributions for each diagnostic category of clinical cases relative to the control group median (depicted as a horizontal black line).

The median deviation of centile scores in each diagnostic category is overlaid as a lollipop plot (white lines with circles corresponding to the median centile score for each group of cases). Pairwise tests for significance were based on Monte Carlo resampling (10,000 permutations) and P values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) correction across all possible case–control differences. Only significant differences from the control group (CN) median (with corrected $P < 0.001$) are highlighted with an asterisk. For a complete overview of all pairwise comparisons, see Supplementary Information [10](#), Supplementary Table [3](#). Groups are ordered by their multivariate distance from the CN group (see **c** and Supplementary Information [10.3](#)). **b**, The CMD is a summary metric that quantifies the aggregate atypicality of an individual scan in terms of all global MRI phenotypes. The schematic shows segmentation of four cerebrum tissue volumes, followed by estimation of univariate centile scores, leading to the orthogonal projection of a single participant’s scan (Sub_x) onto the four respective principal components of the CN (coloured axes and arrows). The CMD for Sub_x is then the sum of its distances from the CN group mean on all four dimensions of the multivariate space. **c**, Probability density plots of CMD across disorders. Vertical black line depicts the median CMD of the control group. Asterisks indicate an FDR-corrected significant difference from the CN group ($P < 0.001$). **d**, Heritability of raw volumetric phenotypes and their centile scores across two twin studies (Adolescent Brain Cognitive Development (ABCD) and Human Connectome Project (HCP)); Supplementary Information [19](#), see Supplementary Information [13](#) for a full overview of statistics for each individual feature in each dataset. Data are mean \pm s.e.m. (although some confidence intervals are too narrow to be seen). MCI, mild cognitive impairment. See Fig. [3](#) for other diagnostic abbreviations. FDR-corrected significance: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Across all major epochs of the lifespan^{[63](#)}, the CMD was consistently greater in cases relative to controls, irrespective of diagnostic category. The largest case–control differences across epochs occurred in late adulthood when risk for dementia increases and in adolescence, which is well-recognized as a period of increased incidence of mental health disorders (Supplementary Information [10.3](#)). In five primary studies covering the lifespan, average centile scores across global tissues were related to two metrics of premature

birth (gestational age at birth: $t = 13.164$, $P < 2 \times 10^{-16}$; birth weight: $t = 36.395$, $P < 2 \times 10^{-16}$; Supplementary Information 12), such that greater gestational age and birth weight were associated with higher average centile scores. Centile scores also showed increased twin-based heritability in two independent studies (total $N = 913$ twin pairs) compared with non-centiled phenotypes (average increase of 11.8 percentage points in narrow sense heritability (h^2) across phenotypes; Fig. 4d, Supplementary Information 13). In summary, centile normalization of brain metrics reproducibly detected case–control differences and genetic effects on brain structure, as well as long-term sequelae of adverse birth outcomes even in the adult brain¹⁰.

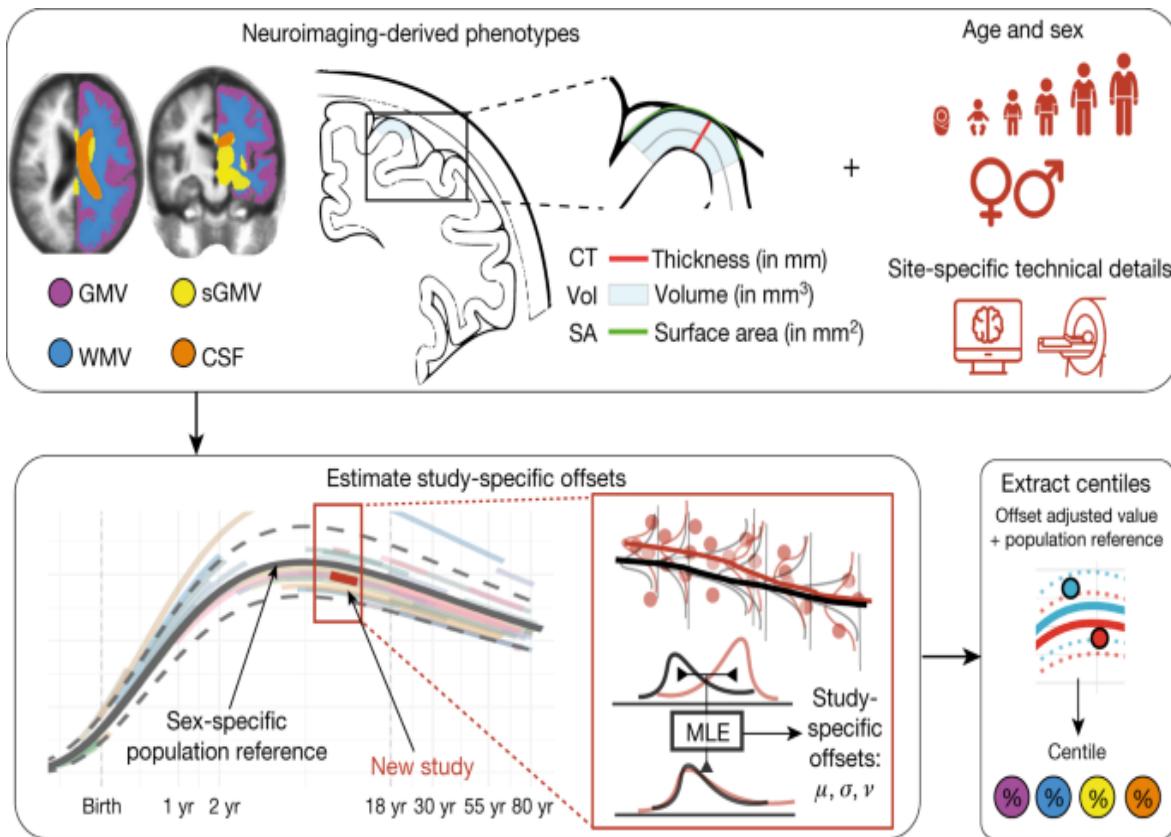
Longitudinal centile changes

Owing to the relative paucity of longitudinal imaging data (about 10% of the reference dataset), normative models were estimated from cross-sectional data collected at a single time point. However, the generalizability of cross-sectional models to longitudinal assessment is important for future research. Within-subject variability of centile scores derived from longitudinally repeated scans, measured with the interquartile range (IQR) (Methods, ‘Longitudinal stability’, Supplementary Information 1.7), was low across both clinical and CN groups (all median IQR < 0.05 centile points), indicating that centile scoring of brain structure was generally stable over time, although there was also some evidence of between-study and cross-disorder differences in within-subject variability (Supplementary Information 14). Notably, individuals who changed diagnostic categories—for example, those who progressed from mild cognitive impairment to Alzheimer’s disease over the course of repeated scanning—showed small but significant increases in within-subject variability of centile scores (Supplementary Information 14, Supplementary Tables 5, 6). Within-subject variability was also slightly higher in samples from younger individuals (Supplementary Information 14), which could reflect increased noise due to the technical or data quality challenges associated with scanning younger individuals, but is also consistent with the evidence of increased variability in earlier development observed across other anthropometric traits⁶⁴.

Centile scoring of new MRI data

A key challenge for brain charts is the accurate centile scoring of out-of-sample MRI data, not represented in the reference dataset used to estimate normative trajectories. We therefore carefully evaluated the reliability and validity of brain charts for centile scoring of such ‘new’ scans. For each new MRI study, we used maximum likelihood to estimate study-specific statistical offsets from the age-appropriate epoch of the normative trajectory; we then estimated centile scores for each individual in the new study benchmarked against the offset trajectory (Fig. 5, Methods, ‘Data-sharing and out-of-sample estimation’, Supplementary Information 1.8). Extensive jack-knife and leave-one-study-out analyses indicated that a study size of $N > 100$ scans was sufficient for stable and unbiased estimation of out-of-sample centile scores (Supplementary Information 4). This study size limit is in line with the size of many contemporary brain MRI research studies. However, these results do not immediately support the use of brain charts to generate centile scores from smaller-scale research studies, or from an individual patient’s scan in clinical practice—this remains a goal for future work. Out-of-sample centile scores proved highly reliable in multiple test-retest datasets and were robust to variations in image processing pipelines (Supplementary Information 4).

Fig. 5: Schematic overview of brain charts, highlighting methods for out-of-sample centile scoring.



Top, brain phenotypes were measured in a reference dataset of MRI scans. GAMLSS modelling was used to estimate the relationship between (global) MRI phenotypes and age, stratified by sex, and controlling for technical and other sources of variation between scanning sites and primary studies. Bottom, the normative trajectory of the median and confidence interval for each phenotype was plotted as a population reference curve. Out-of-sample data from a new MRI study were aligned to the corresponding epoch of the normative trajectory, using maximum likelihood to estimate the study specific offsets (random effects) for three moments of the underlying statistical distributions: mean (μ), variance (σ), and skewness (ν) in an age- and sex-specific manner. Centile scores of each phenotype could then be estimated for each scan in the new study, on the same scale as the reference population curve, while accounting for study-specific ‘batch effects’ on technical or other sources of variation (see Supplementary Information [1.8](#) for details). MLE, maximum likelihood estimation.

Discussion

We have aggregated the largest neuroimaging dataset to date to modernize the concept of growth charts for mapping typical and atypical human brain development and ageing. The approximately 100-year age range enabled the delineation of milestones and critical periods in maturation of the human brain, revealing an early growth epoch across its constituent tissue classes—beginning before 17 post-conception weeks, when the brain is at approximately 10% of its maximum size, and ending by age 3, when the brain is at approximately 80% of the maximum size. Individual centile scores benchmarked by normative neurodevelopmental trajectories were significantly associated with neuropsychiatric disorders as well as with dimensional phenotypes (Supplementary Information [5.2](#), [12](#)). Furthermore, imaging–genetics studies^{[65](#)} may benefit from the increased heritability of centile scores compared with raw volumetric data (Supplementary Information [13](#)). Perhaps most importantly, GAMLSS modelling enabled harmonization across technically diverse studies (Supplementary Information [5](#)), and thus unlocked the potential value of combining primary MRI studies at scale to generate normative, sex-stratified brain growth charts, and individual centile scores of typicality and atypicality.

The analogy to paediatric growth charts is not meant to imply that brain charts are immediately suitable for benchmarking or quantitative diagnosis of individual patients in clinical practice. Even for traditional anthropometric growth charts (height, weight and BMI), there are still important caveats and nuances concerning their diagnostic interpretation in individual children^{[66](#)}; similarly, it is expected that considerable further research will be required to validate the clinical diagnostic utility of brain charts. However, the current results bode well for future progress towards digital diagnosis of atypical brain structure and development^{[67](#)}. By providing an age- and sex-normalized metric, centile scores enable trans-diagnostic comparisons between disorders that emerge at different stages of the lifespan (Supplementary Information [10](#), [11](#)). The generally high stability of centile scores across longitudinal measurements also enabled assessment of brain changes related to diagnostic transition from mild cognitive impairment to Alzheimer’s disease (Supplementary Information [14](#)), which provides one example of how centile scoring could be clinically useful in quantitatively predicting or diagnosing progressive neurodegenerative disorders in the future. Our provision of appropriate normative growth charts and online tools also

creates an immediate opportunity to quantify atypical brain structure in clinical research samples, to leverage available legacy neuroimaging datasets, and to enhance ongoing studies.

Several important caveats are worth highlighting. Even this large MRI dataset was biased towards European and North American populations and European ancestry groups within those populations. This bias is unfortunately common in many clinical and scientific references, including anthropometric growth charts and benchmark genetic datasets, representing an inequity that must be addressed by the global scientific community⁶⁸. In the particular case of brain charts, further increasing ethnic, socioeconomic and demographic diversity in MRI research will enable more population-representative normative trajectories^{69,70} that can be expected to improve the accuracy and strengthen the interpretation of centile scores in relation to appropriate norms²⁶. The available reference data were also not equally distributed across all ages—for example, foetal, neonatal and mid-adulthood (30–40 years of age) epochs were under-represented (Supplementary Information [17–19](#)). Furthermore, although our statistical modelling approach was designed to mitigate study- or site-specific effects on centile scores, it cannot entirely correct for limitations of primary study design, such as ascertainment bias or variability in diagnostic criteria. Our decision to stratify the lifespan models by sex followed the analogous logic of sex-stratified anthropometric growth charts. Males have larger brain-tissue volumes than females in absolute terms (Supplementary Information [16](#)), but this is not indicative of any difference in clinical or cognitive outcomes. Future work would benefit from more detailed and dimensional self-report variables relating to sex and gender⁷¹. The use of brain charts also does not circumvent the fundamental requirement for quality control of MRI data. We have shown that GAMLSS modelling of global structural MRI phenotypes is in fact remarkably robust to inclusion of poor-quality scans (Supplementary Information [2](#)), but it should not be assumed that this level of robustness will apply to future brain charts of regional MRI or functional MRI phenotypes; therefore, the importance of quality control remains paramount.

We have focused primarily on global brain phenotypes, which were measurable in the largest achievable sample, aggregated over the widest age range, with the fewest methodological, theoretical and data-sharing

constraints. However, we have also provided proof-of-concept brain charts for regional grey matter volumetrics, demonstrating plausible heterochronicity of cortical patterning, and illustrating the potential generalizability of this approach to a diverse range of fine-grained MRI phenotypes (Fig. 2, Supplementary Information 8). As ongoing and future efforts provide increasing amounts of high-quality MRI data, we predict an iterative process of improved brain charts for an increasing number of multimodal⁷² neuroimaging phenotypes. Such diversification will require the development, implementation and standardization of additional data quality control procedures²⁷ to underpin robust brain chart modelling. To facilitate further research using our reference charts, we have provided interactive tools to explore these statistical models and to derive normalized centile scores for new datasets across the lifespan at www.brainchart.io.

Methods

Ethics

The research was reviewed by the Cambridge Psychology Research Ethics Committee (PRE.2020.104) and The Children's Hospital of Philadelphia's Institutional Review Board (IRB 20-017874) and deemed not to require PRE or IRB oversight as it consists of secondary analysis of de-identified primary datasets. Informed consent of participants (or their guardians) in primary studies is referenced in Supplementary Information 19 and Supplementary Table 1.

Model generation and specification

To accurately and comprehensively establish standardized brain reference charts across the lifespan, it is crucial to leverage multiple independent and diverse datasets, especially those spanning prenatal and early postnatal life. Here we sought to chart normative brain development and ageing across the largest age-span and largest aggregated neuroimaging dataset to date using a robust and scalable methodological framework^{2,24}. We used GAMLSS² to estimate cross-sectional normative age-related trends from 100 studies, comprising a reference dataset of more than 100,000 scans (see

Supplementary Tables 1.1–1.7 for full demographic information and Supplementary Information 19 for dataset descriptions). We optimised GAMLSS model specification and parameterization to estimate non-linear normative growth curves, their confidence intervals and first derivatives, separately for males and females, allowing for random effects on the mean and higher order moments of the outcome distributions.

The reliability of the models was assessed and endorsed by cross-validation and bootstrap resampling procedures (Supplementary Information 3). We leveraged these normative trajectories to benchmark individual scans by centile scores, which were then investigated as age-normed and sex-stratified measures of diagnostic and longitudinal atypicalities of brain structure across the lifespan.

The GAMLSS approach allowed not only modelling of age-related changes in brain phenotypes but also age related-changes in the variability of phenotypes, and in the form of both linear and nonlinear changes over time, thereby overcoming potential limitations of conventional additive models that only allow additive means to be modelled². In addition, study-specific offsets (mean and variance) for each brain phenotype were also modelled as random effects. These modelling criteria are particularly important in the context of establishing growth reference charts as recommended by the World Health Organization²⁴, as it is reasonable to assume the distribution of higher order moments (for example, variance) changes with age, sex, site/study and pre-processing pipeline, and it is impossible to circumvent some of these issues by collecting standardized data longitudinally for individuals spanning the approximately 100-year age range. Furthermore, recent studies suggest that changes in between-subject variability might intersect with vulnerability for developing a mental health condition⁷⁴. The use of data spanning the entire age range is also critical, as data from partial age-windows can bias estimation of growth charts when extrapolated to the whole lifespan. In short, using a sex-stratified approach²⁴, age, preprocessing pipeline and study were each included in the GAMLSS model estimation of first order (μ) and second order (σ) distribution parameters of a generalized gamma distribution using fractional polynomials to model nonlinear trends. See [Supplementary Information](#) for more details regarding GAMLSS model specification and estimation (Supplementary

Information 1), image quality control (Supplementary Information 2), model stability and robustness (Supplementary Information 3, 4), phenotypic validation against non-imaging metrics (Supplementary Information 3, 5.2), inter-study harmonization (Supplementary Information 5) and assessment of cohort effects (Supplementary Information 6).

More formally, the GAMLSS framework can be specified in the following way:

$$\begin{aligned}
 & \$\$ Y \sim F(\mu, \sigma, \nu, \tau) \$\$ \\
 (1) \quad & \$\$ g_{\mu}(\mu) = X_{\mu}\beta_{\mu} + Z_{\mu}\gamma_{\mu} \\
 & + \sum_i s_{\mu,i}(x_i) \$\$ \\
 & \$\$ g_{\sigma}(\sigma) = X_{\sigma}\beta_{\sigma} + Z_{\sigma}\gamma_{\sigma} \\
 & + \sum_i s_{\sigma,i}(x_i) \$\$ \\
 & \$\$ g_{\nu}(\nu) = X_{\nu}\beta_{\nu} + Z_{\nu}\gamma_{\nu} \\
 & + \sum_i s_{\nu,i}(x_i) \$\$ \\
 & \$\$ g_{\tau}(\tau) = X_{\tau}\beta_{\tau} + Z_{\tau}\gamma_{\tau} \\
 & + \sum_i s_{\tau,i}(x_i) \$\$
 \end{aligned}$$

Here, the outcome vector, \mathbf{Y} , follows a probability distribution \mathbf{F} parameterized by up to four parameters, (μ, σ, ν, τ) . The four parameters, depending on the parameterization of the probability density function, may correspond to the mean, variance, skewness, and kurtosis—that is, the first four moments. However, for many distributions there is not a direct one-to-one correspondence. Each component is linked to a linear equation through a link-function, $\{g\}_{\bullet}()$, and each component equation may include three types of terms: fixed effects, β (with design matrix X); random effects, γ (with design matrix Z); and non-parametric smoothing functions, $s_{\cdot,i}$ applied to the i th covariate for each parameter. The nature of the outcome distribution determines the appropriate link functions and which components are used. In principle any outcome distribution can be used, from well-behaved continuous and discrete outcomes, through to mixtures and truncations.

Here we have used fractional polynomials as a flexible, but not unduly complex, approach to modelling age-related changes in MRI phenotypes. Although non-parametric smoothers are more flexible, they can become

unstable and infeasible, especially in the presence of random effects. Hence, the fractional polynomials enter the model within the X terms, with associated coefficients in β . The GAMLSS framework includes the ability to estimate the most appropriate powers of fractional polynomial expansion within the iterative fitting algorithm, searching across the standard set of powers, $\{p \in \{-2, -1, \text{mathrm}{0.5}, 0, 0.5, 1, 2, 3\}\}$, where the design matrix includes the covariate (in this case, age) raised to the power, namely, (x^p) . Fractional polynomials naturally extend to higher-orders, for example a second-order fractional polynomial of the form, $(x^{p_1} + x^{p_2})$ (see Supplementary Information [1.3](#) for further details).

There are several options for including random effects within the GAMLSS framework depending on the desired covariance structures. We consider the simplest case, including a factor-level (or group-level) random intercept, where the observations are grouped by the study covariate. The random effects are drawn from a normal distribution with zero mean and variance to be estimated, $\gamma \sim N(0, \delta^2)$. The ability to include random effects is fundamental to accounting for co-dependence between observations. It is therefore possible to take advantage of the flexibility of ‘standard’ GAMLSS, as typically used to develop growth charts^{[24,62,75](#)}, while accounting for co-dependence between observations using random effects. The typical applications of GAMLSS assume independent and identically distributed outcomes; however, in this context it is essential to account for within-study covariance implying the observations are no longer independent.

The resulting models were evaluated using several sensitivity analyses and validation approaches. These models of whole-brain and regional morphometric development were robust to variations in image quality, and cross-validated by non-imaging metrics. However, we expect that several sources of variance, including but not limited to MRI data quality and variability of acquisition protocols, may become increasingly important as brain charting methods are applied to more innovative and/or anatomically fine-grained MRI phenotypes. It will be important for future work to remain vigilant about the potential impact of data quality and other sources of noise on robustness and generalizability of both normative trajectories and the centile scores derived from them.

Based on the model selection criteria, detailed in Supplementary Information 1, the final models for normative trajectories of all MRI phenotypes were specified as illustrated below for GMV:

$$\begin{array}{c}
 \text{\$\$}\backslash\text{begin}\{\text{array}\}\{\text{c}\}\{\text{\rm\{G\}}\}\{\text{\rm\{M\}}\}\{\text{\rm\{V\}}\}\backslash\text{sim }\{\text{\rm\{G\}}\}\{\text{\rm\{e\}}\} \\
 \{\text{\rm\{n\}}\}\{\text{\rm\{e\}}\}\{\text{\rm\{r\}}\}\{\text{\rm\{a\}}\}\{\text{\rm\{l\}}\}\{\text{\rm\{i\}}\}\{\text{\rm\{z\}}\}\{\text{\rm\{s\}}\} \\
 \{\text{\rm\{e\}}\}\{\text{\rm\{d\}}\}\backslash,\{\text{\rm\{G\}}\}\{\text{\rm\{a\}}\}\{\text{\rm\{m\}}\}\{\text{\rm\{m\}}\}\{\text{\rm\{a\}}\}(\mu \\
 ,\sigma ,\nu),\{\text{\rm\{w\}}\}\{\text{\rm\{i\}}\}\{\text{\rm\{t\}}\}\{\text{\rm\{h\}}\},\backslash\log (\mu)=\alpha \\
 _{\mu }+\alpha _{\mu ,1}\{\text{\rm\{s\}}\}\{\text{\rm\{e\}}\}\{\text{\rm\{x\}}\})(\{\text{\rm\{s\}}\}\{\text{\rm\{e\}}\} \\
 \{\text{\rm\{x\}}\})+\alpha _{\mu ,2}\{\text{\rm\{v\}}\}\{\text{\rm\{e\}}\}\{\text{\rm\{r\}}\})(\{\text{\rm\{v\}}\} \\
 \{\text{\rm\{e\}}\}\{\text{\rm\{r\}}\})+\beta _{\mu ,1}\{\text{\rm\{a\}}\}\{\text{\rm\{g\}}\} \\
 \{\text{\rm\{e\}}\})^{\wedge -2}+\beta _{\mu ,2}\{\text{\rm\{a\}}\}\{\text{\rm\{g\}}\}\{\text{\rm\{e\}}\})^{\wedge -2}+ \\
 \beta _{\mu ,3}\{\text{\rm\{a\}}\}\{\text{\rm\{g\}}\}\{\text{\rm\{e\}}\})^{\wedge -2},\log (\{\text{\rm\{a\}}\} \\
 \{\text{\rm\{g\}}\}\{\text{\rm\{e\}}\})^{\wedge 2}+\gamma _{\mu ,1}\{\text{\rm\{s\}}\}\{\text{\rm\{t\}}\}\{\text{\rm\{u\}}\} \\
 \{\text{\rm\{d\}}\}\{\text{\rm\{y\}}\})\backslash\log (\sigma)=\alpha _{\sigma }+\alpha _{\sigma ,1} \\
 ,\{\text{\rm\{s\}}\}\{\text{\rm\{e\}}\}\{\text{\rm\{x\}}\})(\{\text{\rm\{s\}}\}\{\text{\rm\{e\}}\}\{\text{\rm\{x\}}\})+\beta _{\sigma ,1} \\
 \{\text{\rm\{a\}}\}\{\text{\rm\{g\}}\}\{\text{\rm\{e\}}\})^{\wedge -2}+\beta _{\sigma ,2} \\
 \{\text{\rm\{a\}}\}\{\text{\rm\{g\}}\}\{\text{\rm\{e\}}\})^{\wedge 3}+\gamma _{\sigma ,1}\{\text{\rm\{s\}}\} \\
 \{\text{\rm\{t\}}\}\{\text{\rm\{u\}}\}\{\text{\rm\{d\}}\}\{\text{\rm\{y\}}\})\backslash\nu =\alpha _{\nu } \\
 \}\backslash\text{end}\{\text{array}\}\$\$ \\
 \end{array}$$

(2)

For each component of the generalized gamma distribution, α terms correspond to fixed effects of the intercept, sex (female or male), and software version used for pre-processing (five categories); β terms correspond to the fixed effects of age, modelled as fractional polynomial functions with the number of terms reflecting the order of the fractional polynomials; and γ terms correspond to the study-level random effects. Note that we have explicitly included the link-functions for μ and σ (since these parameters must be positive) and the identity for ν .

Similarly for the other global MRI phenotypes:

$$\begin{array}{c}
 \text{\$\$}\backslash\text{begin}\{\text{array}\}\{\text{c}\}\{\text{\rm\{W\}}\}\{\text{\rm\{M\}}\}\{\text{\rm\{V\}}\}\backslash\text{sim }\{\text{\rm\{G\}}\}\{\text{\rm\{e\}}\} \\
 \{\text{\rm\{n\}}\}\{\text{\rm\{e\}}\}\{\text{\rm\{r\}}\}\{\text{\rm\{a\}}\}\{\text{\rm\{l\}}\}\{\text{\rm\{i\}}\}\{\text{\rm\{s\}}\}\{\text{\rm\{e\}}\} \\
 \{\text{\rm\{d\}}\}\backslash,\{\text{\rm\{G\}}\}\{\text{\rm\{a\}}\}\{\text{\rm\{m\}}\}\{\text{\rm\{m\}}\}\{\text{\rm\{a\}}\}(\mu ,\sigma
 \end{array}$$

$$\begin{aligned}
& \backslash \nu) \backslash \rm{rm} \{ w \} \backslash \rm{rm} \{ i \} \backslash \rm{rm} \{ t \} \backslash \rm{rm} \{ h \} \backslash \log (\mu) = \alpha _{\mu } + \\
& \alpha _{\mu } \mu , \rm{rm} \{ s \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ x \}) (\rm{rm} \{ s \} \backslash \rm{rm} \{ e \} \\
& \backslash \rm{rm} \{ x \}) + \alpha _{\mu } \mu , \rm{rm} \{ v \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ r \}) (\rm{rm} \{ v \} \\
& \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ r \}) + \beta _{\mu ,1} (\rm{rm} \{ a \} \backslash \rm{rm} \{ g \} \\
& \backslash \rm{rm} \{ e \})) ^{-2} + \beta _{\mu ,2} ((\rm{rm} \{ a \} \backslash \rm{rm} \{ g \} \backslash \rm{rm} \{ e \})) ^3 + \\
& \beta _{\mu ,3} ((\rm{rm} \{ a \} \backslash \rm{rm} \{ g \} \backslash \rm{rm} \{ e \})) ^3 \backslash \log (\rm{rm} \{ a \} \\
& \backslash \rm{rm} \{ g \} \backslash \rm{rm} \{ e \}) + \gamma _{\mu ,s} \mu , \rm{rm} \{ s \} \backslash \rm{rm} \{ t \} \backslash \rm{rm} \{ u \} \backslash \rm{rm} \{ d \} \\
& \backslash \rm{rm} \{ y \}) \backslash \log (\sigma) = \alpha _{\sigma } + \alpha _{\sigma } \sigma , \\
& \rm{rm} \{ s \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ x \}) (\rm{rm} \{ s \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ x \}) + \beta _{\sigma ,1} \\
& ((\rm{rm} \{ a \} \backslash \rm{rm} \{ g \} \backslash \rm{rm} \{ e \})) ^{-2} + \beta _{\sigma ,2} \\
& ((\rm{rm} \{ a \} \backslash \rm{rm} \{ g \} \backslash \rm{rm} \{ e \})) ^3 + \gamma _{\sigma ,s} \rm{rm} \{ s \} \\
& \backslash \rm{rm} \{ t \} \backslash \rm{rm} \{ u \} \backslash \rm{rm} \{ d \} \backslash \rm{rm} \{ y \}) \backslash \nu = \alpha _{\nu } \\
& \backslash \end{array} \$\$\backslash end{array} \$\$ \\
(3)$$

$$\begin{aligned}
& \$\$ \backslash begin{array}{c} \rm{rm} \{ s \} \backslash \rm{rm} \{ G \} \backslash \rm{rm} \{ M \} \backslash \rm{rm} \{ V \} \backslash \sim \rm{rm} \{ G \} \\
& \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ n \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ r \} \backslash \rm{rm} \{ a \} \backslash \rm{rm} \{ l \} \backslash \rm{rm} \{ i \} \backslash \rm{rm} \{ s \} \\
& \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ d \} , \rm{rm} \{ G \} \backslash \rm{rm} \{ a \} \backslash \rm{rm} \{ m \} \backslash \rm{rm} \{ m \} \backslash \rm{rm} \{ a \} (\mu \\
& \sigma , \nu), \rm{rm} \{ w \} \backslash \rm{rm} \{ i \} \backslash \rm{rm} \{ t \} \backslash \rm{rm} \{ h \} \backslash \log (\mu) = \alpha _{\mu } \\
& \alpha _{\mu } \mu + \alpha _{\mu } \mu , \rm{rm} \{ s \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ x \}) (\rm{rm} \{ s \} \backslash \rm{rm} \{ e \} \\
& \backslash \rm{rm} \{ x \}) + \alpha _{\mu } \mu , \rm{rm} \{ v \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ r \}) (\rm{rm} \{ v \} \\
& \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ r \}) + \beta _{\mu ,1} ((\rm{rm} \{ a \} \backslash \rm{rm} \{ g \} \\
& \backslash \rm{rm} \{ e \})) ^{-2} + \beta _{\mu ,2} ((\rm{rm} \{ a \} \backslash \rm{rm} \{ g \} \\
& \backslash \rm{rm} \{ e \})) ^{-2} \backslash \log ((\rm{rm} \{ a \} \backslash \rm{rm} \{ g \} \backslash \rm{rm} \{ e \}) + \beta _{\mu ,3} \\
& ((\rm{rm} \{ a \} \backslash \rm{rm} \{ g \} \backslash \rm{rm} \{ e \})) ^3 + \gamma _{\mu ,s} \mu , \rm{rm} \{ s \} \backslash \rm{rm} \{ t \} \\
& \backslash \rm{rm} \{ u \} \backslash \rm{rm} \{ d \} \backslash \rm{rm} \{ y \}) \backslash \log (\sigma) = \alpha _{\sigma } + \alpha _{\sigma } \sigma , \\
& \rm{rm} \{ s \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ x \}) (\rm{rm} \{ s \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ x \}) + \\
& \beta _{\sigma ,1} ((\rm{rm} \{ a \} \backslash \rm{rm} \{ g \} \backslash \rm{rm} \{ e \})) ^{-2} + \beta _{\sigma ,2} \\
& ((\rm{rm} \{ a \} \backslash \rm{rm} \{ g \} \backslash \rm{rm} \{ e \})) ^{-2} \backslash \log ((\rm{rm} \{ a \} \\
& \backslash \rm{rm} \{ g \} \backslash \rm{rm} \{ e \}) + \gamma _{\sigma ,s} \rm{rm} \{ s \} \backslash \rm{rm} \{ t \} \backslash \rm{rm} \{ u \} \\
& \backslash \rm{rm} \{ d \} \backslash \rm{rm} \{ y \})) \backslash \nu = \alpha _{\nu }, \backslash \end{array} \$\$ \\
(4)$$

$$\begin{aligned}
& \$\$ \backslash begin{array}{c} \rm{rm} \{ V \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ n \} \backslash \rm{rm} \{ t \} \backslash \rm{rm} \{ r \} \backslash \rm{rm} \{ i \} \\
& \backslash \rm{rm} \{ c \} \backslash \rm{rm} \{ l \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ s \} \backslash \sim \rm{rm} \{ G \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ n \} \\
& \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ r \} \backslash \rm{rm} \{ a \} \backslash \rm{rm} \{ l \} \backslash \rm{rm} \{ i \} \backslash \rm{rm} \{ z \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ d \} , \\
& \backslash \rm{rm} \{ G \} \backslash \rm{rm} \{ a \} \backslash \rm{rm} \{ m \} \backslash \rm{rm} \{ m \} \backslash \rm{rm} \{ a \} (\mu , \sigma , \nu), \\
& \rm{rm} \{ w \} \backslash \rm{rm} \{ i \} \backslash \rm{rm} \{ t \} \backslash \rm{rm} \{ h \} , \backslash \log (\mu) = \alpha _{\mu } + \\
& \alpha _{\mu } \mu , \rm{rm} \{ s \} \backslash \rm{rm} \{ e \} \backslash \rm{rm} \{ x \}) (\rm{rm} \{ s \} \backslash \rm{rm} \{ e \})
\end{aligned}$$

$$\begin{aligned}
& \{\rm{x}\}) + \{\alpha\}_{\mu} \{\rm{v}\} \{\rm{e}\} \{\rm{r}\}) (\{\rm{v}\}) \\
& \{\rm{e}\} \{\rm{r}\}) + \{\beta\}_{\mu,1} (\{\rm{a}\} \{\rm{g}\}) \\
& \{\rm{e}\})}^3 + \{\beta\}_{\mu,2} (\{\rm{a}\} \{\rm{g}\}) \\
& \{\rm{e}\})}^3 \backslash \log (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\}) + \{\beta\}_{\mu,3} \\
& (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\})}^3 \backslash \log (\{\rm{a}\} \{\rm{g}\}) \\
& \{\rm{e}\})}^2 + \{\gamma\}_{\mu} \{\rm{s}\} \{\rm{t}\} \{\rm{u}\} \{\rm{d}\} \\
& \{\rm{y}\}) \backslash \log (\sigma) = \{\alpha\}_{\sigma} + \{\alpha\}_{\sigma}, \\
& \{\rm{s}\} \{\rm{e}\} \{\rm{x}\}) (\{\rm{s}\} \{\rm{e}\} \{\rm{x}\}) + \{\beta\}_{\sigma,1} (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\})}^{-2} + \{\beta\}_{\sigma,2} \\
& (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\})}^{-2} \backslash \log (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\}) + \\
& \{\beta\}_{\sigma,3} (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\})}^{-2} \backslash \log (\{\rm{a}\} \\
& \{\rm{g}\} \{\rm{e}\})}^2 \backslash \nu = \{\alpha\}_{\nu}, \end{array} \quad \$\$ \\
(5)$$

$$\begin{aligned}
& \$\$ \begin{array}{c} \{\rm{T}\} \{\rm{C}\} \{\rm{V}\} \sim \{\rm{G}\} \{\rm{e}\} \\
& \{\rm{n}\} \{\rm{e}\} \{\rm{r}\} \{\rm{a}\} \{\rm{l}\} \{\rm{i}\} \{\rm{z}\} \{\rm{e}\} \\
& \{\rm{d}\}, \{\rm{G}\} \{\rm{a}\} \{\rm{m}\} \{\rm{m}\} \{\rm{a}\} (\mu, \sigma \\
& , \nu), \{\rm{w}\} \{\rm{i}\} \{\rm{t}\} \{\rm{h}\} \backslash \log (\mu) = \{\alpha\}_{\mu} \\
& + \{\alpha\}_{\mu} \{\rm{s}\} \{\rm{e}\} \{\rm{x}\}) (\{\rm{s}\} \{\rm{e}\}) \\
& \{\rm{x}\}) + \{\alpha\}_{\mu} \{\rm{v}\} \{\rm{e}\} \{\rm{r}\}) (\{\rm{v}\}) \\
& \{\rm{e}\} \{\rm{r}\}) + \{\beta\}_{\mu,1} (\{\rm{a}\} \{\rm{g}\}) \\
& \{\rm{e}\})}^{-2} + \{\beta\}_{\mu,2} (\{\rm{a}\} \{\rm{g}\}) \\
& \{\rm{e}\})}^{-2} \backslash \log (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\}) + \{\beta\}_{\mu,3} \\
& (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\})}^3 + \{\gamma\}_{\mu} \{\rm{s}\} \{\rm{t}\} \\
& \{\rm{u}\} \{\rm{d}\} \{\rm{y}\}) \backslash \log (\sigma) = \{\alpha\}_{\sigma} + \{\alpha\}_{\sigma}, \\
& \{\sigma\}_{\sigma,1} (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\})}^{-2} + \{\beta\}_{\sigma,2} \\
& (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\})}^{-2} \backslash \log (\{\rm{a}\} \\
& \{\rm{g}\} \{\rm{e}\})}^{-2} \backslash \log ((\{\rm{a}\} \{\rm{g}\} \{\rm{e}\}))}^2 + \{\gamma\}_{\sigma,3} \\
& (\{\rm{a}\} \{\rm{t}\} \{\rm{u}\} \{\rm{d}\} \{\rm{y}\}) \backslash \nu = \{\alpha\}_{\nu} \end{array} \quad \$\$ \\
(6)$$

$$\begin{aligned}
& \$\$ \begin{array}{c} \{\rm{S}\} \{\rm{A}\} \sim \{\rm{G}\} \{\rm{e}\} \{\rm{n}\} \\
& \{\rm{e}\} \{\rm{r}\} \{\rm{a}\} \{\rm{l}\} \{\rm{i}\} \{\rm{s}\} \{\rm{e}\} \{\rm{d}\}, \\
& \{\rm{G}\} \{\rm{a}\} \{\rm{m}\} \{\rm{m}\} \{\rm{a}\} (\mu, \sigma, \nu), \\
& \{\rm{w}\} \{\rm{i}\} \{\rm{t}\} \{\rm{h}\}, \backslash \log (\mu) = \{\alpha\}_{\mu} \\
& + \{\alpha\}_{\mu} \{\rm{s}\} \{\rm{e}\} \{\rm{x}\}) (\{\rm{s}\} \{\rm{e}\}) \end{array} \quad \$\$ \end{aligned}$$

$$\begin{aligned}
& \{\rm{x}\}) + \{\alpha\}_{\mu} \{\rm{v}\} \{\rm{e}\} \{\rm{r}\}) (\{\rm{v}\}) \\
& \{\rm{e}\} \{\rm{r}\}) + \{\beta\}_{\mu,1} (\{\rm{a}\} \{\rm{g}\}) \\
& \{\rm{e}\})}^{\{-2\}} \backslash \backslash \backslash + \{\beta\}_{\mu,2} (\{\rm{a}\} \{\rm{g}\}) \\
& \{\rm{e}\})}^{\{-2\}} \backslash \log (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\}) + \{\beta\}_{\mu,3} \\
& (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\})}^{\{-2\}} \backslash \log (\{\rm{a}\} \{\rm{g}\}) \\
& \{\rm{e}\})}^{\{2\}} + \{\gamma\}_{\mu, \{\rm{s}\} \{\rm{t}\} \{\rm{u}\} \{\rm{d}\}} \\
& \{\rm{y}\}) \backslash \log (\sigma) = \{\alpha\}_{\sigma} + \{\alpha\}_{\sigma}, \\
& \{\rm{s}\} \{\rm{e}\} \{\rm{x}\}) (\{\rm{s}\} \{\rm{e}\} \{\rm{x}\}) + \{\beta\}_{\sigma,1} \\
& (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\})}^{\{-2\}} + \{\beta\}_{\sigma,2} \\
& (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\})}^{\{-2\}} \backslash \log (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\}) + \\
& \{\beta\}_{\sigma,3} (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\})}^{\{-2\}} \backslash \log (\{\rm{a}\} \\
& \{\rm{g}\} \{\rm{e}\})}^{\{2\}} + \{\gamma\}_{\sigma, \{\rm{s}\} \{\rm{t}\} \{\rm{u}\} \\
& \{\rm{d}\} \{\rm{y}\})} \backslash \nu = \{\alpha\}_{\nu}. \end{aligned}$$

(7)

$$\begin{aligned}
& \$\$ \begin{array}{l} \{\rm{C}\} \{\rm{T}\} \sim \{\rm{G}\} \{\rm{e}\} \{\rm{n}\} \\
& \{\rm{e}\} \{\rm{r}\} \{\rm{a}\} \{\rm{l}\} \{\rm{i}\} \{\rm{z}\} \{\rm{e}\} \{\rm{d}\} \backslash, \\
& \{\rm{G}\} \{\rm{a}\} \{\rm{m}\} \{\rm{m}\} \{\rm{a}\} (\mu, \sigma, \nu), \\
& \{\rm{w}\} \{\rm{i}\} \{\rm{t}\} \{\rm{h}\} \backslash \log (\mu) = \{\alpha\}_{\mu} + \\
& \{\alpha\}_{\mu, \{\rm{s}\} \{\rm{e}\} \{\rm{x}\}) (\{\rm{s}\} \{\rm{e}\}) \\
& \{\rm{x}\}) + \{\alpha\}_{\mu, \{\rm{v}\} \{\rm{e}\} \{\rm{r}\}) (\{\rm{v}\}) \\
& \{\rm{e}\} \{\rm{r}\}) + \{\beta\}_{\mu,1} (\{\rm{a}\} \{\rm{g}\}) \\
& \{\rm{e}\})}^{\{-2\}} \backslash \backslash \backslash + \{\beta\}_{\mu,2} (\{\rm{a}\} \{\rm{g}\}) \\
& \{\rm{e}\})}^{\{-2\}} \backslash \log (\{\rm{a}\} \{\rm{g}\} \{\rm{e}\}) + \{\gamma\}_{\mu, \\
& \{\rm{s}\} \{\rm{t}\} \{\rm{u}\} \{\rm{d}\} \{\rm{y}\})} \backslash \log (\sigma) = \{\alpha\}_{\sigma} + \\
& \{\alpha\}_{\sigma,1} (\{\rm{a}\} \{\rm{g}\}) \\
& \{\rm{e}\})}^{\{-1\}} + \{\beta\}_{\sigma,2} (\{\rm{a}\} \{\rm{g}\}) \\
& \{\rm{e}\})}^{\{0.5\}} + \{\gamma\}_{\sigma, \{\rm{s}\} \{\rm{t}\} \{\rm{u}\} \\
& \{\rm{d}\} \{\rm{y}\})} \backslash \nu = \{\alpha\}_{\nu}. \end{array} \$\$ \\
& \end{aligned}$$

(8)

No smoothing terms were used in any GAMLSS models implemented in this study, although the fractional polynomials can be regarded as effectively a parametric form of smoothing. Reliably estimating higher order moments requires increasing amounts of data, hence none of our models specified any age-related fixed-effects or random effects in the $\{\nu\}$ term. However, $\{\alpha\}_{\nu}$ was found to be important in terms of model fit and

hence we have used a generalized gamma distribution (Supplementary Information 1).

Defining developmental milestones

GAMLSS modelling also allowed us to leverage the aggregated life-spanning neuroimaging dataset to derive developmental milestones (that is, peaks of trajectories) and compare them to existing literature. The cerebrum tissue classes from 100 studies (Fig. 1, Supplementary Tables 1.1–1.7, Supplementary Information 18) showed clear, predominantly age-related trends, even prior to any modelling. Comparing these models with multiple non-MRI metrics of brain size demonstrated high correspondence across the lifespan (Supplementary Information 3). Peaks were determined based on the GAMLSS model output (50th centile) for each of the tissue classes and TCV, for both total tissue volumes and rates of change or growth (velocity). A similar series of methodological steps was performed for the set of extended global and regional cortical morphometric phenotypes (Fig. 2, Supplementary Information 7, 8). To further contextualize the neuroimaging trajectories, diagnostic age ranges from previous literature^{73,76} (blue boxes in Fig. 3) were compared with empirical age ranges of patients with a given diagnosis across the aggregated neuroimaging dataset (black boxes in Fig. 3). Note that age of diagnosis is significantly later than age of symptom onset for many disorders⁷³. Developmental milestones were also compared to published work for brain resting metabolic rate⁶¹, from its minimum in infancy to its maximum in early childhood; anthropometric variables (height and weight), which reach a first peak in velocity during infancy and a second peak in velocity in adolescence⁵²; typical acquisition of the six gross motor capabilities⁶²; and pubertal age ranges as defined based on previous reports^{51,53}.

Centile scores and case–control differences

These normative trajectories of brain development and aging also enabled each individual scan to be quantified in terms of its relative distance from the median of the age-normed and sex-stratified distributions provided by the reference model^{67,77} (Fig. 4, Supplementary Information 10, 11).

Individual centile scores were estimated relative to the reference curves, in a

way that is conceptually similar to traditional anthropometric growth charts (Supplementary Information 1). These centiles represent a novel set of population- and age-standardized clinical phenotypes, providing the capacity for cross-phenotype, cross-study and cross-disorder comparison. A single multivariate metric (CMD, Supplementary Information 16) was estimated by combining centile scores on multiple MRI phenotypes for each individual (Fig. 4c). Case-control differences in centile scores were analysed with a bootstrapped (500 bootstraps) non-parametric generalization of Welch's one-way ANOVA. Pairwise, sex stratified, post-hoc comparisons were conducted using non-parametric Monte Carlo permutation tests (10,000 permutations) and thresholded at a Benjamini–Hochberg FDR of $q < 0.05$.

Longitudinal stability

To use centile scores in a diagnostically meaningful or predictive way, they need to be stable across multiple measuring points. To assess this intra-individual stability, we calculated the subject-specific IQR of centiles across timepoints for the datasets that included longitudinal scans ($N = 9,306$, 41 unique studies). Exploratory longitudinal clinical analyses were restricted to clinical groups that had at least 50 subjects with longitudinal data to allow for robust group-wise estimates of longitudinal variability. In addition, there was a subset of individuals with documented clinical progression over the course of longitudinal scans, for instance from mild cognitive impairment to Alzheimer's disease, where we expected an associated change in centile scored brain structure. To test this hypothesis, we assessed whether these individuals showed longitudinal variation of centile scores (as assessed with IQR) with a direction of change consistent with their clinical progression. See Supplementary Information 14 for further details about the longitudinal stability of centile scores.

Data sharing and out-of-sample estimation

We have provided an interactive tool (www.brainchart.io) and made our code and models openly available (<https://github.com/brainchart/Lifespan>). The tool allows the user to visualize the underlying demographics of the primary studies and to explore the normative brain charts in a much more detailed fashion than static images allow. It also provides the opportunity for

interactive exploration of case–control differences in centile scores across many diagnostic categories that is beyond the scope of this paper. Perhaps most significantly, the brain chart interactive tool includes an out-of-sample estimator of model parameters for new MRI data that enables the user to compute centile scores for their own datasets without the computational or data-sharing hurdles involved in adding that data to the reference dataset used to estimate normative charts (Fig. 5). Bias and reliability of out-of-sample centile scoring was extensively assessed and endorsed by resampling and cross-validation studies for ‘new’ studies comprising at least 100 scans. Although already based on the largest and most comprehensive neuroimaging dataset to date, and supporting analyses of out-of-sample data, these normative brain charts will continue to be updated as additional data are made available for aggregation with the reference dataset. See Supplementary Information 1.8, 4 for further details about out-of-sample estimation.

Reporting summary

Further information on research design is available in the [Nature Research Reporting Summary](#) linked to this paper.

Data availability

Model parameters and out-of-sample centile scores are available at www.brainchart.io and on <https://github.com/brainchart/Lifespan>. Summary statistics are available in the Supplementary Tables (Supplementary Tables 1–8). Links to open datasets are also listed on <https://github.com/brainchart/Lifespan>. Availability of other MRI datasets aggregated here is through application procedures individually managed at the discretion of each primary study, with additional information provided in Supplementary Table 1.1 and Supplementary Information 19.

Code availability

All code is available at <https://github.com/brainchart/Lifespan>.

References

1. Cole, T. J. The development of growth references and growth charts. *Ann. Hum. Biol.* **39**, 382–394 (2012).
2. Stasinopoulos, D. & Rigby, R. Generalized additive models for location scale and shape (GAMLSS) in *R. J. Stat. Softw.* **23**, 1–46 (2007).
3. Gilmore, J. H., Knickmeyer, R. C. & Gao, W. Imaging structural and functional brain development in early childhood. *Nat. Rev. Neurosci.* **19**, 123–137 (2018).
4. Tau, G. Z. & Peterson, B. S. Normal development of brain circuits. *Neuropsychopharmacology* **35**, 147–168 (2010).

5. Grydeland, H. et al. Waves of maturation and senescence in micro-structural MRI markers of human cortical myelination over the lifespan. *Cereb. Cortex* **29**, 1369–1381 (2019).
6. Paus, T., Keshavan, M. & Giedd, J. N. Why do many psychiatric disorders emerge during adolescence? *Nat. Rev. Neurosci.* **9**, 947–957 (2008).
7. Jack, C. R. Jr et al. Tracking pathophysiological processes in Alzheimer’s disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol.* **12**, 207–216 (2013).
8. Volpe, J. J. Brain injury in premature infants: a complex amalgam of destructive and developmental disturbances. *Lancet Neurol.* **8**, 110–124 (2009).
9. Nosarti, C. et al. Preterm birth and psychiatric disorders in young adult life. *Arch. Gen. Psychiatry* **69**, E1–E8 (2012).
10. Wheater, E. et al. Birth weight is associated with brain tissue volumes seven decades later but not with MRI markers of brain ageing. *Neuroimage Clin.* **31**, <https://doi.org/10.1016/j.nicl.2021.102776> (2021).
11. Vigo, D., Thornicroft, G. & Atun, R. Estimating the true global burden of mental illness. *Lancet Psychiatry* **3**, 171–178 (2016).
12. Marquand, A. F. et al. Conceptualizing mental disorders as deviations from normative functioning. *Mol. Psychiatry* **24**, 1415–1424 (2019).
13. Brewer, J. B. Fully-automated volumetric MRI with normative ranges: translation to clinical practice. *Behav. Neurol.* **21**, 21–28 (2009).
14. Heuer, K. & Toro, R. Role of mechanical morphogenesis in the development and evolution of the neocortex. *Phys. Life Rev.* **31**, 233–239 (2019).

15. Werling, D. M. et al. Whole-genome and RNA sequencing reveal variation and transcriptomic coordination in the developing human prefrontal cortex. *Cell Rep.* **31**, 107489 (2020).
16. Gandal, M. J., Leppa, V., Won, H., Parikhshak, N. N. & Geschwind, D. H. The road to precision psychiatry: translating genetics into disease mechanisms. *Nat. Neurosci.* **19**, 1397–1407 (2016).
17. Opel, N. et al. Cross-disorder analysis of brain structural abnormalities in six major psychiatric disorders: a secondary analysis of mega- and meta-analytical findings from the ENIGMA consortium. *Biol. Psychiatry* **88**, 678–686 (2020).
18. Li, X. et al. Moving beyond processing and analysis-related variation in neuroscience. Preprint at <https://doi.org/10.1101/2021.12.01.470790> (2021).
19. Peterson, M. R. et al. Normal childhood brain growth and a universal sex and anthropomorphic relationship to cerebrospinal fluid. *J. Neurosurg. Pediatr.* **9**, 458–468 (2021).
20. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
21. Casey, B. J. et al. The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).
22. Zöllei, L., Iglesias, J. E., Ou, Y., Grant, P. E. & Fischl, B. Infant FreeSurfer: an automated segmentation and surface extraction pipeline for T1-weighted neuroimaging data of infants 0–2 years. *Neuroimage* **218**, 116946 (2020).
23. Kim, H. et al. NEOCIVET: towards accurate morphometry of neonatal gyration and clinical applications in preterm newborns. *Neuroimage* **138**, 28–42 (2016).

24. Borghi, E. et al. Construction of the World Health Organization child growth standards: selection of methods for attained growth curves. *Stat. Med.* **25**, 247–265 (2006).
25. Pomponio, R. et al. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *Neuroimage* **208**, 116450 (2020).
26. Dong, H.-M. et al. Charting brain growth in tandem with brain templates at school age. *Sci. Bull.* **65**, 1924–1934 (2020).
27. Zuo, X.-N. et al. Human connectomics across the life span. *Trends Cogn. Sci.* **21**, 32–45 (2017).
28. Chen, H. et al. Quantile rank maps: a new tool for understanding individual brain development. *Neuroimage* **111**, 454–463 (2015).
29. Frangou, S. et al. Cortical thickness across the lifespan: data from 17,075 healthy individuals aged 3–90 years. *Hum. Brain Mapp.* **43**, 431–451 (2021).
30. Habes, M. et al. The Brain Chart of Aging: machine-learning analytics reveals links between brain aging, white matter disease, amyloid burden, and cognition in the iSTAGING consortium of 10,216 harmonized MR scans. *Alzheimers Dement.* **17**, 89–102 (2021).
31. Zhang, H. et al. Growth charts for individualized evaluation of brain morphometry for preschool children. Preprint at <https://doi.org/10.1101/2021.04.08.21255068> (2021).
32. Milham, M., Fair, D., Mennes, M. & Mostofsky, S. The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Front. Syst. Neurosci.* **6**, 62 (2012).
33. Di Martino, A. et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **19**, 659–667 (2014).

34. Snoek, L. et al. AOMIC-PIOP1. *OpenNeuro* <https://doi.org/10.18112/OPENNEURO.DS002785.V2.0.0> (2020).
35. Bilder, R. et al. UCLA Consortium for Neuropsychiatric Phenomics LA5c Study. *OpenNeuro* <https://doi.org/10.18112/OPENNEURO.DS000030.V1.0.0> (2020).
36. Nastase, S. A. et al. Narratives. *OpenNeuro* <https://doi.org/10.18112/OPENNEURO.DS002345.V1.1.4> (2020).
37. Alexander, L. M. et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci. Data* **4**, 170181 (2017).
38. Snoek, L. et al. AOMIC-PIOP2. *OpenNeuro* <https://doi.org/10.18112/OPENNEURO.DS002790.V2.0.0> (2020).
39. Richardson, H., Lisandrelli, G., Riobueno-Naylor, A. & Saxe, R. Development of the social brain from age three to twelve years. *Nat. Commun.* **9**, 1027 (2018).
40. Kuklisova-Murgasova, M. et al. A dynamic 4D probabilistic atlas of the developing brain. *Neuroimage* **54**, 2750–2763 (2011).
41. Snoek, L. et al. AOMIC-ID1000. OpenNeuro <https://doi.org/10.18112/OPENNEURO.DS003097.V1.2.1> (2021).
42. Reynolds, J. E., Long, X., Paniukov, D., Bagshawe, M. & Lebel, C. Calgary preschool magnetic resonance imaging (MRI) dataset. *Data Brief*. **29**, 105224 (2020).
43. Knickmeyer, R. C. et al. A structural MRI study of human brain development from birth to 2 years. *J. Neurosci.* **28**, 12176–12182 (2008).
44. Gilmore, J. H. et al. Individual variation of human cortical structure is established in the first year of life. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **5**, 971–980 (2020).

45. Courchesne, E. et al. Normal brain development and aging: quantitative analysis at in vivo MR imaging in healthy volunteers. *Radiology* **216**, 672–682 (2000).
46. Narvacan, K., Treit, S., Camicioli, R., Martin, W. & Beaulieu, C. Evolution of deep gray matter volume across the human lifespan. *Hum. Brain Mapp.* **38**, 3771–3790 (2017).
47. Desikan, R. S. et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).
48. Vasung, L. et al. Association between quantitative MR markers of cortical evolving organization and gene expression during human prenatal brain development. *Cereb. Cortex* **31**, 3610–3621 (2021).
49. Fjell, A. M. et al. Development and aging of cortical thickness correspond to genetic organization patterns. *Proc. Natl Acad. Sci. USA* **112**, 15462–15467 (2015).
50. Sydnor, V. J. et al. Neurodevelopment of the association cortices: patterns, mechanisms, and implications for psychopathology. *Neuron* **109**, 2820–2846 (2021).
51. Tanner, J. M. *Growth at Adolescence* 2nd edn (Thomas, 1962).
52. Bozzola, M. & Meazza, C. In *Handbook of Growth and Growth Monitoring in Health and Disease* (ed. Preedy, V. R.) 2999–3011 (2012).
53. Tanner, J. M., Whitehouse, R. H. & Takaishi, M. Standards from birth to maturity for height, weight, height velocity, and weight velocity: British children, 1965. I. *Arch. Dis. Child.* **41**, 454–471 (1966).
54. Dubois, J. et al. The early development of brain white matter: a review of imaging studies in fetuses, newborns and infants. *Neuroscience* **276**, 48–71 (2014).

55. Huttenlocher, P. R. & Dabholkar, A. S. Regional differences in synaptogenesis in human cerebral cortex. *J. Comp. Neurol.* **387**, 167–178 (1997).
56. Mountcastle, V. B. The columnar organization of the neocortex. *Brain* **120**, 701–722 (1997).
57. Petanjek, Z. et al. Extraordinary neoteny of synaptic spines in the human prefrontal cortex. *Proc. Natl Acad. Sci. USA* **108**, 13281–13286 (2011).
58. Miller, D. J. et al. Prolonged myelination in human neocortical evolution. *Proc. Natl Acad. Sci. USA* **109**, 16480–16485 (2012).
59. Holland, D. et al. Structural growth trajectories and rates of change in the first 3 months of infant brain development. *JAMA Neurol.* **71**, 1266–1274 (2014).
60. Blüml, S. et al. Metabolic maturation of the human brain from birth through adolescence: insights from in vivo magnetic resonance spectroscopy. *Cereb. Cortex* **23**, 2944–2955 (2013).
61. Kuzawa, C. W. et al. Metabolic costs and evolutionary implications of human brain development. *Proc. Natl Acad. Sci. USA* **111**, 13010–13015 (2014).
62. WHO Multicentre Growth Reference Study Group. WHO Motor Development Study: windows of achievement for six gross motor development milestones. *Acta Paediatr.* **95**, 86–95 (2007).
63. Kang, H. J. et al. Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
64. Sorva, R., Lankinen, S., Tolppanen, E. M. & Perheentupa, J. Variation of growth in height and weight of children. II. After infancy. *Acta Paediatr. Scand.* **79**, 498–506 (1990).

65. Elliott, L. T. et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nat. Neurosci.* **52**, 210–216 (2018).
66. Hendrickson, M. A. & Pitt, M. B. Three areas where our growth chart conversations fall short—room to grow. *JAMA Pediatr.* **176**, 123–124 (2021).
67. Marquand, A. F., Rezek, I., Buitelaar, J. K. & Beckmann, C. F. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biol. Psychiatry* **80**, 552–561 (2016).
68. Shen, F. X. et al. Emerging ethical issues raised by highly portable MRI research in remote and resource-limited international settings. *Neuroimage* **238**, 118210 (2021).
69. Sharma, E. et al. Consortium on Vulnerability to Externalizing Disorders and Addictions (cVEDA): a developmental cohort study protocol. *BMC Psychiatry* **20**, 2 (2020).
70. Liu, S. et al. Chinese Color Nest Project: an accelerated longitudinal brain-mind cohort. *Dev. Cogn. Neurosci.* **52**, 101020 (2021).
71. Shansky, R. M. & Murphy, A. Z. Considering sex as a biological variable will require a global shift in science culture. *Nat. Neurosci.* **24**, 457–464 (2021).
72. Engemann, D. A. et al. Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. *eLife* **9**, e54055 (2020).
73. Solmi, M. et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Mol. Psychiatry* <https://doi.org/10.1038/s41380-021-01161-7> (2021).
74. Wierenga, L. M. et al. Greater male than female variability in regional brain structure across the lifespan. *Hum. Brain Mapp.* **43**, 470–499 (2020).

75. Heude, B. et al. A big-data approach to producing descriptive anthropometric references: a feasibility and validation study of paediatric growth charts. *Lancet Digit Health* **1**, e413–e423 (2019).
76. Erkkinen, M. G., Kim, M.-O. & Geschwind, M. D. Clinical neurology and epidemiology of the major neurodegenerative diseases. *Cold Spring Harb. Perspect. Biol.* **10**, a033118 (2018).
77. Bethlehem, R. A. I. et al. A normative modelling approach reveals age-atypical cortical thickness in a subgroup of males with autism spectrum disorder. *Commun Biol* **3**, 486 (2020).

Acknowledgements

R.A.I.B. was supported by a British Academy Postdoctoral fellowship and by the Autism Research Trust. J. Seidlitz was supported by NIMH T32MH019112-29 and K08MH120564. S.R.W. was funded by UKRI Medical Research Council MC_UU_00002/2 and was supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). E.T.B. was supported by an NIHR Senior Investigator award and the Wellcome Trust collaborative award for the Neuroscience in Psychiatry Network.

A.F.A.-B. was supported by NIMH K08MH120564. Data were curated and analysed using a computational facility funded by an MRC research infrastructure award (MR/M009041/1) to the School of Clinical Medicine, University of Cambridge and supported by the mental health theme of the NIHR Cambridge Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NIH, NHS, the NIHR or the Department of Health and Social Care. We acknowledge the invaluable contribution to this effort made by several openly shared MRI datasets: OpenNeuro (<https://openneuro.org/>), the Healthy Brain Network (<https://healthybrainnetwork.org/>), UK BioBank (<https://www.ukbiobank.ac.uk/>), ABCD (<https://abcdstudy.org/>), the Laboratory of NeuroImaging (<https://loni.usc.edu/>), data made available through the Open Science Framework (<https://osf.io/>), COINS (<http://coins.mrn.org/dx>), the Developing Human Connectome Project (<http://www.developingconnectome.org/>), the Human Connectome Project (<http://www.humanconnectomeproject.org/>), the OpenPain project

(<https://www.openpain.org>), the International Neuroimaging Datasharing Initiative (INDI) (https://fcon_1000.projects.nitrc.org/), and the NIMH Data Archive (<https://nda.nih.gov/>). See Supplementary Information 21 for further notes on the usage of open MRI data and data sharing. Data used in this article were provided by the brain consortium for reliability, reproducibility and replicability (3R-BRAIN)

(<https://github.com/zuoxinian/3R-BRAIN>). Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (<https://adni.loni.usc.edu/aibl-australian-imaging-biomarkers-and-lifestyle-study-of-ageing-18-month-data-now-released/>). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at <https://www.aibl.csiro.au>. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<https://adni.loni.usc.edu/>). The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. More information on the ARWIBO consortium can be found at <https://www.arwibo.it/>. More information on CALM team members can be found at <https://calm.mrc-cbu.cam.ac.uk/team/> and in the Supplementary Information. Further information about the Cam-CAN corporate authorship membership can be found at <https://www.cam-can.org/index.php?content=corpauth#12>. Data used in this article were obtained from the developmental component 'Growing Up in China' of the Chinese Color Nest Project (<http://deepneuro.bnu.edu.cn/?p=163>). Data were downloaded from the Collaborative Informatics and Neuroimaging Suite Data Exchange tool (COINS) (<https://coins.trendscenter.org/>) and data collection was performed at the Mind Research Network. Data used in the preparation of this article were obtained from the IConsortium on Vulnerability to Externalizing Disorders and Addictions (c-VEDA), India (<https://cveda-project.org/>). Details of The ENIGMA Developmental Brain Age working group can be found at <https://github.com/ENIGMA-Developmental-BrainAge/main>. Data used in the preparation of this article were obtained

from the Harvard Aging Brain Study (HABS P01AG036694) (<https://habs.mgh.harvard.edu>). Data used in the preparation of this article were obtained from the IMAGEN consortium (<https://imagen-europe.com/>). Data used in this article were obtained from the Korean Longitudinal Study on Cognitive Aging and Dementia (KLOSCAD) (<https://recode.re.kr>). A full list of NSPN consortium members can be found at <https://www.nspn.org.uk/nspn-team/>. The POND network (<https://pond-network.ca/>) is a Canadian translational network in neurodevelopmental disorders, primarily funded by the Ontario Brain Institute.

Author information

Author notes

1. These authors contributed equally: R. A. I. Bethlehem, J. Seidlitz, S. R. White
2. These authors jointly supervised: E. T. Bullmore, A. F. Alexander-Bloch

Affiliations

1. Autism Research Centre, Department of Psychiatry, University of Cambridge, Cambridge, UK
R. A. I. Bethlehem, B. Auyeung, S. Baron-Cohen, S. A. Bedford, R. Holt & M. V. Lombardo
2. Brain Mapping Unit, Department of Psychiatry, University of Cambridge, Cambridge, UK
R. A. I. Bethlehem & M. G. Kitzbichler
3. Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA

J. Seidlitz, J. W. Vogel, R. E. Gur, R. C. Gur, T. D. Satterthwaite & A. F. Alexander-Bloch

4. Department of Child and Adolescent Psychiatry and Behavioral Science, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

J. Seidlitz & A. F. Alexander-Bloch

5. Lifespan Brain Institute, The Children's Hospital of Philadelphia and Penn Medicine, Philadelphia, PA, USA

J. Seidlitz, C. Chertavian, R. E. Gur, R. C. Gur & A. F. Alexander-Bloch

6. Department of Psychiatry, University of Cambridge, Cambridge, UK

S. R. White, I. M. Goodyer, R. N. Henson, P. B. Jones, M. G. Kitzbichler, N. Medic, S. E. Morgan, R. Romero-Garcia, L. Ronan, J. Suckling, P. E. Vértes, V. Warrier, M. L. Westwater, H. Ziauddeen & E. T. Bullmore

7. MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

S. R. White

8. Lifespan Informatics & Neuroimaging Center, University of Pennsylvania, Philadelphia, PA, USA

J. W. Vogel & T. D. Satterthwaite

9. Department of Psychology, Yale University, New Haven, CT, USA

K. M. Anderson, C. T. Ellis & N. B. Turk-Browne

10. Developmental Imaging, Murdoch Children's Research Institute, Melbourne, Victoria, Australia

C. Adamson, G. Ball & R. Beare

11. Department of Medicine, Monash University, Melbourne, Victoria, Australia

C. Adamson & R. Beare

12. UCL Great Ormond Street Institute for Child Health, London, UK

S. Adler

13. Weill Cornell Institute of Geriatric Psychiatry, Department of Psychiatry, Weill Cornell Medicine, New York, USA

G. S. Alexopoulos

14. Department of Pediatrics University of Toronto, Toronto, Canada

E. Anagnostou

15. Holland Bloorview Kids Rehabilitation Hospital, Toronto, Canada

E. Anagnostou

16. The Clinical Hospital of Chengdu Brain Science Institute, MOE Key Lab for NeuroInformation, University of Electronic Science and Technology of China, Chengdu, China

A. Areces-Gonzalez & D. Paz-Linares

17. University of Pinar del Río “Hermanos Saiz Montes de Oca”, Pinar del Río, Cuba

A. Areces-Gonzalez

18. MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK

D. E. Astle & R. N. Henson

19. Department of Psychology, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK

B. Auyeung

20. Queen's University, Department of Psychiatry, Centre for Neuroscience Studies, Kingston, Ontario, Canada

M. Ayub

21. University College London, Mental Health Neuroscience Research Department, Division of Psychiatry, London, UK

M. Ayub

22. Department of Neuropsychiatry, Seoul National University Bundang Hospital, Seongnam, Korea

J. Bae

23. Department of Paediatrics, University of Melbourne, Melbourne, Victoria, Australia

G. Ball

24. Cambridge Lifetime Asperger Syndrome Service (CLASS), Cambridgeshire and Peterborough NHS Foundation Trust, Cambridge, UK

S. Baron-Cohen

25. Centre for Addiction Medicine, National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, India

V. Benegal

26. Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

F. Beyer, A. Villringer & A. V. Witte

27. Department of Human Genetics, South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley, Edinburg, TX, USA

J. Blangero

28. MRC Centre for Reproductive Health, University of Edinburgh, Edinburgh, UK

M. Blesa Cábez, J. P. Boardman & G. Sullivan

29. Fetal and Neonatal Institute, Division of Neonatology, Children's Hospital Los Angeles, Department of Pediatrics, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

M. Borzage

30. McGill Centre for Integrative Neuroscience, Ludmer Centre for Neuroinformatics and Mental Health, Montreal Neurological Institute, Montreal, Quebec, Canada

J. F. Bosch-Bayard

31. McGill University, Montreal, Quebec, Canada

J. F. Bosch-Bayard & M. M. Chakravarty

32. Department of Brain Sciences, Imperial College London, London, UK

N. Bourke & D. Sharp

33. Care Research and Technology Centre, Dementia Research Institute, London, UK

N. Bourke

34. Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia

Institute of Technology, and Emory University, Atlanta, GA, USA

V. D. Calhoun

35. Computational Brain Anatomy (CoBrA) Laboratory, Cerebral Imaging Centre, Douglas Mental Health University Institute, Montreal, Quebec, Canada

M. M. Chakravarty

36. Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

C. Chen & R. T. Shinohara

37. Normandie Univ, UNICAEN, INSERM, U1237, PhIND
“Physiopathology and Imaging of Neurological Disorders”, Institut Blood and Brain @ Caen-Normandie, Cyceron, Caen, France

G. Chetelat, M. Delarue, B. Landeau & L. Paly

38. Singapore Institute for Clinical Sciences, Agency for Science, Technology and Research, Singapore, Singapore

Y. S. Chong

39. Department of Obstetrics and Gynaecology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Y. S. Chong

40. Centre for Medical Image Computing (CMIC), University College London, London, UK

J. H. Cole

41. Dementia Research Centre (DRC), University College London, London, UK

J. H. Cole

42. Department of Psychiatry, Trinity College, Dublin, Ireland

A. Corvin

43. Cerebral Imaging Centre, Douglas Mental Health University Institute, Verdun, Quebec, Canada

M. Costantino

44. Undergraduate program in Neuroscience, McGill University, Montreal, Quebec, Canada

M. Costantino

45. Department of Neuroscience, University of California, San Diego, San Diego, CA, USA

E. Courchesne & K. Pierce

46. Autism Center of Excellence, University of California, San Diego, San Diego, CA, USA

E. Courchesne

47. Institute of Neurodegenerative Disorders, CNRS UMR5293, CEA, University of Bordeaux, Bordeaux, France

F. Crivello & B. Mazoyer

48. Melbourne Neuropsychiatry Centre, University of Melbourne, Melbourne, Victoria, Australia

V. L. Cropley, M. A. Di Biase, J. Lv & A. Zalesky

49. The Hospital for Sick Children, Toronto, Ontario, Canada

J. Crosbie, C. F. Hammill, Z. Pausova & R. J. Schachar

50. Department of Psychiatry, School of Medicine, Pontificia Universidad Católica de Chile, Santiago, Chile

N. Crossley

51. Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

N. Crossley & P. McGuire

52. Instituto Milenio Intelligent Healthcare Engineering, Santiago, Chile

N. Crossley

53. Child and Adolescent Psychiatry Department, Robert Debré University Hospital, AP-HP, Paris, France

R. Delorme

54. Human Genetics and Cognitive Functions, Institut Pasteur, Paris, France

R. Delorme

55. Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

S. Desrivieres

56. Cerebral Imaging Centre, McGill Department of Psychiatry, Douglas Mental Health University Institute, Montreal, QC, Canada

G. A. Devenyi

57. Department of Psychiatry, McGill University, Montreal, QC, Canada

G. A. Devenyi

58. Department of Psychiatry, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

M. A. Di Biase

59. Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK

R. Dolan

60. Wellcome Centre for Human Neuroimaging, London, UK

R. Dolan & K. Wagstyl

61. Division of Developmental Paediatrics, Department of Paediatrics and Child Health, Red Cross War Memorial Children's Hospital, Cape Town, South Africa

K. A. Donald

62. Neuroscience Institute, University of Cape Town, Cape Town, South Africa

K. A. Donald & N. A. Groenewold

63. Center for Neuroimaging, Cognition & Genomics (NICOG), School of Psychology, National University of Ireland Galway, Galway, Ireland

G. Donohoe

64. Weil Family Brain and Mind Research Institute, Department of Psychiatry, Weill Cornell Medicine, New York, NY, USA

K. Dunlop & C. Lynch

65. Centre for the Developing Brain, King's College London, London, UK

A. D. Edwards

66. Evelina London Children's Hospital, London, UK

A. D. Edwards

67. MRC Centre for Neurodevelopmental Disorders, London, UK

A. D. Edwards

68. Institute of Child Development, Department of Pediatrics, Masonic Institute for the Developing Brain, University of Minnesota, Minneapolis, MN, USA

J. T. Elison, D. A. Fair & E. Feczko

69. Haskins Laboratories, New Haven, CT, USA

C. T. Ellis

70. Department of Psychiatry, Center for Behavior Genetics of Aging, University of California, San Diego, La Jolla, CA, USA

J. A. Elman, C. E. Franz & W. S. Kremen

71. Desert-Pacific Mental Illness Research Education and Clinical Center, VA San Diego Healthcare, San Diego, CA, USA

L. Eyler

72. Department of Psychiatry, University of California San Diego, Los Angeles, CA, USA

L. Eyler

73. Department of Psychiatry, University of Cambridge, and Wellcome Trust MRC Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK

P. C. Fletcher

74. Cambridgeshire and Peterborough NHS Foundation Trust, Cambridge,
UK

P. C. Fletcher, P. B. Jones, J. Suckling & H. Ziauddeen

75. Department of Clinical, Educational and Health Psychology,
University College London, London, UK

P. Fonagy

76. Anna Freud National Centre for Children and Families, London, UK

P. Fonagy

77. Cuban Center for Neuroscience, La Habana, Cuba

L. Galan-Garcia & M. J. Valdes-Sosa

78. Computational Radiology Laboratory, Boston Children's Hospital,
Boston, MA, USA

A. Gholipour & S. K. Warfield

79. Department of Child and Adolescent Psychiatry, University of
California, San Diego, San Diego, CA, USA

J. Giedd

80. Department of Psychiatry, University of California San Diego, San
Diego, CA, USA

J. Giedd

81. Department of Psychiatry, University of North Carolina, Chapel Hill,
NC, USA

J. H. Gilmore

82. Department of Psychiatry, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA

D. C. Glahn, K. Im, S. R. Mathias & A. Rodrigue

83. Harvard Medical School, Boston, MA, USA

D. C. Glahn, K. A. Johnson, A. P. Schultz & R. A. Sperling

84. Division of Newborn Medicine and Neuroradiology, Fetal Neonatal Neuroimaging and Developmental Science Center, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

P. E. Grant

85. Department of Paediatrics and Child Health, Red Cross War Memorial Children's Hospital, SA-MRC Unit on Child & Adolescent Health, University of Cape Town, Cape Town, South Africa

N. A. Groenewold & H. J. Zar

86. Weill Cornell Institute of Geriatric Psychiatry, Department of Psychiatry, Weill Cornell Medicine, New York, NY, USA

F. M. Gunning & L. W. Victoria

87. Mouse Imaging Centre, Toronto, Ontario, Canada

C. F. Hammill

88. Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Lund University, Malmö, Sweden

O. Hansson

89. Memory Clinic, Skåne University Hospital, Malmö, Sweden

O. Hansson

90. Department of Neurology, Icahn School of Medicine at Mount Sinai,
New York, NY, USA

T. Hedden

91. Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

T. Hedden

92. Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Psychiatry and Psychotherapy, Charité Campus Mitte, Berlin, Germany

A. Heinz

93. Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

K. Heuer

94. Université de Paris, Paris, France

K. Heuer & R. Toro

95. Department of Psychiatry, University of Cape Town, Cape Town, South Africa

J. Hoare

96. Department of Integrative Medicine, NIMHANS, Bengaluru, India

B. Holla

97. Accelerator Program for Discovery in Brain disorders using Stem cells (ADBS), Department of Psychiatry, NIMHANS, Bengaluru, India

B. Holla

98. Departments of Psychology and Psychiatry, Yale University, New Haven, CT, USA

A. J. Holmes

99. Radiology Research, Children's Hospital of Philadelphia, Philadelphia, PA, USA

H. Huang & M. Ouyang

100. The Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

H. Huang

101. Department of Psychiatry and Mental Health, Clinical Neuroscience Institute, University of Cape Town, Cape Town, South Africa

J. Ipser

102. Department of Radiology, Mayo Clinic, Rochester, MN, USA

C. R. Jack Jr & D. T. Jones

103. Department of Psychiatry, Universidade Federal de São Paulo, São Paulo, Brazil

A. P. Jackowski

104. National Institute of Developmental Psychiatry, Beijing, China

A. P. Jackowski

105. Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China

T. Jia

106. Key Laboratory of Computational Neuroscience and BrainInspired Intelligence (Fudan University), Ministry of Education, Shanghai, China

T. Jia

107. Centre for Population Neuroscience and Precision Medicine (PONS), Institute of Psychiatry, Psychology and Neuroscience, SGDP Centre, King's College London, London, UK

T. Jia

108. Harvard Aging Brain Study, Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

K. A. Johnson, A. P. Schultz & R. A. Sperling

109. Center for Alzheimer Research and Treatment, Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA

K. A. Johnson & R. A. Sperling

110. Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

K. A. Johnson

111. Department of Neurology, Mayo Clinic, Rochester, MN, USA

D. T. Jones

112. Department of Psychiatry, Icahn School of Medicine, Mount Sinai, NY, USA

R. S. Kahn

113. Department of Clinical Medicine, Department of Psychiatry and Turku Brain and Mind Center, FinnBrain Birth Cohort Study, University of Turku and Turku University Hospital, Turku, Finland

H. Karlsson, L. Karlsson & J. J. Tuulari

114. Centre for Population Health Research, Turku University Hospital and University of Turku, Turku, Finland

H. Karlsson & L. Karlsson

115. Institute of Development, Aging and Cancer, Tohoku University, Seirycho, Aobaku, Sendai, Japan

R. Kawashima, Y. Taki & B. Thyreau

116. Queen's University, Departments of Psychology and Psychiatry, Centre for Neuroscience Studies, Kingston, Ontario, Canada

E. A. Kelley

117. Neuropsychiatric Epidemiology Unit, Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, the Sahlgrenska Academy, Centre for Ageing and Health (AGECAP) at the University of Gothenburg, Gothenburg, Sweden

S. Kern, I. Skoog & A. Zettergren

118. Region Västra Götaland, Sahlgrenska University Hospital, Psychiatry, Cognition and Old Age Psychiatry Clinic, Gothenburg, Sweden

S. Kern & I. Skoog

119. Department of Brain and Cognitive Sciences, Seoul National University College of Natural Sciences, Seoul, South Korea

K. W. Kim

120. Department of Neuropsychiatry, Seoul National University Bundang Hospital, Seongnam, South Korea

K. W. Kim

121. Department of Psychiatry, Seoul National University College of Medicine, Seoul, South Korea

K. W. Kim

122. Institute of Human Behavioral Medicine, SNU-MRC, Seoul, South Korea

K. W. Kim

123. Section on Developmental Neurogenomics, Human Genetics Branch, National Institute of Mental Health, Bethesda, MD, USA

F. Lalonde & A. Raznahan

124. Department of Brain & Cognitive Sciences, Seoul National University College of Natural Sciences, Seoul, South Korea

S. Lee

125. Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada

J. Lerch

126. Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neuroscience, University of Oxford, Oxford, UK

J. Lerch

127. Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada

J. D. Lewis

128. The Clinical Hospital of Chengdu Brain Science Institute, University of Electronic Science and Technology of China, Chengdu, China

J. Li, W. Liao & P. A. Valdes-Sosa

129. Department of Psychiatry and Brain and Mind Research Institute,
Weill Cornell Medicine, New York, NY, USA

C. Liston

130. Laboratory for Autism and Neurodevelopmental Disorders, Center for
Neuroscience and Cognitive Systems @UniTn, Istituto Italiano di
Tecnologia, Rovereto, Italy

M. V. Lombardo

131. School of Biomedical Engineering and Brain and Mind Centre, The
University of Sydney, Sydney, New South Wales, Australia

J. Lv

132. Department of Psychology, University of Texas, Austin, TX, USA

T. T. Mallard

133. Department of Psychiatry and Neuropsychology, School of Mental
Health and Neuroscience, EURON, Maastricht University Medical
Centre, Maastricht, The Netherlands

M. Marcelis

134. Institute for Mental Health Care Eindhoven (GGzE), Eindhoven, The
Netherlands

M. Marcelis

135. McConnell Brain Imaging Centre, Montreal Neurological Institute,
McGill University, Montreal, Quebec, Canada

R. D. Markello, B. Misic & L. Vasung

136. Ludmer Centre for Neuroinformatics and Mental Health, Douglas Mental Health University Institute, Montreal, Quebec, Canada

B. Mazoyer & M. J. Meaney

137. Singapore Institute for Clinical Sciences, Singapore, Singapore

M. J. Meaney

138. Bordeaux University Hospital, Bordeaux, France

A. Mechelli

139. Department of Computer Science and Technology, University of Cambridge, Cambridge, UK

S. E. Morgan

140. The Alan Turing Institute, London, UK

S. E. Morgan & P. E. Vértes

141. Department of Psychology, School of Business, National College of Ireland, Dublin, Ireland

D. Mothersill

142. School of Psychology and Center for Neuroimaging and Cognitive Genomics, National University of Ireland Galway, Galway, Ireland

D. Mothersill

143. Department of Psychiatry, Trinity College Dublin, Dublin, Ireland

D. Mothersill

144. Department of Psychiatry, School of Medicine, Oregon Health and Science University, Portland, OR, USA

J. Nigg

145. Center for Sleep and Cognition, Yong Loo Lin School of Medicine,
National University of Singapore, Singapore, Singapore

M. Q. W. Ong, X. Qian & J. H. Zhou

146. Department of Pediatrics, Washington University in St Louis, St Louis,
MO, USA

C. Ortinau

147. Alzheimer Center Amsterdam, Department of Neurology, Amsterdam
Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC,
Amsterdam, The Netherlands

R. Ossenkoppele

148. Lund University, Clinical Memory Research Unit, Lund, Sweden

R. Ossenkoppele

149. Robarts Research Institute and The Brain and Mind Institute,
University of Western Ontario, London, Ontario, Canada

L. Palaniyappan

150. Department of Psychiatry, Federal University of Sao Poalo
(UNIFESP), Sao Poalo, Brazil

P. M. Pan

151. National Institute of Developmental Psychiatry for Children and
Adolescents (INPD), Sao Poalo, Brazil

P. M. Pan & A. Zugman

152. Melbourne Neuropsychiatry Centre, Department of Psychiatry, The
University of Melbourne and Melbourne Health, Carlton South,

Victoria, Australia

C. Pantelis

153. Melbourne School of Engineering, The University of Melbourne,
Parkville, Victoria, Australia

C. Pantelis

154. Florey Institute of Neuroscience and Mental Health, Parkville,
Victoria, Australia

C. Pantelis

155. Department of Psychiatry, Schulich School of Medicine and Dentistry,
Western University, London, Ontario, Canada

M. M. Park

156. Department of Psychiatry, Faculty of Medicine and Centre Hospitalier
Universitaire Sainte-Justine, University of Montreal, Montreal,
Quebec, Canada

T. Paus

157. Departments of Psychiatry and Psychology, University of Toronto,
Toronto, Ontario, Canada

T. Paus

158. Departments of Physiology and Nutritional Sciences, University of
Toronto, Toronto, Ontario, Canada

Z. Pausova

159. Cuban Neuroscience Center, Havana, Cuba

D. Paz-Linares

160. Department of Psychiatry, Faculty of Medicine, McGill University,
Montreal, Quebec, Canada

A. Pichet Binette & S. Villeneuve

161. Douglas Mental Health University Institute, Montreal, Quebec,
Canada

A. Pichet Binette & S. Villeneuve

162. School of Psychology, Southwest University, Chongqing, China

J. Qiu

163. Department of Biomedical Engineering, The N.1 Institute for Health,
National University of Singapore, Singapore, Singapore

A. Qiu & A. Pichet Binette

164. Department of Clinical Neurosciences, University of Cambridge,
Cambridge, UK

T. Rittman, K. A. Tsvetanov & A. Pichet Binette

165. Department of Neurology, Harvard Medical School, Boston, MA,
USA

C. K. Rollins

166. Department of Neurology, Boston Children's Hospital, Boston, MA,
USA

C. K. Rollins

167. Instituto de Biomedicina de Sevilla (IBiS) HUVR/CSIC/Universidad
de Sevilla, Dpto. de Fisiología Médica y Biofísica, Seville, Spain

R. Romero-Garcia

168. Department of Psychology and Neuroscience Institute, University of Chicago, Chicago, IL, USA

M. D. Rosenberg

169. Department of Paediatrics and Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK

D. H. Rowitch

170. Department of Psychiatry, Universidade Federal do Rio Grande do Sul (UFRGS), Hospital de Clinicas de Porto Alegre, Porto Alegre, Brazil

G. A. Salum

171. National Institute of Developmental Psychiatry (INPD), São Paulo, Brazil

G. A. Salum

172. Otto Hahn Group Cognitive Neurogenetics, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

H. L. Schaare

173. Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Juelich, Juelich, Germany

H. L. Schaare

174. Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, MA, USA

A. P. Schultz

175. Centre for Population Neuroscience and Stratified Medicine (PONS), Institute for Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai, China

G. Schumann

176. PONS-Centre, Charite Mental Health, Dept of Psychiatry and Psychotherapy, Charite Campus Mitte, Berlin, Germany

G. Schumann

177. Wallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, Gothenburg, Sweden

M. Schöll

178. Department of Psychiatry and Neurochemistry, University of Gothenburg, Gothenburg, Sweden

M. Schöll

179. Dementia Research Centre, Queen's Square Institute of Neurology, University College London, London, UK

M. Schöll

180. Care Research and Technology Centre, UK Dementia Research Institute, London, UK

D. Sharp

181. Center for Biomedical Image Computing and Analytics, Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

R. T. Shinohara

182. Departments of Neurology, Pediatrics, and Radiology, Washington University School of Medicine, St Louis, MO, USA

C. D. Smyser

183. SA MRC Unit on Risk and Resilience in Mental Disorders, Dept of Psychiatry and Neuroscience Institute, University of Cape Town, Cape Town, South Africa

D. J. Stein

184. Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

A. Stoliczyn & H. C. Whalley

185. Department of Neuroscience, Institut Pasteur, Paris, France

R. Toro & N. Traut

186. Center for Research and Interdisciplinarity (CRI), Université Paris Descartes, Paris, France

N. Traut

187. Department of Psychology, University of Cambridge, Cambridge, UK

K. A. Tsvetanov

188. Wu Tsai Institute, Yale University, New Haven, CT, USA

N. B. Turk-Browne

189. Department of Clinical Medicine, University of Turku, Turku, Finland

J. J. Tuulari

190. Turku Collegium for Science, Medicine and Technology, University of Turku, Turku, Finland

J. J. Tuulari

191. Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, U1219, CHU Bordeaux, Bordeaux, France

C. Tzourio

192. Faculty of Dental Medicine and Oral Health Sciences, McGill University, Montreal, Quebec, Canada

É. Vachon-Presseau

193. Alan Edwards Centre for Research on Pain (AECRP), McGill University, Montreal, Quebec, Canada

P. A. Valdes-Sosa

194. Institute for Neuroscience and Medicine 7, Forschungszentrum Jülich, Jülich, Germany

S. L. Valk

195. Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

S. L. Valk

196. Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, The Netherlands

T. van Amelsvoort

197. Department of Biostatistics, Vanderbilt University, Nashville, TN, USA

S. N. Vandekar

198. Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

S. Villeneuve

199. Clinic for Cognitive Neurology, University of Leipzig Medical Center, Leipzig, Germany

A. Villringer, A. V. Witte & X. N. Zuo

200. State Key Laboratory of Cognitive Neuroscience and Learning,
Beijing Normal University, Beijing, China

Y. S. Wang, N. Yang & X. N. Zuo

201. Developmental Population Neuroscience Research Center,
IDG/McGovern Institute for Brain Research, Beijing Normal
University, Beijing, China

Y. S. Wang, N. Yang & X. N. Zuo

202. National Basic Science Data Center, Beijing, China

Y. S. Wang, N. Yang & X. N. Zuo

203. Research Center for Lifespan Development of Brain and Mind,
Institute of Psychology, Chinese Academy of Sciences, Beijing, China

Y. S. Wang & N. Yang

204. Division of Clinical Geriatrics, Center for Alzheimer Research,
Department of Neurobiology, Care Sciences and Society, Karolinska
Institutet, Stockholm, Sweden

E. Westman

205. Faculty of Medicine, CRC 1052 ‘Obesity Mechanisms’, University of
Leipzig, Leipzig, Germany

A. V. Witte

206. Department of Electrical and Computer Engineering, National
University of Singapore, Singapore, Singapore

B. Yeo & J. H. Zhou

207. Centre for Sleep and Cognition and Centre for Translational MR Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

B. Yeo

208. N.1 Institute for Health & Institute for Digital Medicine, National University of Singapore, Singapore, Singapore

B. Yeo

209. Integrative Sciences and Engineering Programme (ISEP), National University of Singapore, Singapore, Singapore

H. Yun

210. Department of Biomedical Engineering, University of Melbourne, Melbourne, Victoria, Australia

H. J. Zar

211. Center for Translational Magnetic Resonance Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

J. H. Zhou

212. Wellcome Trust-MRC Institute of Metabolic Science, University of Cambridge, Cambridge, UK

H. Ziauddeen

213. National Institute of Mental Health (NIMH), National Institutes of Health (NIH), Bethesda, MD, USA

A. Zugman

214. Department of Psychiatry, Escola Paulista de Medicina, São Paulo, Brazil

A. Zugman

215. Key Laboratory of Brain and Education, School of Education Science,
Nanning Normal University, Nanning, China

X. N. Zuo

216. Memory Disorders Clinic, Austin Health, Melbourne, Victoria,
Australia

C. Rowe

217. University Hospitals and University of Geneva, Geneva, Switzerland

G. B. Frisoni

218. IRCCS Fatebenefratelli, The National Centre for Alzheimer's and
Mental Diseases, Brescia, Italy

G. B. Frisoni

Consortia

3R-BRAIN

- X. N. Zuo

AIBL

- C. Rowe

Alzheimer's Disease Neuroimaging Initiative

- C. R. Jack Jr

Alzheimer's Disease Repository Without Borders Investigators

- G. B. Frisoni

CALM Team

- D. E. Astle

Cam-CAN

- R. N. Henson

CCNP

- Y. S. Wang
- , N. Yang
- & X. N. Zuo

COBRE

- V. D. Calhoun

cVEDA

- B. Holla

ENIGMA Developmental Brain Age Working Group

- J. H. Cole
- , N. Bourke
- , H. C. Whalley
- , D. C. Glahn
- , J. Seidlitz
- , R. A. I. Bethlehem
- & A. F. Alexander-Bloch

Developing Human Connectome Project

- A. D. Edwards

FinnBrain

- H. Karlsson
- , L. Karlsson
- , J. D. Lewis
- & J. J. Tuulari

Harvard Aging Brain Study

- K. A. Johnson
- , K. A. Johnson
- , K. A. Johnson
- & K. A. Johnson

IMAGEN

- S. Desrivieres
- , A. Heinz
- , T. Jia
- & G. Schumann

KNE96

- J. Bae
- , K. W. Kim
- & S. Lee

The Mayo Clinic Study of Aging

- C. R. Jack Jr
- & D. T. Jones

NSPN

- E. T. Bullmore
- , R. Dolan
- , P. Fonagy
- , I. M. Goodyer
- & P. B. Jones

POND

- E. Anagnostou
- , M. Ayub
- , J. Crosbie
- , C. F. Hammill
- , E. A. Kelley
- , J. Lerch
- & R. J. Schachar

The PREVENT-AD Research Group

- A. Pichet Binette
- & S. Villeneuve

VETSA

- J. A. Elman
- , C. E. Franz
- & W. S. Kremen

Contributions

R.A.I.B., J.S., S.R.W., E.T.B. and A.F.A.-B. designed the study, conducted analyses, wrote and edited the manuscript. J.V. and K.M.A. helped to design the study and contributed to data analysis. All other authors made substantial contributions to the conception or design of the work, the

acquisition, analysis or interpretation of data, the creation of new software used in the work, or drafted or substantively revised the Article.

Corresponding authors

Correspondence to [R. A. I. Bethlehem](#) or [J. Seidlitz](#).

Ethics declarations

Competing interests

E.T.B. serves on the scientific advisory board of Sosei Heptares and as a consultant for GlaxoSmithKline, Boehringer Ingelheim and Monument Therapeutics. G.S.A. has served on advisory boards of Eisai and Janssen and in speakers bureaus of Allergan, Takeda and Lundbeck. K.M.A. is an employee of Neumora Therapeutics. P.B.J. has consulted for MSD. L. Palaniyappan reports personal fees from Janssen Canada for participating in an Advisory Board (2019) and Continuous Professional Development events (2017–2020), Otsuka Canada for Continuous Professional Development events (2017–2020), SPMM Course Limited, UK for preparing educational materials for psychiatrists and trainees (2010 onwards), Canadian Psychiatric Association for Continuous Professional Development events (2018–2019); book royalties from Oxford University Press (2009 onwards); institution-paid investigator-initiated educational grants with no personal remunerations from Janssen Canada, Sunovion and Otsuka Canada (2016–2019); travel support to attend a study investigator's meeting organized by Boehringer-Ingelheim (2017); travel support from Magstim Limited (UK) to speak at an academic meeting (2014); none of these activities are related to this work. T.R. has received honoraria from Oxford Biomedica. A.P.S. has consulted for Janssen, Biogen, Qynapse, and NervGen. R.T.S. has received consulting income from Octave Bioscience and compensation for scientific review duties from the American Medical Association, the US Department of Defense, the Emerson Collective, and the National Institutes of Health. R.A.S. has consulted for Janssen, AC Immune, NervGen and Genentech. D.J.S. has received research grants and/or consultancy honoraria from Discovery Vitality, Johnson & Johnson,

Lundbeck, Sanofi, Servier, Takeda and Vistagen. J. Suckling has consulted for GW Pharmaceuticals, Claritas HealthTech, Fundacion La Caixa and Fondazione Cariplo. All other authors declare no competing interests.

Peer review

Peer review information

Nature thanks Michael Harms, Michael Milham and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer review reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information

Supplementary Information

This file includes common nomenclature, Supplementary Methods, sensitivity analyses, supplementary analyses, reference database details, replication/validation datasets, a note on data sharing, affiliations, Acknowledgements and supplementary references.

Reporting Summary

Supplementary Table 1

Demographic information. SI1.1, sample size and acquisition parameters for each individual study. SI1.2–SI1.42, demographic information per study and sex for each individual brain imaging phenotype (that is, SI1.2 has demographics for all control subjects included in the GMV model, SI1.3 for

all subjects included in the WMV model, etc.). Thus these contain all demographics and sample sizes related to the models presented in Figs. 1, 2.

Supplementary Table 2

Milestones. SI2.1, information on volumetric peaks and rate of change peaks for the global neuroimaging phenotypes as presented in Figs. 1–3. SI2.2, information on volumetric peaks and rate of change peaks for the regional phenotypes as presented in Fig. 2.

Supplementary Table 3

Case–control centile comparisons. SI3.1–3.7, statistics and sample size for case–control comparisons presented in Fig. 4a (that is, for clinical groups with $N > 500$) and SI10. Each table reports the sample sizes, test-statistics, effect sizes and corrected and uncorrected P -values per case–control pair and per sex. Corrected P -values represent P -values adjusted using Benjamini–Hochberg FDR correction for multiple comparisons across all case–control pairs of a specific sex within a specific feature. SI3.8–3.14, same statistics as 3.1–3.7 for all clinical groups with $N > 250$. SI3.15–3.21, same statistics as 3.1–3.7 for all clinical groups with $N > 100$. SI3.22–3.28, same statistics as 3.1–3.7 for all clinical groups with $N > 50$.

Supplementary Table 4

Centile distributions. SI4.1–4.7, statistics on centile distributions of global neuroimaging phenotypes, presenting the estimated number of peaks and the accompanying Hartigan’s dip-test statistic for unimodality per sex per phenotype.

Supplementary Table 5

Case–control IQR comparisons. SI5.1–5.7, statistics of the case–control comparisons of the interquartile range for subjects that had longitudinal data. Each table reports the sample sizes, test-statistics, effect sizes and corrected and uncorrected P -values per case–control pair and per sex. Corrected P -values represent P -values adjusted using Benjamini–Hochberg

FDR correction for multiple comparisons across all case–control pairs of a specific sex within a specific feature.

Supplementary Table 6

Case–control IQR comparisons for subjects with diagnostic changes. SI6.1–6.7, statistics of the case–control comparisons of the interquartile range for subjects that had longitudinal data and changed diagnostic labels during the course of their longitudinal assessment. Each table reports the sample sizes, test-statistics, effect sizes and corrected and uncorrected *P*-values per case–control pair and per sex. Corrected *P*-values represent *P*-values adjusted using Benjamini–Hochberg FDR correction for multiple comparisons across all case–control pairs of a specific sex within a specific feature. SI6.8.614, demographics for the groups compared in each phenotype.

Supplementary Table 7

Leave-one-study-out statistics. SI7.1–7.7, statistics of the per study stability of centiles based on the leave-one-study-out (LOSO) analysis. For each study and each feature the Pearson correlation (and corresponding *t*-value), uncorrected *P*-value, confidence intervals and degrees of freedom are listed comparing the centiles estimated within the model versus outside the model (that is, using the out-of-sample estimation method).

Supplementary Table 8

Centile sex differences. SI8.1–8.7, statistics of linear models assessing the effects of sex on centile scores within the clinical cohorts (patient $N > 500$), including beta estimates, standard errors, degrees of freedom, *t*-value and uncorrected *P*-values for all main effects and interactions.

Peer Review File

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Reprints and Permissions

About this article

Cite this article

Bethlehem, R.A.I., Seidlitz, J., White, S.R. *et al.* Brain charts for the human lifespan. *Nature* **604**, 525–533 (2022). <https://doi.org/10.1038/s41586-022-04554-y>

- Received: 09 June 2021
- Accepted: 16 February 2022
- Published: 06 April 2022
- Issue Date: 21 April 2022
- DOI: <https://doi.org/10.1038/s41586-022-04554-y>

Share this article

Anyone you share the following link with will be able to read this content:

[Get shareable link](#)

Sorry, a shareable link is not currently available for this article.

[Copy to clipboard](#)

Provided by the Springer Nature SharedIt content-sharing initiative

Further reading

- [Your brain expands and shrinks over time — these charts show how](#)

◦ Max Kozlov

Nature (2022)

[Charting brain development across the human lifespan](#)

- Heather Wood

Nature Reviews Neurology Research Highlight 20 Apr 2022

This article was downloaded by **calibre** from <https://www.nature.com/articles/s41586-022-04554-y>.

| [Section menu](#) | [Main menu](#) |

[Download PDF](#)

- Article
- [Published: 28 February 2022](#)

Molecular basis of receptor binding and antibody neutralization of Omicron

- [Qin Hong^{1,2}](#) na1,
- [Wenyu Han^{1,2}](#) na1,
- [Jiawei Li^{1,2}](#) na1,
- [Shiqi Xu](#) [ORCID: orcid.org/0000-0003-0345-2814](#)³ na1,
- [Yifan Wang](#) [ORCID: orcid.org/0000-0001-9966-682X](#)^{1,2} na1,
- [Cong Xu](#)¹,
- [Zuyang Li^{1,2}](#),
- [Yanxing Wang](#) [ORCID: orcid.org/0000-0002-9523-0030](#)¹,
- [Chao Zhang](#) [ORCID: orcid.org/0000-0002-8658-9009](#)³,
- [Zhong Huang](#) [ORCID: orcid.org/0000-0002-2161-1359](#)³ &
- [Yao Cong](#) [ORCID: orcid.org/0000-0002-7164-8694](#)^{1,2}

[Nature](#) volume 604, pages 546–552 (2022)

- 13k Accesses
- 1 Citations
- 27 Altmetric
- [Metrics details](#)

Subjects

- [Cryoelectron microscopy](#)
- [Protein vaccines](#)

Abstract

The SARS-CoV-2 Omicron variant exhibits striking immune evasion and is spreading rapidly worldwide. Understanding the structural basis of the high transmissibility and enhanced immune evasion of Omicron is of high importance. Here, using cryo-electron microscopy, we present both the closed and the open states of the Omicron spike (S) protein, which appear more compact than the counterparts of the G614 strain¹, potentially related to enhanced inter-protomer and S1–S2 interactions induced by Omicron residue substitution. The closed state showing dominant population may indicate a conformational masking mechanism for the immune evasion of Omicron. Moreover, we captured three states for the Omicron S–ACE2 complex, revealing that the substitutions on the Omicron RBM result in new salt bridges and hydrogen bonds, more favourable electrostatic surface properties, and an overall strengthened S–ACE2 interaction, in line with the observed higher ACE2 affinity of Omicron S than of G614. Furthermore, we determined the structures of Omicron S in complex with the Fab of S3H3, an antibody that is able to cross-neutralize major variants of concern including Omicron, elucidating the structural basis for S3H3-mediated broad-spectrum neutralization. Our findings shed light on the receptor engagement and antibody neutralization or evasion of Omicron and may also inform the design of broadly effective vaccines against SARS-CoV-2.

Main

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has undergone considerable evolution since its initial discovery, leading to the emergence of several variants of concern (VOCs) including Alpha^{2,3,4,5,6}, Beta^{5,6,7,8,9,10}, Gamma¹¹ and Delta^{12,13}. These variants that have multiple mutations on their S protein show enhanced transmissibility and resistance to antibody neutralization¹³. Recently, a new variant named Omicron (B.1.1.529), which was first reported in South Africa in November 2021, was classified as the fifth VOC by the World Health Organization (WHO)¹⁴.

Omicron bears 37 mutations in its S protein relative to the original SARS-CoV-2 strain^{15,16}. As a consequence, Omicron has been observed to extensively escape neutralization by previously developed neutralizing monoclonal antibodies (mAbs) or sera from vaccinated or convalescent individuals^{15,17,18,19,20,21,22}. Among all of the Omicron S mutations, 15 are present in the receptor-binding domain (RBD) that mediates binding of the virus to its host cell receptor, angiotensin-converting enzyme 2 (ACE2), which is also a major target for neutralizing antibodies^{23,24,25,26,27}. However, Omicron still uses ACE2 as its entry receptor²². Moreover, the Omicron S appears to have an increased binding affinity to ACE2 relative to the wild-type (WT) S^{15,16,28}.

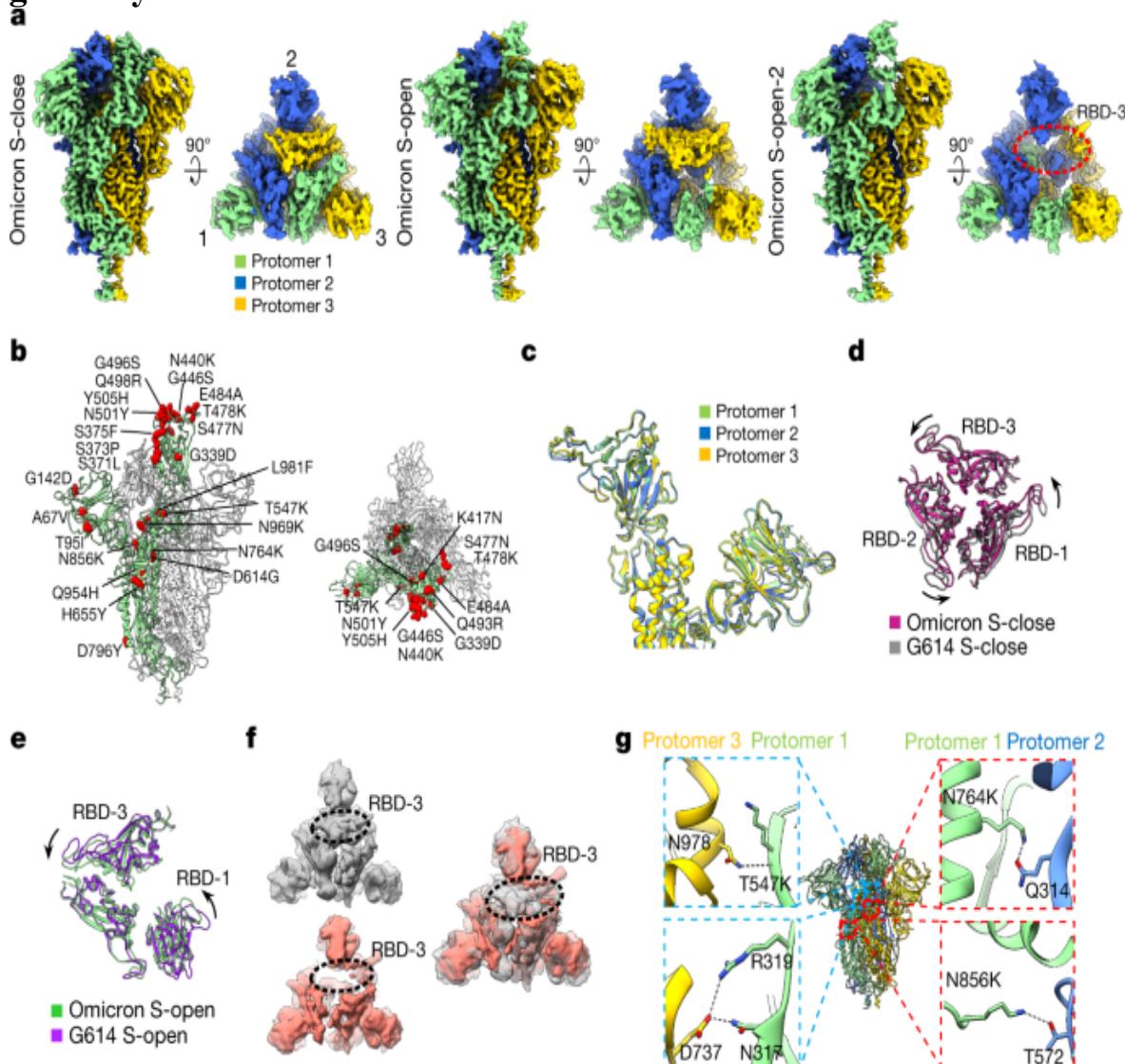
The high transmissibility and greatly enhanced resistance to antibody neutralization observed for Omicron makes this VOC particularly threatening. Therefore, further understanding of the nature of Omicron is of substantial importance and may help in developing countermeasures against this VOC. To address this from a structural aspect of how Omicron binds to the ACE2 receptor and how it recognizes or evades neutralizing antibodies raised to the original virus, here we performed cryo-electron microscopy (cryo-EM) analyses on the Omicron S trimer and its complexes with ACE2 or with neutralizing mAbs.

Closed and open states of Omicron S protein

We prepared a prefusion-stabilized trimeric S protein of the SARS-CoV-2 Omicron variant (Extended Data Fig. 1) and determined its cryo-EM structures. Three cryo-EM maps, including an all RBD down conformation (termed Omicron S-close), an RBD-1 up open conformation (termed Omicron S-open), and an RBD-1 up while RBD-3 disordered open conformation (termed Omicron S-open-2), were determined at 3.08, 3.40 and 3.41 Å resolution, respectively (Fig. 1a,b, Extended Data Fig. 2a–e, Extended Data Table 1). For the Omicron S-close state, the three protomers are well resolved and they display similar conformation with their RBDs in the down position (Fig. 1c, Extended Data Fig. 2d). The Omicron S-close appears more twisted and compact in the RBDs than the G614 S-close structure¹ (Fig. 1d). In addition, in the Omicron S-open state, the RBDs are

slightly more twisted and compact than that of the G614 S-open state¹ (Fig. 1e). There is no linoleic acid (termed LA) in the Omicron S-close, S-open and S-open-2 maps, and LA has been suggested to lead to more compact RBDs²⁹. Thus, the Omicron S trimer is more compact than that of G614, and this is not caused by binding of LA. Moreover, in the Omicron S-open-2 map, the RBD-3 density appears mostly disordered, indicating an extremely dynamic nature of RBD-3 (Fig. 1a). Further 3D variability analysis³⁰ on the Omicron S trimer dataset revealed an intrinsic rising up motion of RBD-1, which could alter the original RBD-1–RBD-3 contact and destabilize RBD-3, making it extremely dynamic and may transiently rise up (Fig. 1f, [Supplementary Video 1](#)).

Fig. 1: Cryo-EM structures of the SARS-CoV-2 Omicron S trimer.



a, Cryo-EM maps of the Omicron S-close, S-open and S-open-2 states. Protomers 1, 2 and 3 are shown in light green, royal blue and gold, respectively; this colour scheme is followed throughout. In S-open-2, the mostly disordered RBD-3 is indicated by a red dotted ellipsoid. **b**, Atomic model of the Omicron S-open, with mutations labelled and indicated by a red sphere. **c**, Side view of the overlaid protomers of the Omicron S-close. **d**, Top view of the overlaid RBDs of the Omicron S-close (violet red) and the G614 S-close (PDB: [7KRQ](#); dark grey), indicating a twist of the Omicron S-close relative to that of G614. **e**, Top view of the overlaid RBDs of the Omicron S-open (light green) and the G614 S-open (PDB: [7KRR](#); purple), indicating a twist of the Omicron S-open relative to that of G614. **f**, One representative 3D variability analysis motion of the Omicron S. The two left maps illustrate the top view of two extremes in the motion, with the RBD-3 indicated by a black dotted ellipsoid; the top view of the overlaid two extreme maps is shown on the right. **g**, Newly formed hydrogen bonds (black dashed line) in the interfaces of protomer 1–protomer 3 (blue dashed line box) and protomer 1–protomer 2 (red dashed line box) of the S-close state (see also Extended Data Fig. [2f](#)).

The population distribution of the Omicron closed and open states (S-open and S-open-2) is about 60.8% and 39.2%, respectively (Extended Data Fig. [2a](#)), displaying a considerable population shift to the closed state than that of the Beta and Kappa S protein (both around 50.0:50.0% open-transition ratio) or that of the Delta S protein (75.3:24.7% open-transition ratio) from our recent studies^{[10,31](#)}, with these S structures obtained in the same sample preparation and data processing schemes. Thus, the Omicron S trimer appears more prone to the closed state and potentially stabilized relative to the counterparts of the Delta, Kappa, Beta or G614 variants. The population distribution of the closed and open states of these S trimers varies among different studies, which could potentially be due to subtle difference in the chemical condition used by different research groups^{[32,33,34,35](#)}.

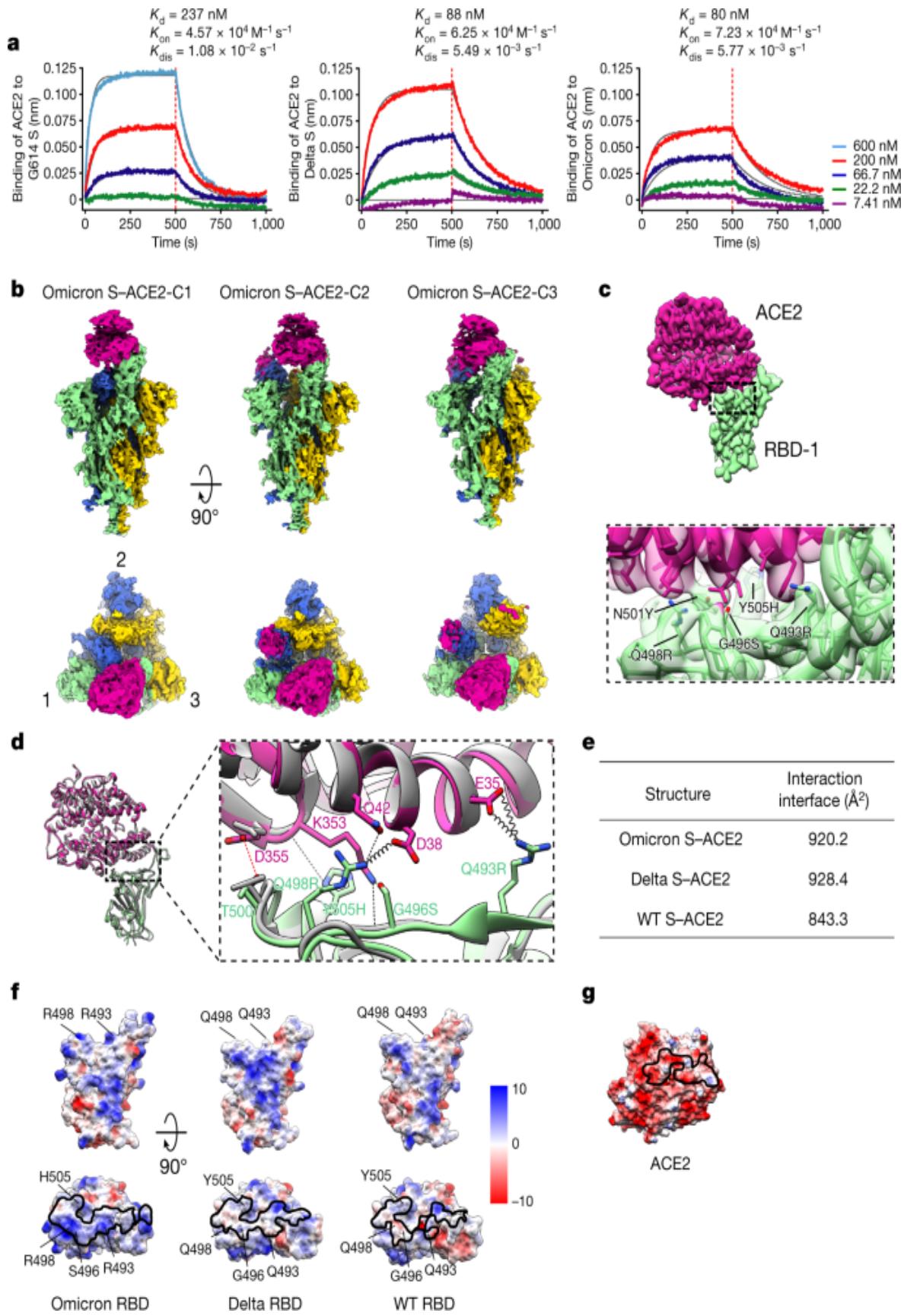
Inspection of the protomer interaction interface of Omicron S-close revealed new hydrogen bond interactions induced by the unique Omicron substitutions in the SD1 and S2 regions (Fig. [1g](#), Extended Data Fig. [2f](#), Extended Data Table [2](#)). Specifically, T547K from the SD1 subdomain of protomer 1 forms a new hydrogen bond with N978 from the S2 subunit of protomer 3, potentially enhancing the S1–S2 interaction between the two

protomers; N856K or N764K from protomer 1 can form hydrogen bonds with T572 or Q314 from protomer 2, respectively. We also observed multiple new such interactions between N317 or R319 of protomer 1 and D737 of protomer 3 (Fig. 1g, Extended Data Fig. 2f). The SD1 and SD2 of S1 is the hinge for RBD upward rotation^{31,35,36,37,38}, and disturbing the SD1–S2 interface could destabilize the interface and increase S1 mobility³⁶. Collectively, these extra hydrogen bonds mainly induced by Omicron substitutions contribute greatly to strengthen the inter-protomer and S1–S2 interactions, markedly stabilizing the Omicron S trimer and inhibiting its transformation towards the fusion-prone open state and subsequent shielding of S1.

Structure of enhanced S–ACE2 interaction

Compared with the WT strain, Omicron bears 15 mutations in the RBD region, nine of which are located in the receptor-binding motif (RBM)¹⁵. We assessed whether these mutations affect the human ACE2 receptor-binding ability of the Omicron S trimer by performing biolayer interferometry assay. We found that the ACE2-binding affinity of the Omicron S (dissociation constant (K_d) = 80 nM) is comparable to that of the Delta S (K_d = 88 nM) but is about threefold higher than that of the G614 S (K_d = 237 nM; Fig. 2a), consistent with other recent reports^{15,16,28}.

Fig. 2: Structural basis of the enhanced Omicron S–ACE2 interaction.



a, Measurement of the binding affinity between the ACE2 monomer and the S trimer of the G614 (left), Delta (middle) and Omicron (right) variants using biolayer interferometry. Biotinylated S trimers were loaded onto streptavidin sensors and were then allowed to interact with different concentrations of ACE2 (shown on the right). Raw sensograms and fitting curves are shown in colour and grey, respectively. Association and dissociation phases are divided by the red dashed lines. K_{dis} , dissociation rate; K_{on} , ‘on-rate’. **b**, Cryo-EM maps of the Omicron S–ACE2 complex in three distinct conformational states. In the S–ACE2-C2 and S–ACE2-C3 maps, the density of RBD-2-associated or RBD-3-associated ACE2 appears weaker than that of the stably associated ACE2 on RBD-1 (see also Extended Data Fig. 3c). ACE2 is shown in violet red. This colour scheme is followed throughout. **c**, Density map of the focus-refined Omicron RBD-1–ACE2 and the zoomed-in view of the RBD–ACE2 interaction interface, showing the side chain densities of Q493R, G496S, Q498R, N501Y and Y505H on the RBM. **d**, The substituted residues R493, S496, R498 and H505 of the Omicron RBM form new interactions with E35, D38, Q42 and K353 of ACE2 (the spring represents the salt bridge, and the black dashed line represents the hydrogen bond) relative to that in the WT RBD–ACE2 (PDB: [6M0J](#); dark grey). A newly formed hydrogen bond without substitution is shown by a red dashed line. **e**, Interaction interface areas between ACE2 and the RBD of Omicron, Delta (PDB: [7W9I](#)) and WT (PDB: [6M0J](#)), analysed using PISA. **f**, The electrostatic surface properties of Omicron, Delta (PDB: [7W9I](#)) and WT (PDB: [6M0J](#)) RBDs, with the mutated residues indicated. The black outlines depict the footprint of ACE2 on the RBD. **g**, The electrostatic surface property of ACE2, with residues in proximity to RBD-1 (less than 4 Å) indicated (related to Extended Data Table 3).

[Source data](#)

Next, we carried out cryo-EM analysis on the Omicron S in complex with the human ACE2 peptidase domain (Extended Data Fig. 3a). We obtained three cryo-EM maps of the S trimer engaged with ACE2, including a one RBD-up state (termed Omicron S–ACE2-C1), a two RBD-up state (termed Omicron S–ACE2-C2) and an all three RBD-up state (termed Omicron S–ACE2-C3) at 3.69, 3.70 and 4.04 Å resolution, respectively (Fig. 2b,

Extended Data Fig. 4, Extended Data Table 1). The population distribution among Omicron S-ACE2-C1, S-ACE2-C2 and S-ACE2-C3 is about 43.9%, 41.2% and 14.9%, respectively (Extended Data Fig. 3b), displaying an obvious higher one RBD-up C1 population (43.9%) than that of the Beta, Kappa or Delta variants (C1 population ranges from 8.3% to 14.1%) observed in our recent studies^{10,31}, with these S-ACE2 structures obtained in the same sample preparation and processing schemes. Together, the Omicron S trimer exhibits less ability to transform to the more RBD-up C2 or C3 states in the presence of ACE2 than that of the Beta, Kappa and Delta variants.

We further focus-refined the stably associated Omicron RBD-1-ACE2 region and obtained a 3.67 Å resolution structure (Fig. 2c, Extended Data Figs. 3, 4), which revealed many new interactions between the RBM substitutions and ACE2 compared with that of the WT RBD-ACE2 (ref. 26). Specifically, the RBM Q493R and Q498R result in three new salt bridges with the ACE2 E35 and D38, respectively; the RBM G496S and Y505H with ACE2 K353, the RBM Q498R with ACE2 Q42, and the RBM S477N with ACE2 Q19 also form new hydrogen bonds, respectively (Fig. 2d, Extended Data Table 3), generally in line with recent studies^{16,28,39,40,41,42}. Moreover, we observed an extra hydrogen bond between RBM T500 and ACE2 D355 (Fig. 2d). Our previous study showed that Y505A obviously decreased the binding affinity of ACE2 (ref. 35), thus the Omicron Y505H substitution may maintain or even enhance ACE2 binding. Meanwhile, the K417N substitution, which occurred in Omicron as well as in Beta and Delta, is known to markedly reduce the binding of ACE2 through abolishing multiple salt bridges and/or hydrogen bonds with ACE2 D30 (refs. 26,43,44). Together, these newly formed RBM-ACE2 interactions may compensate for the loss of some of the original RBM-ACE2 interactions due to the residue changes introduced into the Omicron RBM.

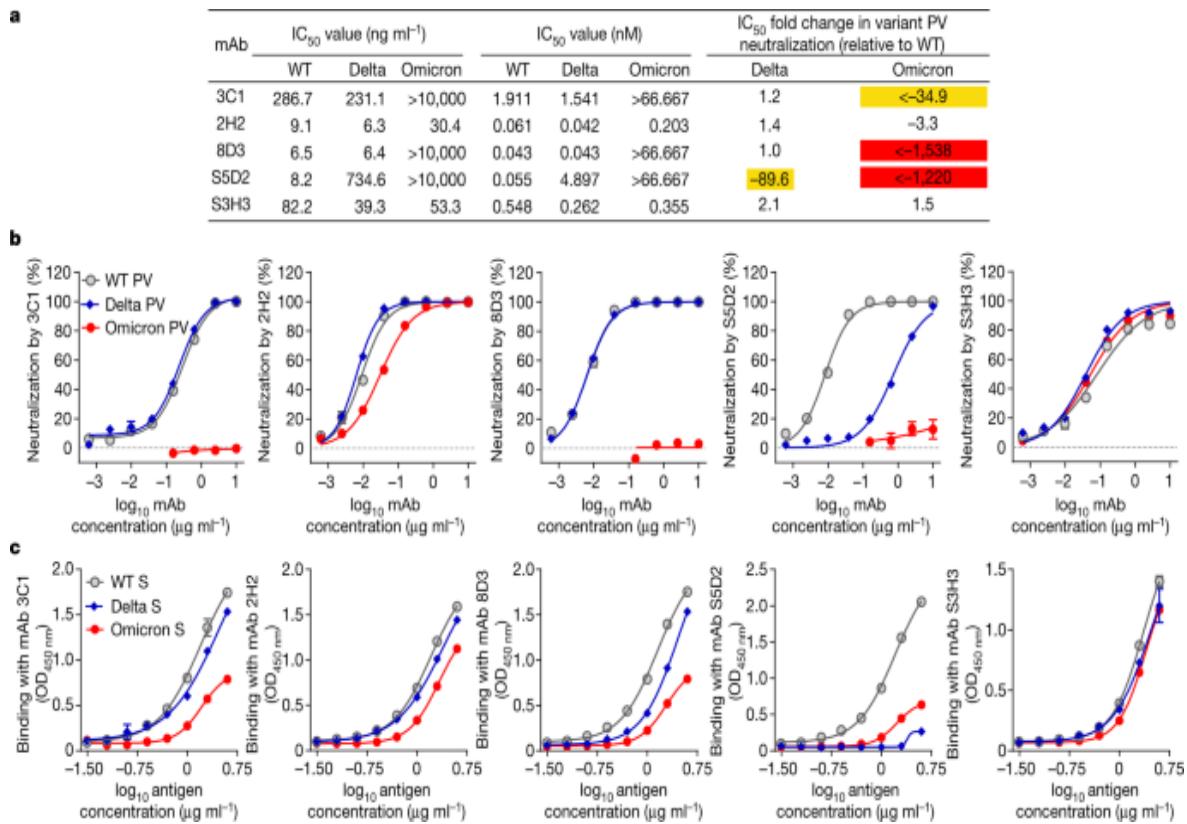
Further inspection of the surface property showed that the substitutions in RBM, especially Q493R, G496S, Q498R and Y505H, render the substituted site within the ACE2 interaction footprint more positively charged, which could strengthen the interaction of RBM with the overall negatively charged ACE2 in the interaction interface (Fig. 2f,g). Corroborating this, the Omicron RBD-ACE2 interaction area (920.2 Å²) was enlarged compared to

that of the WT (843.3 \AA^2), whereas it was comparable to that of the Delta RBD–ACE2 (928.4 \AA^2)³¹ (Fig. 2e). This is also in agreement with our biolayer interferometry data showing that the ACE2-binding affinity of the Omicron S is similar to that of the Delta S but is higher than that of the G614 S (Fig. 2a).

Omicron sensitivity to neutralizing mAbs

We compared five previously generated mAbs—2H2, 3C1, 8D3, S5D2 and S3H3 (refs. 45,46)—for neutralization of the WT (Wuhan-Hu-1 strain), Delta or Omicron pseudoviruses (Fig. 3a,b). The half-maximal inhibitory concentration (IC_{50}) values of the mAbs 3C1, 2H2, 8D3 and S3H3 against Delta were comparable (less than 2.5-fold variation) to the corresponding ones against WT, whereas S5D2 was still neutralizing to Delta ($\text{IC}_{50} = 734.6\text{ ng ml}^{-1}$) but was about 90-fold less potent. In Omicron neutralization tests, 3C1, 8D3 and S5D2 lost neutralization activity ($\text{IC}_{50} > 10\text{ \mu g ml}^{-1}$). However, 2H2 and S3H3 remained highly effective against Omicron with IC_{50} s being 30.4 and 53.3 ng ml^{-1} , respectively, despite observing a 3.3-fold increase (relative to the WT) in the IC_{50} value for 2H2. Thus, 2H2 and S3H3 are two potent neutralizing mAbs against Omicron.

Fig. 3: Neutralization and binding activities of the mAbs against Omicron and Delta variants.



The mAbs were raised to WT RBD or S trimer proteins. **a**, Neutralization IC₅₀ values and fold changes in neutralization potency for the Delta and Omicron variant pseudoviruses (PVs) compared to the WT pseudovirus. A minus sign (−) denotes decrease. The yellow highlighting indicates a more than tenfold decrease; the red highlighting denotes a more than 1,000-fold decrease. **b**, Neutralization of the mAbs towards WT, Delta and Omicron pseudoviruses. All mAbs were fourfold serially diluted. Data are expressed as mean ± s.e.m. of four replicate wells. The horizontal black dotted lines indicate 0% neutralization. **c**, Binding activities of the mAbs to recombinant S trimers of the WT, Delta and Omicron strains were tested by ELISA. Serially diluted S trimer proteins were coated onto the ELISA wells. Data are mean ± s.d. of triplicate wells. Neutralization and ELISA data are representative of two independent experiments with similar results.

Source data

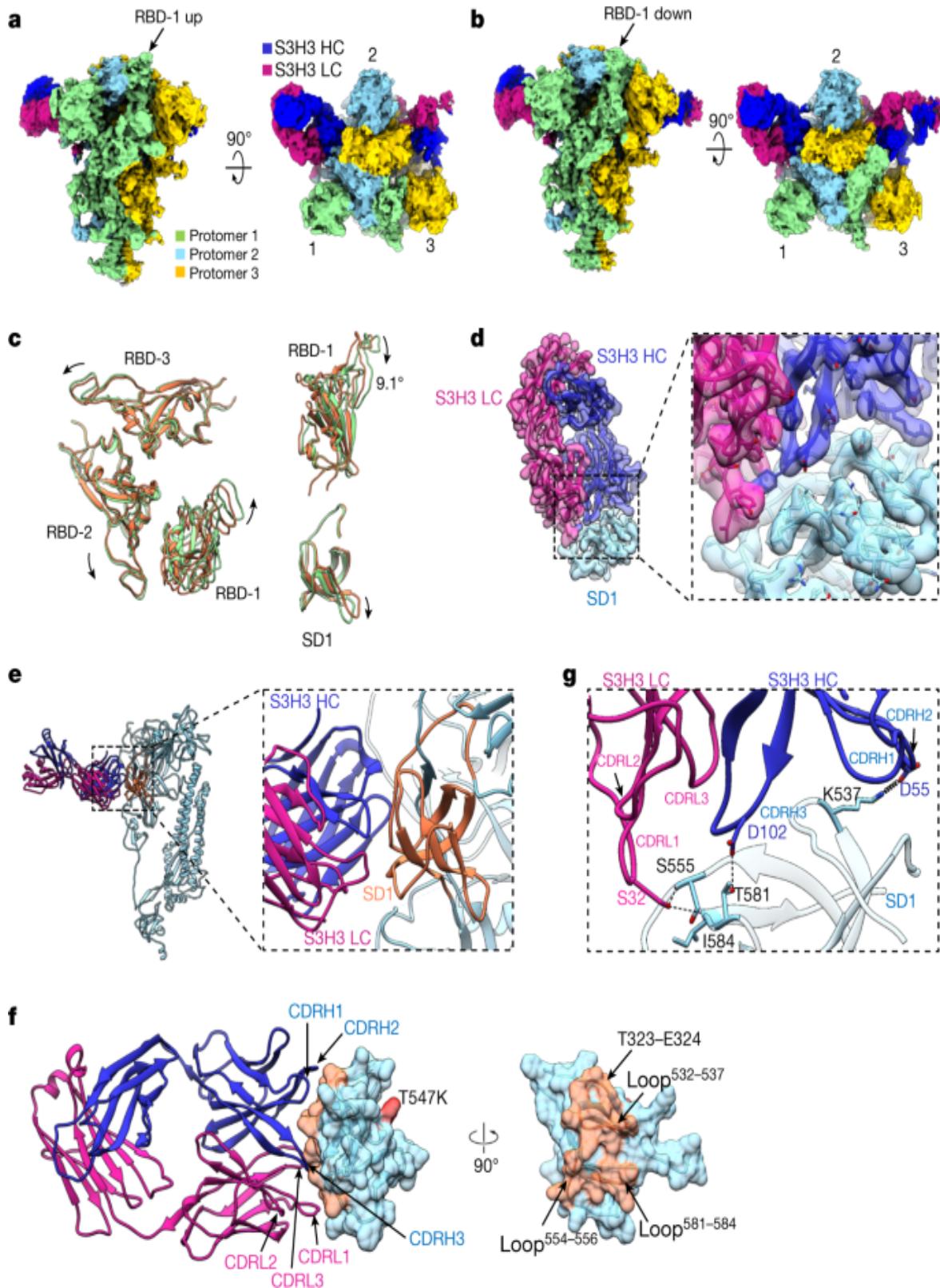
We then compared the binding ability of the five mAbs to the WT, Delta and Omicron S proteins by ELISA (Fig. 3c). For mAb S5D2, its binding to the Delta S and to the Omicron S was nearly abolished; for mAbs 3C1 and 8D3,

their reactivity profile with the Delta S closely resembled that towards the WT S but their binding to the Omicron S was substantially reduced; for mAb 2H2, its binding curve to the Omicron S was similar to those towards the WT S and the Delta S, despite the binding efficiency to the Omicron S being slightly lower; meanwhile, mAb S3H3 produced nearly identical binding curves to the three S proteins. Overall, the antigen-binding ability of the mAbs was in good agreement with their neutralization potency towards specific variant pseudovirus (Fig. 3). Collectively, the above data showed that Omicron remains sensitive to binding and neutralization by the mAbs 2H2 and S3H3, whereas it displays resistance to 3C1, 8D3 and S5D2.

Structure of the Omicron S–S3H3 complex

The mAb S3H3 is a unique neutralizing antibody that binds to the SD1 region of the WT S⁴⁶. Our cryo-EM study revealed two states of the Omicron S in complex with S3H3 Fab. Both structures showed two engaged Fab densities on the SD1 of protomer 2 and protomer 3, but with the RBD-1 in the up (termed Omicron S-open–S3H3) or down (termed Omicron S-close–S3H3) conformations (Fig. 4a, b). The two maps were resolved to the resolution of 3.48 Å and 3.64 Å, respectively (Extended Data Fig. 5, Extended Data Table 1). Compared with the Omicron S-open, the S trimer in S-open–S3H3 exhibits a slight twist and the RBD-1 displays a 9.1° downward rotation (Fig. 4c), making the S trimer seemingly less ‘open’ as a whole. SD1 shows a slight downward rotation (Fig. 4c).

Fig. 4: Cryo-EM analyses on the Omicron S–S3H3 Fab complex.



a, b, Side and top views of the cryo-EM map of the Omicron S-open–S3H3 (**a**) and S-close–S3H3 complex (**b**), with the heavy chain (HC) and light chain (LC) of S3H3 Fab in medium blue and violet red, respectively. This colour scheme is followed throughout this figure. **c**, Conformational comparison between Omicron S-open–S3H3 (light green) and Omicron S-open (orange), indicating a slight twist of the RBDs of S-open–S3H3 and the downward rotations of RBD-1 (up to 9.1°) and SD1. **d**, Model map fitting of the focus-refined Omicron SD1–S3H3 structure, and the zoomed-in view of the Omicron SD1–S3H3 interaction interface. The side chain densities at the interface were well resolved. **e**, The S3H3 binding on SD1 of protomer 2. **f**, The structural elements involved in the interaction between S3H3 Fab and SD1 are labelled. The SD1 T547K substitution is also shown in red. The residues of SD1 in proximity to S3H3 (less than 4 Å) are indicated and coloured in transparent orange (related to Extended Data Table 4). **g**, The SD1–S3H3 interaction interface analysed using PISA, with the major involved structural elements labelled (the spring represents the salt bridge, and the black dashed line represents the hydrogen bond).

We further focus-refined the SD1–S3H3 Fab region and obtained a map at 3.61 Å resolution, with most of the side chain densities well resolved (Fig. 4d). The heavy chain of S3H3 Fab contributes more to the interactions with SD1 than the light chain does, that is, all three heavy-chain complementarity determining regions (CDRs) of S3H3 and its CDRL1 and CDRL3 interact with T323–E324 and the three loops (loop^{532–537}, loop^{554–556} and loop^{581–584}) of SD1 (Fig. 4e,f, Extended Data Table 4). Specifically, S32 of CDRL1 forms hydrogen bonds with S555 and I584 of SD1, respectively, D102 of CDRH3 forms a hydrogen bond with T581 from loop^{581–584}, and D55 of CDRH2 forms a salt bridge with K537 from loop^{532–537} (Fig. 4g, Extended Data Table 4), thus constituting an intense interaction network between S3H3 Fab and SD1. A single mutation, T547K, is present in the SD1 of Omicron; however, this mutation was located outside the footprint of S3H3 (Fig. 4f), and thus will not affect the Omicron S–S3H3 interaction. Collectively, S3H3 binds to the extremely conserved SD1 region, therefore retains binding and neutralizing activity towards major VOCs including Omicron.

Discussion

In this study, we performed cryo-EM and biochemical analysis on the Omicron S trimer and its complex with the ACE2 receptor. We captured both the closed and the open states of the Omicron S trimer (Fig. 1a). In contrast to the S trimer of the Delta, Beta and Kappa variants^{10,31}, the Omicron S-close and S-open structures appear more twisted and compact than their counterpart of the G614 strain¹ (Fig. 1d,e), which may hinder its spike transformation towards the fusion-prone open state and shielding of S1. This could be related to enhanced inter-protomer and S1–S2 interactions induced by unique Omicron substitutions (T547K, N856K and N764K in SD1 and S2) (Fig. 1g).

Corroborating to the enhanced inter-protomer and S1–S2 interactions of Omicron, our cryo-EM analysis revealed that for the Omicron S trimer, the dominantly populated conformation is the closed state with all the RBDs buried, possibly leading to ‘conformational masking’ that may prevent antibody binding and neutralization at sites of receptor binding, similar to that described for the HIV-1 envelope^{47,48}. Such an Omicron conformational masking mechanism could potentially affect antibodies that bind to the up RBDs (such as classes 1, 2 and 4 RBD antibodies⁴⁹), contributing to the observed extensive neutralization escape by Omicron. However, for the Delta S trimer, our recent work showed that the dominant population is in the open state, indicating that the conformational masking mechanism may be less effective for the Delta variant^{31,35}. This could contribute to the striking immune evasion of the Omicron variant^{15,17,18,19,20,21,22}.

We also captured three states for the Omicron S–ACE2 complex (Fig. 2b). Unlike the Delta S, which tends to bind to three ACE2 in majority³¹, Omicron mainly binds to one or two ACE2 (Extended Data Fig. 3b). Further focus-refining of the RBD-1–ACE2 structure demonstrated that the substitutions on the RBM of Omicron (especially Q493R, G496S, Q498R, S477N and Y505H) result in the formation of new salt bridges and hydrogen bonds, and more complementary electrostatic surface properties (Fig. 2d, f,g), which together may compensate the abolished original RBM–ACE2 interactions^{26,43,44}, leading to enhanced interactions with ACE2 and potentially enhanced transmissibility of the Omicron variant.

The present study showed that 2H2 and S3H3 retain potent neutralization towards Omicron and Delta (Fig. 3). Further structural study revealed a unique binding epitope of S3H3 within the SD1 region (Fig. 4). The binding of S3H3 to S trimer may function as a ‘lock’ to block the releasing of S1 from S2, resulting in inhibition of virus entry. The SD1 region targeted by S3H3 is extremely conserved among SARS-CoV-2 variants (Fig. 4f), thus explaining the cross-neutralization ability of S3H3 towards Omicron, Delta and other variants⁴⁶. These findings also suggest a possibility to design SD1-based broad-spectrum SARS-CoV-2 vaccines. It is somewhat surprising that 2H2, whose epitope largely overlaps with RBM⁴⁵, remained highly neutralizing against Omicron, despite the loss of the neutralization potency of this antibody to the Kappa and Beta variants due to their E484Q or E484K substitution¹⁰. Docking of our previous RBD-bound 2H2 Fab structure⁴⁵ onto our current Omicron RBD-1 structure from the S-open state revealed that the 2H2 Fab could be reasonably well accommodated without major clashes with RBD (Extended Data Fig. 6a). In particular, the E484A substitution in Omicron does not appear to create steric hindrance with 2H2 due to the smaller size of the Ala side chain. In addition, the Omicron RBM substitutions, such as Q493R, Q498R and Y505H, do not introduce extra clashes between RBM and 2H2, and also make the epitope surface more positively charged than the WT RBD-1 (ref. 45) (Extended Data Fig. 6b), potentially allowing better interaction with the 2H2 paratope, which tends to be more negatively charged (Extended Data Fig. 6c).

We found that Omicron could escape three RBD antibodies, including 8D3, S5D2 and 3C1 (refs. 45,46). Both 8D3 and S5D2 are class 1 RBD antibodies and they share similar epitopes centred around loop^{477–489} (refs. 31,46). Docking of the RBD-bound 8D3 or S5D2 Fab structures^{31,46} onto the Omicron RBD-1 structure revealed that several Omicron residues, especially S477N, may potentially clash with the 8D3 Fab (Extended Data Fig. 6d); the S477N and T478K substitutions in Omicron may break the hydrogen bond network between RBD and S5D2 Fabs (Extended Data Fig. 6e). 3C1, whose epitope mainly involves the relatively stable RBD core⁴⁵, belongs to the class 3 RBD antibody⁴⁹. Analysis of the docked 3C1 structure on the Omicron RBD-1 showed that the S375F substitution may contribute to the escape of Omicron from 3C1 binding and neutralization through altering the interaction interface (Extended Data Fig. 6f).

In summary, the present study reveals that the Omicron spike acquires an increased RBM–ACE2 interaction network contributed by new hydrogen bonds and salt bridges and more favourable surface properties, and therefore maintains a strong affinity to ACE2, providing a possible explanation to the high infectivity and transmissibility of Omicron. In addition, our study suggests that, besides individual residue substitutions in RBD antibody epitopes, a conformational masking mechanism may also contribute to the extensive antibody evasion by Omicron. Moreover, our work demonstrates that Omicron remains sensitive to S3H3, an antibody that targets the extremely conserved SD1 region. Our findings provide structural insights into how Omicron maintains high transmissibility while greatly evading immunity, and may also inform design of broadly effective vaccines against emerging variants.

Methods

Protein expression and purification

To express the SARS-CoV-2 Omicron variant S glycoprotein ectodomain, the mammalian codon-optimized gene-coding SARS-CoV-2 (hCoV-19 Botswana R42B90_BHP_000842207 2021, GISAID ID: EPI_ISL_6752027) S glycoprotein ectodomain with double proline substitutions and ‘GSAS’ substitution at the furin cleavage site was cloned into the pcDNA 3.1+ vector³⁵. A C-terminal T4 fibritin trimerization motif, a TEV protease cleavage site, a FLAG tag and a His tag were cloned downstream of the S glycoprotein ectodomain (Extended Data Fig. 1a). The constructs of prefusion-stabilized S proteins of SARS-CoV-2 G614 and Delta (B.1.617.2) variants were prepared as previously reported³¹. A gene encoding the human ACE2 peptidase domain (Q18–D615) with an N-terminal IL-10 signal peptide and a C-terminal His tag was cloned into the pcDNA 3.4 vector³⁵. The recombinant proteins were prepared as previously reported in a published protocol³⁵. In brief, the constructs were transiently transfected into HEK293F cells (Thermo Fisher) using polyethylenimine (PEI). Note that the HEK293F cell line has not recently been tested for mycoplasma contamination. Three days after transfection, the supernatants were harvested by centrifugation, and then passed through 0.45-μm filter

membrane. The clarified supernatants were added with 20 mM Tris-HCl pH 7.5, 200 mM NaCl, 20 mM imidazole and 4 mM MgCl₂, and incubated with Ni-NTA resin at 4 °C for 1 h. The Ni-NTA resin was recovered and washed with 20 mM Tris-HCl pH 7.5, 200 mM NaCl and 20 mM imidazole. The protein was eluted by 20 mM Tris-HCl pH 7.5, 200 mM NaCl and 250 mM imidazole.

Biolayer interferometry assay

Before the biolayer interferometry (BLI) assay, Ni-NTA-purified recombinant S trimer proteins of the G614, Delta and Omicron SARS-CoV-2 variants were further purified by gel-filtration chromatography using a Superose 6 increase 10/300 GL column (GE Healthcare) pre-equilibrated with PBS. Then, 70 µg of purified S trimer proteins of the G614, Delta and Omicron variants were separately biotinylated using the EZ-Link Sulfo-NHS-LC-LC-Biotin kit (Thermo Fisher) and then purified by Zeba spin desalting columns (Thermo Fisher).

Binding affinities of S trimers to ACE2 were determined by BLI analysis on an Octet Red96 instrument (Pall FortéBio). In brief, biotinylated S trimer proteins (approximately 40 µg ml⁻¹) were immobilized onto streptavidin (SA) biosensors (Pall FortéBio). After washing with kinetic buffer (0.01 M PBS with 0.02% Tween 20 and 0.1% bovine serum albumin), these sensors were incubated with threefold serial dilutions of ACE2 monomer protein for 500 s. Subsequently, the biosensors were allowed to dissociate in kinetic buffer for 500 s. The data were analysed using the Octet Data Analysis 11.0 software to calculate affinity constants.

Neutralization

Luciferase (Luc)-expressing pseudoviruses bearing SARS-CoV-2 S proteins were constructed based on the HIV-1 backbone. In brief, HEK 293T cells (American Type Culture Collection) in a 10-cm dish were co-transfected using PEI (polysciences) with 10 µg pCMV-dR8.91 packaging plasmid, 10 µg recombinant pLVX-IRES-ZsGreen1 plasmid containing the luciferase reporter gene, and 2 µg recombinant pVAX1 plasmids encoding the SARS-CoV-2 S proteins. The cells were incubated with the transfection mixture for

6 h, and then 5 ml of fresh DMEM medium with 10% FBS was added to each dish. After incubation overnight, the media in the dishes were replaced with fresh DMEM medium (10% FBS). At 48 h post-transfection, the culture supernatant was harvested and frozen at -80°C before use.

All mAbs were fourfold serially diluted and tested by a pseudovirus neutralization assay with human ACE2-overexpressing HEK 293T cells (293T-hACE2), which were generated in our previous study and verified by western blot, following our previous protocol⁴⁵. Note that the 293T-hACE2 cell line has not recently been tested for mycoplasma contamination. Two days after pseudovirus infection, luciferase activity was measured. Data were analysed by non-linear regression using GraphPad Prism 8 to calculate the IC₅₀.

ELISA

To test the binding activities of recombinant Omicron S protein with our previously developed anti-SARS-CoV-2 mAbs^{45,46}, recombinant S trimer proteins from WT⁴⁵, Delta or Omicron SARS-CoV-2 strains were twofold serially diluted and coated onto ELISA plates at 37°C for 2 h. The plates were blocked with 5% milk in PBS-Tween 20 (PBST) at 37°C for 1 h. After washing with PBST, the plates were incubated with 50 ng per well of each of the anti-SARS-CoV-2 mAbs^{45,46} at 37°C for 2 h, followed by horseradish peroxidase (HRP)-conjugated anti-mouse IgG (1:5,000 dilution; Sigma) at 37°C for 1 h. After washing and colour development, absorbance was measured at 450 nm. ELISA data were analysed by non-linear regression using GraphPad Prism 8. Note that for neutralization and ELISA assays, no statistical method was used to predetermine sample size, and no blinding or randomization protocols were used.

Omicron S–S3H3 Fab complex formation

The Omicron S–S3H3 Fab complex was prepared following our previously reported protocol⁴⁶. In brief, purified S3H3 IgG was incubated with papain (300:1 w/w) in PBS buffer (in the presence of 20 mM l-cysteine and 1 mM EDTA) for 3 h at 37°C . The reaction was quenched by 20 mM iodoacetamide. Fab was purified by running over a HiTrap DEAE FF

column (GE Healthcare) pre-equilibrated with PBS. Omicron S protein was incubated with S3H3 Fab in a 1:6 molar ratio on ice for 1 h. The Omicron S–S3H3 Fab complex was purified by size-exclusion chromatography using a Superose 6 increase 10/300 GL column (GE Healthcare) in 20 mM Tris-HCl pH 7.5, 200 mM NaCl and 4% glycerol. The complex peak fractions were concentrated and assessed by SDS-PAGE.

Cryo-EM

Sample preparation

To prepare the cryo-EM sample of the Omicron S trimer, a 2.2 μ l aliquot of the sample was applied on a plasma-cleaned holey carbon grid (R 1.2/1.3, Cu, 200 mesh; Quantifoil). The grid was blotted with Vitrobot Mark IV (Thermo Fisher Scientific) at 100% humidity and 8 °C, and then plunged into liquid ethane cooled by liquid nitrogen. To prepare the cryo-EM sample of the Omicron S–ACE2 complex, the purified Omicron S trimer was incubated with ACE2 in a 1:4 molar ratio on ice for 20 min and then vitrified using the same condition. The purified Omicron S–S3H3 complex was vitrified using the same procedure as for the Omicron S sample.

Data collection

Cryo-EM movies of the samples were collected on a Titan Krios electron microscope (Thermo Fisher Scientific) operated at an accelerating voltage of 300 kV. For the three datasets, the movies were collected at a magnification of \times 64,000 and recorded on a K3 direct electron detector (Gatan) operated in the counting mode (yielding a pixel size of 1.093 Å), and under a low-dose condition in an automatic manner using EPU 2.11 software (Thermo Fisher Scientific). Each frame was exposed for 0.1 s, and the total exposure time was 3 s, leading to a total accumulated dose of 50.2 e⁻/Å² on the specimen.

3D reconstruction

For each dataset, the motion correction of image stack was performed using the embedded module Motioncor2 in Relion 3.1 (refs. [35,50,51](#)), and CTF

parameters were determined using CTFFIND4.1.8 (ref. [52](#)) before further data processing. Unless otherwise described, the data processing was performed in Relion 3.1.

For the Omicron S dataset (Extended Data Fig. [2](#)), 600,845 particles remained after reference-free 2D classification in cryoSPARC v3.3.1 (ref. [30](#)). After two rounds of 3D classification and further focused 3D classification on the RBD-1 region, we obtained an Omicron S-close map from 108,509 particles and an RBD-1-up open conformation from 69,873 particles. We then preformed focused 3D classification on the RBD-3 region of the open-state dataset and obtained an Omicron S-open map from 30,967 particles and a S-open-2 map from 38,906 particles. After Bayesian polishing and CTF refinement, the Omicron S-close, S-open and S-open-2 datasets were independently loaded into cryoSPARC v3.3.1 (ref. [30](#)) and refined to the resolution of 3.08 Å, 3.40 Å and 3.41 Å, respectively, using non-uniform refinement. The overall resolutions for all of the cryo-EM maps in this study were determined based on the gold-standard criterion using a Fourier shell correlation (FSC) of 0.143. Moreover, we performed 3D variability analysis on the Omicron S trimer dataset containing 178,382 particles in cryoSPARC v3.3.1 to capture its continuous conformational dynamics^{[30](#)}.

For the Omicron S–ACE2 dataset (Extended Data Fig. [3](#)), 1,268,072 particles remained after reference-free 2D classification. After two rounds of 3D classification and further focused 3D classification on the RBD-1–ACE2 region, we obtained an Omicron S–ACE2 map from 141,538 particles. After Bayesian polishing and CTF refinement, the map was reconstructed to 3.53 Å resolution. We then focused on the RBD-2 region for further classification and obtained two conformations, with RBD-2 in the ‘down’ position (43.9%), termed S–ACE2-C1, or in the ‘up’ position. The RBD-2 up dataset was further focused 3D classified on the RBD-3 region. We then obtained two conformations with RBD-3 in the down or up position, termed S–ACE2-C2 and S–ACE2-C3, respectively. The three datasets were independently loaded into cryoSPARC v3.3.1 and refined using non-uniform refinement to 3.69 Å, 3.70 Å and 4.04 Å resolution, respectively. Here, after obtaining the 3.53 Å resolution map of Omicron S–ACE2, we performed further local refinement on the RBD-1–ACE2 region (indicated by a black

dotted ellipsoid in Extended Data Fig. 3a) in cryoSPARC v3.3.1 to acquire a 3.67 Å resolution map of this region.

For the Omicron S–S3H3 dataset (Extended Data Fig. 5), a similar data processing procedure was adapted as described for the Omicron S dataset to obtain a 3.56 Å resolution S–S3H3 map from 238,121 particles. We then carried out focused 3D classification on the RBD-1 region, followed by non-uniform refinement in cryoSPARC v3.3.1, and obtained a 3.48 Å resolution S-open–S3H3 map from 162,221 particles and a 3.64 Å resolution S-close–S3H3 map from 75,900 particles. In addition, after obtaining the 3.56 Å resolution map, we performed focused 3D classification on the S3H3–SD1 region of protomer 2 (highlighted by an orange dotted ellipsoid in Extended Data Fig. 5), leading to a dataset of 101,192 particles, which was further local refined on the S3H3–SD1 region in cryoSPARC v3.3.1, deducing a 3.61 Å resolution map of this region. All of the obtained maps were post-processed through deepEMhancer⁵³.

Atomic model building

To build an atomic model for the Omicron S-open structure, we used the atomic model of Delta S-open (PDB: [7W92](#)) from our previous study as the initial model³¹. We first fit the model into our Omicron S-open map in Chimera by rigid body fitting, then manually substituted the mutations of the Omicron variant in COOT⁵⁴. Subsequently, we flexibly refined the model against our Omicron S-open map using ROSETTA⁵⁵. Finally, we used the phenix.real_space_refine module in Phenix 1.19.2-4158 for the S trimer model refinement against the map⁵⁶. For the S-close model, we utilized the down protomer from our recent Delta S-transition (PDB: [7W94](#))³¹ structure as the initial template, and followed a similar procedure described above for model refinement. For the Omicron S–ACE2 and the local refined RBD-1–ACE2 structures, we used the Delta S–ACE2 model (PDB: [7W98](#) and [7W9I](#))³¹ as the initial template, and followed a similar procedure described above for model refinement. For better fitting in the dynamic ACE2 region of the Omicron S–ACE2-C1, C2, and C3 maps, we merged the better resolved Omicron RBD-1–ACE2 model with the other portion of the original model to make a complete model, and refined it against the corresponding unsharpened map. For the Omicron S–S3H3 and the local

refined RBD-1–S3H3 structures, we utilized our recent Beta S–S3H3 model (PDB: [7WDF](#) and [7WD8](#))⁴⁶ as the template, and followed a similar procedure described above for model refinement. The atomic models were validated using the Phenix.molprobity command in Phenix. Analyses of the interaction interface were conducted through the PISA server⁵⁷.

UCSF Chimera and ChimeraX were applied for figure generation, rotation measurement and Coulombic potential surface analysis^{58,59}.

Reporting summary

Further information on research design is available in the [Nature Research Reporting Summary](#) linked to this paper.

Data availability

The cryo-EM maps determined for the Omicron S trimer have been deposited at the Electron Microscopy Data Bank with the accession codes [EMD-32556](#), [EMD-32854](#) and [EMD-32855](#), and the associated atomic models have been deposited in the PDB with the accession codes [7WK2](#), [7WVN](#) and [7WVO](#) for S-close, S-open and S-open-2 states, respectively. For the S-ACE2 dataset, related cryo-EM maps have been deposited in the Electron Microscopy Data Bank with the accession codes [EMD-32558](#), [EMD-32856](#), [EMD-32857](#) and [EMD-32560](#), and the associated models have been deposited in the PDB with the accession codes [7WK4](#), [7WVP](#), [7WVQ](#) and [7WK6](#) for S-ACE2-C1, S-ACE2-C2, S-ACE-C3 and RBD-1–ACE2, respectively. For the S–S3H3 Fab dataset, related cryo-EM maps have been deposited in the Electron Microscopy Data Bank with the accession codes [EMD-32562](#), [EMD-32563](#) and [EMD-32564](#), and the associated models have been deposited in the PDB with the accession codes [7WK8](#), [7WK9](#) and [7WKA](#) for SD1–S3H3, S-open–S3H3 and S-close–S3H3, respectively. The structures used for the initial templates or structural analysis in this work include PDB IDs: [7W92](#), [7W94](#), [7W98](#), [7W9I](#), [7WDF](#), [7WD8](#), [7KRQ](#), [7KRR](#), [6M0J](#), [7DK4](#), [7WCR](#), [7W9F](#) and [7DCC](#). For gel source images, see Supplementary Fig. 1. [Source data](#) are provided with this paper.

References

1. Zhang, J. et al. Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science* **372**, 525–530 (2021).
2. Grabowski, F., Preibisch, G., Gizinski, S., Kochanczyk, M. & Lipniacki, T. SARS-CoV-2 variant of concern 202012/01 has about twofold replicative advantage and acquires concerning mutations. *Viruses* **13**, 392 (2021).
3. Wise, J. Covid-19: new coronavirus variant is identified in UK. *Brit. Med. J.* **371**, m4857 (2020).
4. Davies, N. G. et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, eabg3055 (2021).
5. Gobeil, S. M. et al. Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity. *Science* **373**, eabi6226 (2021).
6. Cai, Y. et al. Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. *Science* **373**, 642–648 (2021).
7. Yuan, M. et al. Structural and functional ramifications of antigenic drift in recent SARS-CoV-2 variants. *Science* **373**, 818–823 (2021).
8. Tegally, H. et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
9. Msomi, N., Mlisana, K. & de Oliveira, T., Network for Genomic Surveillance in South Africa writing group. A genomics network established to respond rapidly to public health threats in South Africa. *Lancet Microbe* **1**, e229–e230 (2020).
10. Wang, Y. et al. Conformational dynamics of the Beta and Kappa SARS-CoV-2 spike proteins and their complexes with ACE2 receptor revealed by cryo-EM. *Nat. Commun.* **12**, 7345 (2021).

11. Voloch, C. M. et al. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *J. Virol.* **95**, e00119–21 (2021).
12. Singh, J., Rahman, S. A., Ehtesham, N. Z., Hira, S. & Hasnain, S. E. SARS-CoV-2 variants of concern are emerging in India. *Nat. Med.* **27**, 1131–1133 (2021).
13. Winger, A. & Caspari, T. The spike of concern—the novel variants of SARS-CoV-2. *Viruses* **13**, 1002 (2021).
14. WHO. Enhancing readiness for Omicron (B.1.1.529): technical brief and priority actions for Member States. *WHO* [https://www.who.int/publications/m/item/enhancing-readiness-for-omicron-\(b.1.1.529\)-technical-brief-and-priority-actions-for-member-states](https://www.who.int/publications/m/item/enhancing-readiness-for-omicron-(b.1.1.529)-technical-brief-and-priority-actions-for-member-states) (2021).
15. Cameroni, E. et al. Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron antigenic shift. *Nature* **602**, 664–670 (2021).
16. Mannar, D. et al. SARS-CoV-2 Omicron variant: antibody evasion and cryo-EM structure of spike protein-ACE2 complex. *Science* **375**, 760–764 (2022).
17. Hoffmann, M. et al. The Omicron variant is highly resistant against antibody-mediated neutralization—implications for control of the COVID-19 pandemic. *Cell* **185**, 447–456.e11 (2021).
18. Carreño, J. M. et al. Activity of convalescent and vaccine serum against SARS-CoV-2 Omicron. *Nature* **602**, 682–688 (2021).
19. Liu, L. et al. Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602**, 676–681 (2021).
20. Cao, Y. et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657–663 (2022).
21. Planas, D. et al. Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. *Nature* **602**, 671–675 (2022).

22. Cele, S. et al. Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. *Nature* **602**, 654–656 (2022).
23. Tang, T., Bidon, M., Jaimes, J. A., Whittaker, G. R. & Daniel, S. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Res.* **178**, 104792 (2020).
24. Rabaan, A. A. et al. SARS-CoV-2, SARS-CoV, and MERS-CoV: a comparative overview. *Infez. Med.* **28**, 174–184 (2020).
25. Wang, Q. et al. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* **181**, 894–904.e9 (2020).
26. Lan, J. et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
27. Shang, J. et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
28. Yin, W. et al. Structures of the Omicron spike trimer with ACE2 and an anti-Omicron antibody. *Science* **375**, 1048–1053 (2022).
29. Toelzer, C. et al. Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein. *Science* **370**, 725–730 (2020).
30. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
31. Wang, Y. et al. Structural basis for SARS-CoV-2 Delta variant recognition of ACE2 receptor and broadly neutralizing antibodies. *Nat. Commun.* **13**, 871 (2022).
32. Cai, Y. et al. Distinct conformational states of SARS-CoV-2 spike protein. *Science* **369**, 1586–1592 (2020).
33. Walls, A. C. et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281–292.e6 (2020).

34. Wrapp, D. et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
35. Xu, C. et al. Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM. *Sci. Adv.* **7**, eabe5575 (2021).
36. Henderson, R. et al. Controlling the SARS-CoV-2 spike glycoprotein conformation. *Nat. Struct. Mol. Biol.* **27**, 925–933 (2020).
37. Benton, D. J. et al. Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature* **588**, 327–330 (2020).
38. Berger, I. & Schaffitzel, C. The SARS-CoV-2 spike protein: balancing stability and infectivity. *Cell Res.* **30**, 1059–1060 (2020).
39. Cui, Z. et al. Structural and functional characterizations of infectivity and immune evasion of SARS-CoV-2 Omicron. *Cell* **185**, 860–871.e13 (2022).
40. Han, P. et al. Receptor binding and complex structures of human ACE2 to spike RBD from Omicron and Delta SARS-CoV-2. *Cell* **185**, 630–640 (2022).
41. Lan, J. et al. Structural and computational insights into the SARS-CoV-2 Omicron RBD-ACE2 interaction. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.01.03.474855> (2022).
42. McCallum, M. et al. Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement. *Science* **375**, 864–868 (2022).
43. Mannar, D. et al. Structural analysis of receptor binding domain mutations in SARS-CoV-2 variants of concern that modulate ACE2 and antibody binding. *Cell Rep.* **37**, 110156 (2021).
44. Laffeber, C., de Koning, K., Kanaar, R. & Lebbink, J. H. G. Experimental evidence for enhanced receptor binding by rapidly spreading SARS-CoV-2 variants. *J. Mol. Biol.* **433**, 167058 (2021).

45. Zhang, C. et al. Development and structural basis of a two-mAb cocktail for treating SARS-CoV-2 infections. *Nat. Commun.* **12**, 264 (2021).
46. Xu, S. et al. Mapping cross-variant neutralizing sites on the SARS-CoV-2 spike protein. *Emerg. Microbes Infect.* **11**, 351–367 (2021).
47. Kwong, P. D. et al. HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature* **420**, 678–682 (2002).
48. Munro, J. B. et al. Conformational dynamics of single HIV-1 envelope trimers on the surface of native virions. *Science* **346**, 759–763 (2014).
49. Barnes, C. O. et al. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **588**, 682–687 (2020).
50. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
51. Fernandez-Leiro, R. & Scheres, S. H. W. A pipeline approach to single-particle processing in RELION. *Acta Crystallogr. D Struct. Biol.* **73**, 496–502 (2017).
52. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
53. Sanchez-Garcia, R. et al. DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *Commun. Biol.* **4**, 874 (2021).
54. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
55. DiMaio, F. et al. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods* **12**, 361–365 (2015).

56. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
57. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
58. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
59. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).

Acknowledgements

We are grateful to the staff at the NCPSS Electron Microscopy facility, Database and Computing facility, and Protein Expression and Purification facility for instrument support and technical assistance. This work was supported by grants from the Strategic Priority Research Program of CAS (XDB37040103 and XDB29040300), National Key R&D Program of China (2017YFA0503503 and 2020YFC0845900), the NSFC (32130056 and 31872714), Shanghai Academic Research Leader (20XD1404200), Shanghai Pilot Program for Basic Research—CAS, Shanghai Branch (JCYJ--SHFY-2022-008) and the European Union’s Horizon 2020 research and innovation programme under grant agreement (no. 101003589). C.Z. is supported by the Youth Innovation Promotion Association of CAS and Shanghai Rising-Star Program (21QA1410000). C.X. is supported by the China National Postdoctoral Program for Innovative Talents (BX2021310), Shanghai Super Postdoctoral Incentive Plan.

Peer review information

Nature thanks Nadav Elad and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Author information

Author notes

1. These authors contributed equally: Qin Hong, Wenyu Han, Jiawei Li, Shiqi Xu, Yifan Wang

Affiliations

1. State Key Laboratory of Molecular Biology, National Center for Protein Science Shanghai, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai, China

Qin Hong, Wenyu Han, Jiawei Li, Yifan Wang, Cong Xu, Zuyang Li, Yanxing Wang & Yao Cong

2. University of Chinese Academy of Sciences, Beijing, China

Qin Hong, Wenyu Han, Jiawei Li, Yifan Wang, Zuyang Li & Yao Cong

3. CAS Key Laboratory of Molecular Virology and Immunology, Institut Pasteur of Shanghai, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai, China

Shiqi Xu, Chao Zhang & Zhong Huang

Contributions

Y.C. and Z.H. designed the experiments. Yanxing Wang expressed and purified the proteins with the assistance of Z.L. and S.X. Q.H. and W.H. performed cryo-EM data acquisitions. Q.H., W.H., J.L. and Yifan Wang performed cryo-EM reconstructions and model buildings. C.Z. and S.X. performed biochemical analyses. J.L., Q.H., W.H., Yifan Wang, C.X. and C.Z. analysed the data. Y.C., Z.H., Q.H., J.L., W.H., Yanxing Wang and C.Z. wrote the manuscript.

Corresponding authors

Correspondence to [Yanxing Wang](#), [Chao Zhang](#), [Zhong Huang](#) or [Yao Cong](#).

Ethics declarations

Competing interests

Institut Pasteur of Shanghai is in the process of applying for a patent application (no. 202110713023.7) covering anti-SARS-CoV-2 mAb S3H3 that lists Z.H., S.X. and C.Z. as inventors. The other authors declare no competing interests.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

[Extended Data Fig. 1 Purification of Omicron variant S and S-S3H3 Fab complex.](#)

a, Schematic diagram of the Omicron variant S organization in this study (left, positions of all mutations are indicated), and the model of a SARS-CoV-2 S protomer (right) with mutation sites of the Omicron variant shown as red sphere. **b**, SDS-PAGE analysis of the purified Omicron variant S protein. Representative images of two independent experiments are shown. For gel source data, related to Supplementary Fig. 1. **c**, Size-exclusion chromatogram and SDS-PAGE analysis of the formed Omicron S-S3H3 Fab complex. Representative images of two independent experiments are shown. For gel source data, related to Supplementary Fig. 1.

[Extended Data Fig. 2 Cryo-EM analysis on the Omicron S trimer.](#)

a, Data processing workflow for structure determination of the Omicron S trimer. A representative original micrograph, reference-free 2D class averages, and the population distributions of the Omicron S-close, S-open, and S-open-2 states are also presented. **b**, Resolution assessment of the Omicron S-close, S-open, and S-open-2 maps by FSC at 0.143 criterion. **c**, **d**, Angular distribution (**c**) and local resolution evaluation (**d**) of the S-close, S-open, and S-open-2 maps. **e**, Model-map fitting for each of the three structures. **f**, Model and map fitting in the inter-protomer interaction area of Omicron S-close. Protomer 1, 2, and 3 are colored in light green, royal blue and gold, respectively.

[Extended Data Fig. 3 Cryo-EM data processing procedure for the Omicron S-ACE2 complex.](#)

a, Data processing procedure for the Omicron S-ACE2 dataset. A representative original micrograph and the reference-free 2D class averages are also presented. The masks for the focused 3D classification on RBD-2 and RBD-3 regions are shown as pink and orange transparent surfaces, respectively. **b**, Population distribution of the Omicron S-ACE2-C1/-C2/-C3 states. **c**, Lower threshold rendering of the Omicron S-ACE2-C1/-C2/-C3 maps. ACE2 density (violet red) also associates with RBD-2 or RBD-3 in the C2 and C3 states, although appearing weaker than that of the stably associated ACE2 on RBD-1, indicating a dynamic ACE2 association with RBD-2–3.

[Extended Data Fig. 4 Cryo-EM analysis on the Omicron S-ACE2 complex.](#)

a, Resolution assessment of the cryo-EM maps by FSC at 0.143 criterion. **b**, Angular distribution of the Omicron S-ACE2-C1/-C2/-C3 and RBD-1-ACE2 maps. **c**, **d**, Local resolution evaluation (**c**) and model-map fitting (**d**) for the Omicron S-ACE2 maps and the RBD-1-ACE2 map.

[Extended Data Fig. 5 Cryo-EM analysis on the Omicron S-S3H3 Fab complex.](#)

a, Data processing workflow for the Omicron S-S3H3 Fab complex. A representative original micrograph and the reference-free 2D class averages are also presented. **b**, Resolution assessment of the cryo-EM maps by FSC at 0.143 criterion. **c**, Local resolution evaluation of the Omicron S-S3H3 and SD1-S3H3 maps. **d**, Angular distribution for the three Omicron S-S3H3 complexes. **e**, Model-map fitting for the Omicron S-S3H3 complexes.

Extended Data Fig. 6 Structural basis of 2H2 antibody retaining neutralization potency and 8D3/S5D2/3C1 antibodies losing their neutralization potency against Omicron.

a, Model of the Omicron RBD-2H2 Fab interaction interface, obtained by docking of the RBD-bound 2H2 (PDB ID: [7DK4](#)) onto the RBD-1 structure from our current Omicron S-open. **b**, The electrostatic surface properties of Omicron and WT RBDs, with the black line depicting the footprint of 2H2 on RBD-1. The residues with substantial electrical charge changes are indicated. **c**, The electrostatic surface property of 2H2 Fab, with the black line depicting the footprint of RBD-1 on 2H2. **d-f**, Docking of RBD-bound 8D3/S5D2/3C1 (PDB ID: [7W9F](#)/[7WCR](#)/[7DCC](#), respectively) onto the Omicron RBD-1 structure. Residue S477N that may potentially clash with the 8D3 Fab is labeled (**d**). Residues S477N and T478K that may break the H-bond network between RBD and S5D2 Fab are labeled (**e**). Residue S375F that may contribute to the Omicron's escape of 3C1 binding and neutralization is labeled (**f**). All substituted residues in Omicron RBD-1 are colored in red.

Extended Data Table 1 Cryo-EM data collection and refinement statistics for Omicron S, Omicron S-ACE2, and Omicron S-S3H3
Extended Data Table 2 Omicron S-close protomer 1/2 and protomer 1/3 interactions

Extended Data Table 3 Interaction interface analysis (upper table) and footprint distance measurement (lower table) for Omicron RBD-1-ACE2 structure

Extended Data Table 4 Interaction interface analysis (upper table) and footprint distance measurement (lower table) for Omicron SD1-S3H3 structure

Supplementary information

[Supplementary Figure 1](#)

Uncropped gel images for Extended Data Fig. 1. **a**, Uncropped gel image for Extended Data Fig. 1b. **b**, Uncropped gel image for Extended Data Fig. 1c. The cropped section used for the Extended Data figures are indicated by the red rectangles.

[Reporting Summary](#)

[Peer Review File](#)

[Supplementary Video 1](#)

A representative motion mode revealed by 3DVA of the Omicron S trimer dataset.

Source data

[Source Data Fig. 2](#)

[Source Data Fig. 3](#)

Rights and permissions

[Reprints and Permissions](#)

About this article

Cite this article

Hong, Q., Han, W., Li, J. *et al.* Molecular basis of receptor binding and antibody neutralization of Omicron. *Nature* **604**, 546–552 (2022).

<https://doi.org/10.1038/s41586-022-04581-9>

- Received: 11 January 2022
- Accepted: 24 February 2022
- Published: 28 February 2022
- Issue Date: 21 April 2022
- DOI: <https://doi.org/10.1038/s41586-022-04581-9>

Share this article

Anyone you share the following link with will be able to read this content:

[Get shareable link](#)

Sorry, a shareable link is not currently available for this article.

[Copy to clipboard](#)

Provided by the Springer Nature SharedIt content-sharing initiative

This article was downloaded by **calibre** from <https://www.nature.com/articles/s41586-022-04581-9>

| [Section menu](#) | [Main menu](#) |

[Download PDF](#)

- Article
- Open Access
- [Published: 03 March 2022](#)

Antibody evasion properties of SARS-CoV-2 Omicron sublineages

- [Sho Iketani](#) ORCID: [orcid.org/0000-0002-3733-9556^{1,2 na1}](https://orcid.org/0000-0002-3733-9556),
- [Lihong Liu^{1 na1}](#),
- [Yicheng Guo](#) ORCID: [orcid.org/0000-0001-5357-660X^{1 na1}](https://orcid.org/0000-0001-5357-660X),
- [Liyuan Liu^{3 na1}](#),
- [Jasper F.-W. Chan](#) ORCID: [orcid.org/0000-0001-6336-6657^{4,5 na1}](https://orcid.org/0000-0001-6336-6657),
- [Yiming Huang³](#),
- [Maple Wang¹](#),
- [Yang Luo](#) ORCID: [orcid.org/0000-0003-3277-8792¹](https://orcid.org/0000-0003-3277-8792),
- [Jian Yu¹](#),
- [Hin Chu](#) ORCID: [orcid.org/0000-0003-2855-9837^{4,5}](https://orcid.org/0000-0003-2855-9837),
- [Kenn K.-H. Chik^{4,5}](#),
- [Terrence T.-T. Yuen⁴](#),
- [Michael T. Yin^{1,6}](#),
- [Magdalena E. Sobieszczyk^{1,6}](#),
- [Yaoxing Huang](#) ORCID: [orcid.org/0000-0001-6270-1644¹](https://orcid.org/0000-0001-6270-1644),
- [Kwok-Yung Yuen](#) ORCID: [orcid.org/0000-0002-2083-1552^{4,5}](https://orcid.org/0000-0002-2083-1552),
- [Harris H. Wang](#) ORCID: [orcid.org/0000-0003-2164-4318^{3,7}](https://orcid.org/0000-0003-2164-4318),
- [Zizhang Sheng¹](#) &
- [David D. Ho](#) ORCID: [orcid.org/0000-0003-1627-149X^{1,2,6}](https://orcid.org/0000-0003-1627-149X)

[Nature](#) volume **604**, pages 553–556 (2022)

- 40k Accesses

- 8 Citations
- 883 Altmetric
- [Metrics details](#)

Subjects

- [Antiviral agents](#)
- [SARS-CoV-2](#)
- [Vaccines](#)
- [Viral immune evasion](#)

Abstract

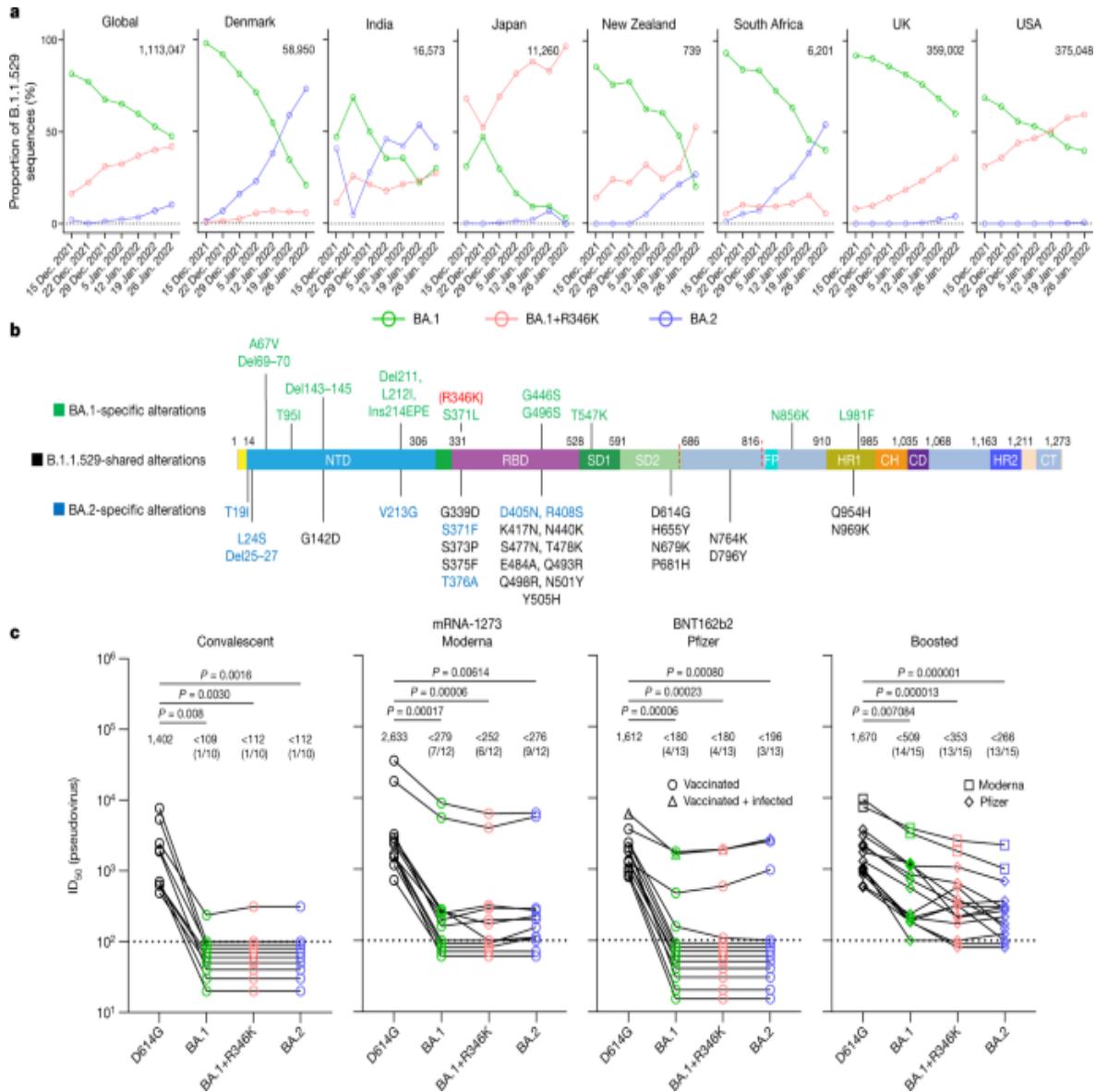
The identification of the Omicron (B.1.1.529.1 or BA.1) variant of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Botswana in November 2021¹ immediately caused concern owing to the number of alterations in the spike glycoprotein that could lead to antibody evasion. We² and others^{3,4,5,6} recently reported results confirming such a concern. Continuing surveillance of the evolution of Omicron has since revealed the rise in prevalence of two sublineages, BA.1 with an R346K alteration (BA.1+R346K, also known as BA.1.1) and B.1.1.529.2 (BA.2), with the latter containing 8 unique spike alterations and lacking 13 spike alterations found in BA.1. Here we extended our studies to include antigenic characterization of these new sublineages. Polyclonal sera from patients infected by wild-type SARS-CoV-2 or recipients of current mRNA vaccines showed a substantial loss in neutralizing activity against both BA.1+R346K and BA.2, with drops comparable to that already reported for BA.1 (refs. 2,3,5,6). These findings indicate that these three sublineages of Omicron are antigenically equidistant from the wild-type SARS-CoV-2 and thus similarly threaten the efficacies of current vaccines. BA.2 also exhibited marked resistance to 17 of 19 neutralizing monoclonal antibodies tested, including S309 (sotrovimab)⁷, which had retained appreciable activity against BA.1 and BA.1+R346K (refs. 2,3,4,6). This finding shows that no authorized monoclonal antibody therapy could adequately cover all sublineages of the

Omicron variant, except for the recently authorized LY-CoV1404 (bebtelovimab).

Main

The rise of the Omicron (B.1.1.529) variant to become the dominant variant of SARS-CoV-2 globally has been remarkable⁸. Continuing surveillance of its evolution in the population in December 2021 and January 2022 has revealed that the proportion of the original form, BA.1, has been decreasing steadily whereas the proportions of two other sublineages have increased noticeably (Fig. 1a). In fact, the BA.1+R346K sublineage now accounts for about 40% of Omicron sequences globally, and about 35–60% in New Zealand, the UK and the USA. On the other hand, the BA.2 sublineage accounts for only about 10% of Omicron sequences globally, but it is not only on the rise but also the dominant form in countries such as Denmark, India and South Africa. These three sublineages of Omicron share 21 alterations in the spike protein, wherein BA.2 contains 8 unique alterations and BA.1 contains 13 unique alterations (Fig. 1b). Of course, BA.1+R346K has one alteration more than BA.1. Given these differences, their antigenic properties cannot be assumed to be the same or similar.

Fig. 1: BA.2 exhibits a similar serum neutralization profile to those of BA.1 sublineages.



a, Proportions of BA.1, BA.1+R346K and BA.2 in B.1.1.529 sequences on GISAID over the latter half of December 2021 and January 2022. The value in the upper right corner of each box denotes the cumulative number of Omicron sequences. **b**, Alterations in the B.1.1.529 lineage. NTD, N-terminal domain; RBD, receptor-binding domain; SD1, subdomain 1; SD2, subdomain 2; FP, fusion peptide; HR1, heptad repeat 1; CH, central helix; CD, connector domain; HR2, heptad repeat 2; CT, cytoplasmic tail. **c**, Pseudovirus neutralization by convalescent and vaccinee sera. $n = 10, 12, 13$ and 15 biologically independent serum samples, respectively, for convalescent, mRNA-1273, BNT162b2 and boosted groups. The values above the points indicate geometric means. The numbers in parentheses

denote the numbers of samples above the limit of detection (LOD) of 100. Values below the LOD are arbitrarily plotted to allow for visualization of each sample. *P* values were determined by a two-sided Friedman test followed by Dunn's multiple comparisons test.

Serum neutralization of sublineages

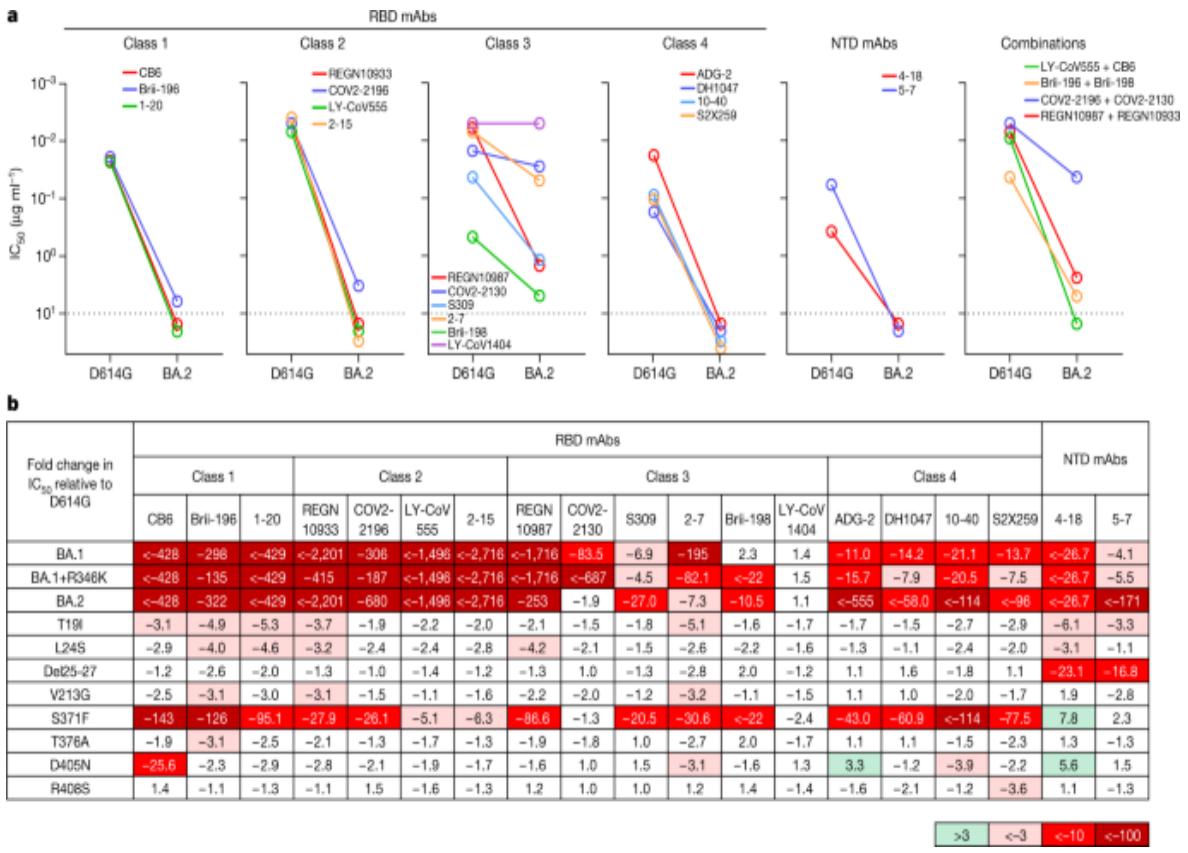
Therefore, we first investigated the sensitivity of the Omicron sublineages to neutralization by polyclonal sera from convalescent individuals or individuals given mRNA vaccines, with or without a booster shot. These serum samples, as well as the pseudovirus neutralization assay used, were identical to ones previously reported². The wild-type D614G pseudovirus was included as a comparator. As was observed and reported for BA.1 (refs. 2,3,5,6), a marked and significant loss of neutralizing activity of the serum against BA.1+R346K and BA.2 relative to D614G was noted, with neutralizing titres for numerous samples dropping below the limit of detection (Fig. 1c). The loss of neutralizing activity against BA.1+R346K or BA.2 sublineages was less prominent for sera obtained from individuals who received a booster vaccination (Fig. 1c, right panel), consistent with reported findings for BA.1 (refs. 2,3,6). Among these samples, the mean serum neutralizing titres against Omicron sublineages were significantly lower than the mean titre for D614G; although the mean titre was slightly lower for BA.2, the difference from that of the BA.1 sublineages did not reach statistical significance ($P = 0.242$). Finally, we confirmed the pseudovirus neutralization data by testing a separate set of sera from individuals given mRNA vaccines for neutralization of authentic viruses (Extended Data Fig. 1 and Extended Data Table 1). As above, neutralizing titres dropped significantly against authentic BA.2 virus relative to D614G.

Antibody neutralization of sublineages

To further examine antigenic differences in the spike protein of these Omicron sublineages, a panel of 19 neutralizing monoclonal antibodies was used as probes. Seventeen were directed to different epitope clusters (classes 1–4) in the receptor-binding domain (RBD), whereas two were directed to the N-terminal domain (NTD). These antibodies included REGN10987

(imdevimab)⁹, REGN10933 (casirivimab)⁹, COV2-2196 (tixagevimab)¹⁰, COV2-2130 (cilgavimab)¹⁰, LY-CoV555 (bamlanivimab)¹¹, CB6 (etesevimab)¹², Brii-196 (amubarvimab)¹³, Brii-198 (romlusevimab)¹³, S309 (sotrovimab)⁷, LY-CoV1404 (bebtelovimab)¹⁴, ADG-2 (ref. ¹⁵), DH1047 (ref. ¹⁶) and S2X259 (ref. ¹⁷), as well as 1-20, 2-15, 2-7, 4-18, 5-7 (ref. ¹⁸) and 10-40 (ref. ¹⁹) from our group. Overall, 17 of the 19 monoclonal antibodies were either totally inactive or severely impaired in neutralizing BA.2 (Fig. ^{2a}), similar to previous findings for BA.1 and BA.1+R346K (ref. ²) but with important differences (Fig. ^{2b}). All class 4 antibodies tested lost greater neutralizing potency against BA.2 versus BA.1 sublineages. Two class 3 antibodies, COV2-2130 and 2-7, retained decent activity against BA.2 but had almost no activity against BA.1 viruses. S309 or sotrovimab lost 27-fold neutralizing activity against BA.2; this is important because it is an authorized monoclonal antibody that was found to retain activity against the original form of Omicron^{2,3,4}. LY-CoV1404, the most recently authorized monoclonal antibody, remained potent in neutralizing all Omicron sublineages, suggesting that there is still a patch in this antibody-binding region that is unaffected by all spike alterations found in SARS-CoV-2 variants so far. Although there was a lack of an observable difference among the Omicron sublineages in neutralization by polyclonal sera (Fig. ^{1c}), important antigenic differences do exist when probed by monoclonal antibodies. BA.1 seems to be more resistant to class 3 antibodies than BA.2 (except for S309), whereas BA.2 is more resistant to all class 4 antibodies tested. Our recent study² showed that previous SARS-CoV-2 variants, such as Beta (B.1.351) and Delta (B.1.617.2), evolved to resist class 1, class 2 and NTD antibodies first, and then the Omicron variant seemingly has further evolved to resist class 3 and class 4 antibodies in addition. Our current findings suggest that the Omicron sublineages may have diverged under slightly different pressure from class 3 and class 4 antibodies to the RBD.

Fig. 2: BA.2 differs in resistance profile to monoclonal antibodies.



a, Pseudovirus neutralization by monoclonal antibodies (mAbs). Values above the LOD of $10 \mu\text{g ml}^{-1}$ (dotted line) are arbitrarily plotted to allow for visualization of each sample. **b**, Fold change in IC₅₀ values relative to D614G of neutralization of Omicron variants, as well as point mutants unique to BA.2.

Alterations conferring antibody resistance

Finally, we constructed each of the eight BA.2-specific spike alterations alone as pseudoviruses and tested them using the same panel of 19 monoclonal antibodies (Fig. 2b). S371F broadly affected most of the RBD-directed antibodies, similar to what was observed for S371L in BA.1 (ref. 2) but with a greater negative impact, perhaps due to the bulkier side chain of phenylalanine. Notably, S371F seems to be responsible for the loss in potency of S309, although this alteration was not observed previously as a marker for clinical resistance to sotrovimab²⁰. CB6 was adversely affected by the D405N alteration, probably owing to its position in the epitope of this

antibody¹². It is not clear how T19I and L24S alterations in the NTD subtly impaired the neutralizing activity of class 1 antibodies to the RBD.

Discussion

In summary, we have comprehensively evaluated the antigenic properties of two sublineages of the Omicron variant, BA.1+R346K and BA.2, and we believe that our results have important clinical implications. First, polyclonal sera showed a substantial loss in neutralizing activity against both sublineages, with drops comparable to that against BA.1 (Fig. [1c](#)). These three sublineages of Omicron, therefore, seem to be antigenically equidistant from the wild-type SARS-CoV-2, probably threatening the efficacies of current coronavirus disease 2019 (COVID-19) vaccines to a similar extent. The present study, however, does not address the antigenic distance between BA.1 and BA.2, the determination of which will require cross-neutralization experiments using sublineage-specific sera. Second, monoclonal antibodies were affected in a disparate manner for the different Omicron sublineages. For clinically approved or authorized antibodies, S309 (sotrovimab) retained activity against both BA.1 and BA.1+R346K, but its activity against BA.2 has dropped 27-fold (Fig. [2b](#)) to a 50% inhibitory concentration (IC_{50}) of about $1 \mu\text{g ml}^{-1}$ (Fig. [2a](#)). COV2-2130 (cilgavimab) and its combination with COV2-2196 (tixagevimab) retained activity against BA.2, but this antibody combination is authorized only for preventive use. Only the recently authorized LY-CoV1404 (bebtelovimab) could adequately treat all sublineages of the Omicron variant. As COVID-19 treatment options are narrowed by the emergence of more and more variants, it is imperative that we continue to devise novel strategies to contain this ever-evolving pathogen.

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Serum samples

For the pseudovirus neutralization experiments, identical samples from a previous study were utilized². For the authentic virus neutralization experiments, the samples are described in Extended Data Table 1. All collections were conducted under protocols reviewed and approved by the Institutional Review Board of Columbia University. All of the participants provided written informed consent.

Antibodies

Antibodies were expressed as previously described¹⁸. Briefly, Vh and Vl genes for each antibody were codon optimized and synthesized (GenScript), and then inserted into mammalian expression vectors. These plasmids were transiently transfected into Expi293 cells (Thermo Fisher) using polyethylenimine and cultured for 5 days, and then the antibody was purified by affinity chromatography using rProtein A Sepharose (GE). REGN10933, REGN10987, COV2-2130 and COV2-2196 were provided by Regeneron Pharmaceuticals, Brii-196 and Brii-198 were provided by Brii Biosciences, and CB6 was provided by B. Zhang and P. Kwong (NIAID).

Cells

Expi293 cells were obtained from Thermo Fisher (catalogue number A14527), Vero E6 cells were obtained from ATCC (catalogue number CRL-1586), HEK293T cells were obtained from ATCC (catalogue number CRL-3216), and Vero-E6-TMPRSS2 cells were obtained from JCRB (catalogue number JCRB1819). All cells were purchased from authenticated vendors and morphology was visually confirmed before use. All cell lines tested mycoplasma negative.

Pseudovirus production

Spike expression constructs for variant SARS-CoV-2 spikes were produced by an in-house gene synthesis method as previously described². Constructs were confirmed by sequencing, and then transfected into HEK293T cells using Lipofectamine 3000 (Thermo Fisher) according to the manufacturer's

instructions. Cells were washed 24 h post-transfection with complete medium (DMEM + 10% FBS + penicillin/streptomycin) and then infected with rVSV-G-pseudotyped ΔG-luciferase (G*ΔG-luciferase, Kerafast). Cells were thoroughly washed 2 h post-infection with complete medium, and then incubated for a further 24 h at 37 °C under 5% CO₂. Pseudoviruses were then collected and incubated with anti-VSV-G hybridoma supernatant for 1 h at 37 °C (I1-Hybridoma, ATCC) to neutralize residual rVSV-G. The titre of each pseudovirus was determined by serially diluting the virus in complete medium in 96-well plates, and then incubating with 40,000 Vero E6 cells for approximately 12 h at 37 °C under 5% CO₂. Following infection, luminescence was quantified using the Luciferase Assay System (Promega) according to the manufacturer's instructions and measured with a SpectraMax i3x Multi-Mode Microplate Reader (Molecular Devices) using SoftMax Pro 7.0.2 (Molecular Devices), and then the titre was determined by comparison to control wells with cells alone. Pseudoviruses were aliquoted and stored at –80 °C until use.

Pseudovirus neutralization assay

Neutralization assays were conducted in 96-well plates by serially diluting sera or antibodies in complete medium, starting at 1:100 dilution or 10 µg ml⁻¹, respectively, and incubating with pseudoviruses for 1 h at 37 °C. Following incubation, 40,000 Vero E6 cells were added to each well, and further incubated for approximately 12 h at 37 °C under 5% CO₂.

Luminescence was quantified using the Luciferase Assay System according to the manufacturer's instructions and measured with a SpectraMax i3x Multi-Mode Microplate Reader using SoftMax Pro 7.0.2. Neutralization was determined by comparison to control wells with cells alone and with virus alone. IC₅₀ values were calculated by fitting a nonlinear five-parameter dose–response curve in GraphPad Prism version 9.2.

Authentic virus isolation and propagation

SARS-CoV-2 variants D614G (GISAID: EPI_ISL_497840) and BA.2 (GISAID: EPI_ISL_9845731) were isolated from respiratory tract specimens of patients with COVID-19 in Hong Kong by J.F.-W.C., K.-Y.Y.

and colleagues at the Department of Microbiology, The University of Hong Kong. The viruses were propagated in Vero-E6-TMPRSS2 cells and the sequence was confirmed by next-generation sequencing before use.

Authentic virus neutralization assay

Vero-E6-TMPRSS2 cells were seeded in 96-well plates in complete medium overnight at 37 °C under 5% CO₂ to establish a monolayer. The following day, sera were serially diluted starting at 1:500 dilution in 96-well plates in triplicate in DMEM + 2% FBS and then incubated with 0.01 MOI of either virus at 37 °C for 1 h. Afterwards, the mixture was overlaid onto cells and further incubated at 37 °C under 5% CO₂ for approximately 72 h. Cytopathic effects were then visually assessed in all wells and scored as either negative or positive for infection by comparison to control uninfected or infected wells in a blinded manner. Neutralization curves and IC₅₀ values were derived by fitting a nonlinear five-parameter dose–response curve to the data in GraphPad Prism version 9.2.

Reporting summary

Further information on research design is available in the [Nature Research Reporting Summary](#) linked to this paper.

Data availability

All experimental data are provided in the manuscript. Omicron prevalence analyses utilized sequences submitted to and available from GISAID (ref. ⁸). The sequences of the authentic viruses used in this study have been deposited to GISAID (<https://www.gisaid.org/>) under the accession numbers EPI_ISL_497840 (D614G) and EPI_ISL_9845731 (BA.2). Materials use in this study will be made available under an appropriate Materials Transfer Agreement.

References

1. Viana, R. et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **https://doi.org/10.1038/s41586-022-04411-y** (2022).
2. Liu, L. et al. Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602**, 676–681 (2022).
3. Cameroni, E. et al. Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron antigenic shift. *Nature* **602**, 664–670 (2022).
4. Cao, Y. et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657–663 (2022).
5. Cele, S. et al. Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. *Nature* **602**, 654–656 (2022).
6. Planas, D. et al. Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. *Nature* **602**, 671–675 (2022).
7. Pinto, D. et al. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature* **583**, 290–295 (2020).
8. Shu, Y. & McCauley, J. GISAID: global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
9. Hansen, J. et al. Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science* **369**, 1010–1014 (2020).
10. Zost, S. J. et al. Potently neutralizing and protective human antibodies against SARS-CoV-2. *Nature* **584**, 443–449 (2020).
11. Jones, B. E. et al. The neutralizing antibody, LY-CoV555, protects against SARS-CoV-2 infection in nonhuman primates. *Sci. Transl. Med.* **13**, eabf1906 (2021).
12. Shi, R. et al. A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature* **584**, 120–124 (2020).

13. Ju, B. et al. Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* **584**, 115–119 (2020).
14. Westendorf, K. et al. LY-CoV1404 (bebtelovimab) potently neutralizes SARS-CoV-2 variants. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.04.30.442182> (2022).
15. Rappazzo, C. G. et al. Broad and potent activity against SARS-like viruses by an engineered human monoclonal antibody. *Science* **371**, 823–829 (2021).
16. Li, D. et al. In vitro and in vivo functions of SARS-CoV-2 infection-enhancing and neutralizing antibodies. *Cell* **184**, 4203–4219 (2021).
17. Tortorici, M. A. et al. Broad sarbecovirus neutralization by a human monoclonal antibody. *Nature* **597**, 103–108 (2021).
18. Liu, L. et al. Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **584**, 450–456 (2020).
19. Liu, L. et al. Isolation and comparative analysis of antibodies that broadly neutralize sarbecoviruses. *Sci. Transl. Med.* (in the press).
20. Rockett, R. et al. Resistance mutations in SARS-CoV-2 Delta variant after sotrovimab use. *N. Engl. J. Med.* <https://doi.org/10.1056/NEJMc2120219> (2022).

Acknowledgements

We thank Regeneron Pharmaceuticals and B. Zhang and P. Kwong (NIAID) for antibodies. This study was supported by funding from the Gates Foundation, the JPB Foundation, A. and P. Cherng, S. Yin, C. Ludwig, D. and R. Wu, Health@InnoHK, the National Science Foundation (MCB-2032259) and the NIH SARS-CoV-2 Assessment of Viral Evolution Program.

Author information

Author notes

1. These authors contributed equally: Sho Iketani, Lihong Liu, Yicheng Guo, Liyuan Liu, Jasper F.-W. Chan

Affiliations

1. Aaron Diamond AIDS Research Center, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

Sho Iketani, Lihong Liu, Yicheng Guo, Maple Wang, Yang Luo, Jian Yu, Michael T. Yin, Magdalena E. Sobieszczyk, Yaoxing Huang, Zizhang Sheng & David D. Ho

2. Department of Microbiology and Immunology, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

Sho Iketani & David D. Ho

3. Department of Systems Biology, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

Liyuan Liu, Yiming Huang & Harris H. Wang

4. State Key Laboratory of Emerging Infectious Diseases, Carol Yu Centre for Infection, Department of Microbiology, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong Special Administrative Region, China

Jasper F.-W. Chan, Hin Chu, Kenn K.-H. Chik, Terrence T.-T. Yuen & Kwok-Yung Yuen

5. Centre for Virology, Vaccinology and Therapeutics, Hong Kong Science and Technology Park, Hong Kong Special Administrative Region, China

Jasper F.-W. Chan, Hin Chu, Kenn K.-H. Chik & Kwok-Yung Yuen

6. Division of Infectious Diseases, Department of Medicine, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

Michael T. Yin, Magdalena E. Sobieszczuk & David D. Ho

7. Department of Pathology and Cell Biology, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

Harris H. Wang

Contributions

D.D.H. conceived this project. S.I. and Lihong Liu conducted pseudovirus neutralization experiments. Y.G. and Z.S. conducted bioinformatic analyses. Liyuan Liu and Yiming Huang constructed the spike expression plasmids. J.F.-W.C., H.C., K.K.-H.C., and T.T.-T.Y. conducted authentic virus neutralization experiments. M.W. aided sample collections. Y.L. managed the project. J.Y. expressed and purified antibodies. M.T.Y. and M.E.S. provided clinical samples. Yaoxing Huang contributed to discussions. K.-Y.Y., H.H.W., and D.D.H. directed and supervised the project. S.I., Lihong Liu and D.D.H. analysed the results and wrote the manuscript.

Corresponding author

Correspondence to [David D. Ho](#).

Ethics declarations

Competing interests

S.I., Lihong Liu, J.Y., Yaoxing Huang and D.D.H. are inventors on patent applications (WO2021236998) or provisional patent applications (63/271,627) filed by Columbia University for a number of SARS-CoV-2 neutralizing antibodies described in this manuscript. Both sets of applications are under review. D.D.H. is a co-founder of TaiMed Biologics

and RenBio, consultant to WuXi Biologics and Brii Biosciences, and board director for Vicarious Surgical.

Peer review

Peer review information

Nature thanks the anonymous reviewers for their contribution to the peer review of this work.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

[Extended Data Fig. 1 Serum neutralization of authentic viruses.](#)

Authentic virus neutralization by vaccinee sera. $n = 10$ and 13 biologically independent serum samples, respectively, for two doses mRNA and three doses mRNA groups. Values above points indicate the geometric mean. Numbers in parentheses denote the number of samples above the limit of detection (LOD) of 500 . Values below the LOD and those that overlap are plotted to allow for visualization of each sample. P values were determined by two-sided Wilcoxon matched-pairs signed rank test.

Extended Data Table 1 Demographics and vaccination information for serum samples from vaccinated individuals used in authentic virus neutralization experiments

Supplementary information

[Reporting Summary](#)

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Reprints and Permissions

About this article

Cite this article

Iketani, S., Liu, L., Guo, Y. *et al.* Antibody evasion properties of SARS-CoV-2 Omicron sublineages. *Nature* **604**, 553–556 (2022).
<https://doi.org/10.1038/s41586-022-04594-4>

- Received: 06 February 2022
- Accepted: 28 February 2022
- Published: 03 March 2022
- Issue Date: 21 April 2022
- DOI: <https://doi.org/10.1038/s41586-022-04594-4>

Share this article

Anyone you share the following link with will be able to read this content:

[Get shareable link](#)

Sorry, a shareable link is not currently available for this article.

[Copy to clipboard](#)

Provided by the Springer Nature SharedIt content-sharing initiative

Further reading

- [**Remdesivir in Coronavirus Disease 2019 patients treated with anti-CD20 monoclonal antibodies: a case series**](#)
 - Susanne Rüfenacht
 - Pascal Gantenbein
 - Werner C. Albrich

Infection (2022)

- [**Antikörpertherapie bei Patienten mit COVID-19**](#)
 - M. Augustin
 - M. Hallek
 - S. Nitschmann

Der Internist (2022)

- [**Serum neutralization of SARS-CoV-2 Omicron sublineages BA.1 and BA.2 in patients receiving monoclonal antibodies**](#)
 - Timothée Bruel
 - Jérôme Hadjadj
 - Olivier Schwartz

Nature Medicine (2022)

This article was downloaded by **calibre** from <https://www.nature.com/articles/s41586-022-04594-4>

| [Section menu](#) | [Main menu](#) |

[Download PDF](#)

- Article
- Open Access
- [Published: 13 April 2022](#)

Nonlinear control of transcription through enhancer–promoter interactions

- [Jessica Zuin](#)¹ na1,
- [Gregory Roth](#) [ORCID: orcid.org/0000-0003-0677-3293](#)¹ na1,
- [Yinxiu Zhan](#) [ORCID: orcid.org/0000-0001-7702-6207](#)¹,
- [Julie Cramard](#) [ORCID: orcid.org/0000-0003-1041-8913](#)¹,
- [Josef Redolfi](#)¹,
- [Ewa Piskadlo](#)¹,
- [Pia Mach](#)^{1,2},
- [Mariya Kryzhanovska](#)¹,
- [Gergely Tihanyi](#) [ORCID: orcid.org/0000-0003-3729-6709](#)^{1,2},
- [Hubertus Kohler](#)¹,
- [Mathias Eder](#)³,
- [Christ Leemans](#)³,
- [Bas van Steensel](#) [ORCID: orcid.org/0000-0002-0284-0404](#)³,
- [Peter Meister](#) [ORCID: orcid.org/0000-0002-6230-4216](#)⁴,
- [Sebastien Smallwood](#)¹ &
- [Luca Giorgetti](#) [ORCID: orcid.org/0000-0002-9664-9087](#)¹

[Nature](#) volume 604, pages 571–577 (2022)

- 18k Accesses
- 207 Altmetric

- [Metrics details](#)

Subjects

- [Chromatin](#)
- [Transcriptional regulatory elements](#)

Abstract

Chromosome structure in mammals is thought to regulate transcription by modulating three-dimensional interactions between enhancers and promoters, notably through CTCF-mediated loops and topologically associating domains (TADs)^{1,2,3,4}. However, how chromosome interactions are actually translated into transcriptional outputs remains unclear. Here, to address this question, we use an assay to position an enhancer at large numbers of densely spaced chromosomal locations relative to a fixed promoter, and measure promoter output and interactions within a genomic region with minimal regulatory and structural complexity. A quantitative analysis of hundreds of cell lines reveals that the transcriptional effect of an enhancer depends on its contact probabilities with the promoter through a nonlinear relationship. Mathematical modelling suggests that nonlinearity might arise from transient enhancer–promoter interactions being translated into slower promoter bursting dynamics in individual cells, therefore uncoupling the temporal dynamics of interactions from those of transcription. This uncovers a potential mechanism of how distal enhancers act from large genomic distances, and of how topologically associating domain boundaries block distal enhancers. Finally, we show that enhancer strength also determines absolute transcription levels as well as the sensitivity of a promoter to CTCF-mediated transcriptional insulation. Our measurements establish general principles for the context-dependent role of chromosome structure in long-range transcriptional regulation.

Main

Transcriptional control in mammals critically depends on enhancers, which control tissue specificity and developmental timing of many genes⁵.

Enhancers are often located hundreds of kilobases away from target promoters and are thought to control gene expression by interacting with the promoters in the three-dimensional space of the nucleus. Chromosome conformation capture (3C) methods⁶ revealed that enhancer–promoter interactions predominantly occur within sub-megabase domains known as topologically associating domains (TADs). These mainly arise from nested looping interactions between sites that are bound by the DNA-binding protein CTCF that act as barriers for the loop extrusion activity of cohesin⁷.

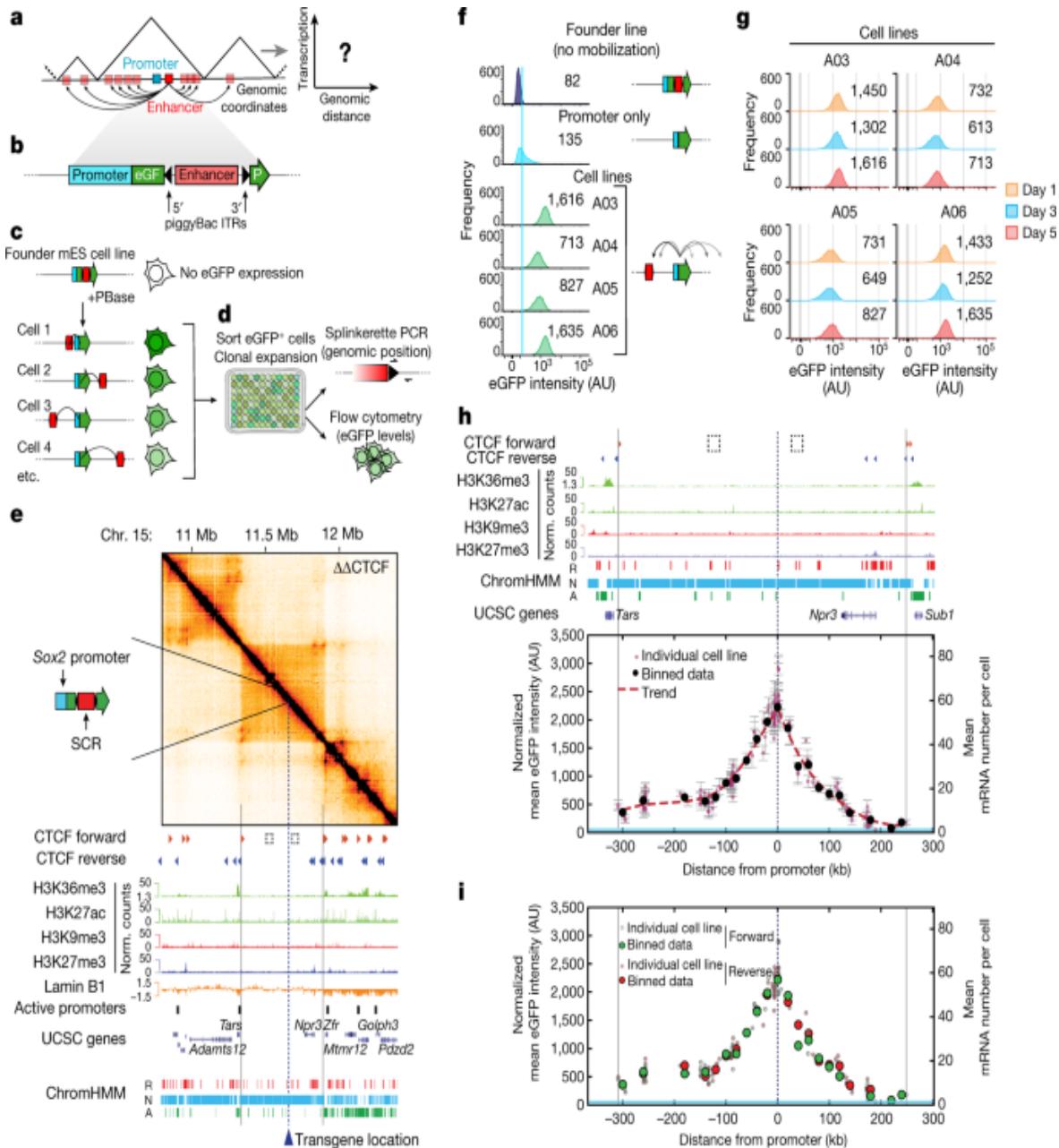
TAD boundaries and CTCF loops are thought to favour enhancer–promoter communication within specific genomic regions and disfavour it with respect to surrounding sequences^{1,3,4,8}. However, this view has recently been challenged by reports that disruption of TAD boundaries^{9,10} or depletion of CTCF and cohesin^{11,12} do not lead to systematic changes in gene expression, and that some regulatory sequences can act across TAD boundaries¹³. The manipulation of single CTCF sites has also been reported to result in variable effects on gene expression^{2,4,10,14,15,16,17,18}. The very notion that physical proximity is required for transcriptional regulation has been questioned by the observed lack of correlation between transcription and proximity in single cells^{19,20}. Thus, it is highly debated whether there are indeed general principles that determine how physical interactions enable or prevent enhancer action²¹. Enhancer–promoter genomic distance might also contribute to transcriptional regulation^{22,23}, but it is unclear whether an enhancer acts uniformly within a TAD^{24,25}, or whether its effect depends on the genomic distance from a promoter^{23,26}.

Enhancer action depends on genomic distance

Addressing these questions requires a quantitative understanding of the relationship between transcription and enhancer–promoter interactions in conditions in which confounding effects by additional regulatory and structural interactions are minimized. Here we provide such a description using an experimental assay in which an enhancer is mobilized from an initial location and reinserted at large numbers of genomic positions with respect to a promoter. This enables the measurement of transcription levels as a function of the enhancer location and, therefore, of enhancer–promoter

contact frequencies (Fig. 1a). Specifically, we generated mouse embryonic stem (mES) cells carrying a transgene in which a promoter drives the expression of enhanced green fluorescent protein (eGFP). The *eGFP* transcript is split in two by a piggyBac transposon containing the cognate enhancer of the promoter (Fig. 1b). After expression of the PBase transposase, the transposon is excised and reintegrated randomly into the genome, but preferentially in the vicinity of the initial site²⁷. Excision leads to reconstitution of functional eGFP of which the expression is used to isolate clonal cell lines by sorting single eGFP⁺ cells (Fig. 1c, d). This enables the rapid generation of hundreds of cell lines, each with the enhancer in a distinct genomic position. Enhancer position and eGFP expression are then determined in every cell line (Fig. 1d).

Fig. 1: Enhancer action depends on the genomic distance from the promoter and is constrained by TAD boundaries.



a, Mobilization of an enhancer around its target promoter to measure transcription as a function of their genomic distance. **b**, Schematic of the transgene: a promoter drives transcription of an *eGFP* gene split by a piggyBac-enhancer cassette. ITR, inverted terminal repeats. **c**, After expression of PBase, the piggyBac-enhancer cassette is excised and randomly reinserted, occasionally leading to eGFP expression. **d**, Sorting of single *eGFP*⁺ cells results in cell lines in which the enhancer drives transcription from a single position. Splinkerette PCR and flow cytometry analysis are used to determine the enhancer position and promoter

expression levels. **e**, Capture-C (6.4 kb resolution) analysis and genomic datasets in mES cells across 2.6 Mb centred around the selected TAD with both of the internal CTCF motifs deleted (dashed squares; $\Delta\Delta$ CTCF, double CTCF site deletion). The dashed line indicates the position of the future insertion of the transgene carrying the *Sox2* promoter and SCR. A, active; N, neutral; R, repressive; Chr, chromosome. **f**, Representative flow cytometry profiles from founder mES cells, a promoter-only control cell line and eGFP⁺ cell lines with mobilized SCR. The light blue line indicates the mean eGFP levels in the promoter-only line. The numbers show the median eGFP intensities. AU, arbitrary units. **g**, eGFP levels in individual eGFP⁺ cell lines over cell passages. The numbers show the median eGFP values. **h**, Normalized mean eGFP intensities in individual eGFP⁺ cell lines as a function of SCR genomic position. The red dots are data from 135 individual cell lines; data are mean \pm s.d. $n = 3$ measurements on different days. The black dots show the average values within equally spaced 20 kb bins. The dashed red line shows the spline interpolation of average values. Mean mRNA numbers were inferred using smRNA-FISH calibration (Extended Data Fig. [1h](#)). The light blue area shows the interval between the mean \pm s.d. of eGFP levels in three promoter-only cell lines. **i**, Data as in **h**, colour-coded according to SCR genomic orientation.

To minimize confounding effects, we integrated the transgene within a 560 kb TAD on chromosome 15 carrying minimal regulatory and structural complexity. This TAD does not contain expressed genes or active enhancers, is mostly composed of ‘neutral’ chromatin^{[28](#)} except for a repressive ~80 kb region at its 3' side (Extended Data Fig. [1a](#)), and displays minimal structure mediated by two internal forward CTCF sites (Extended Data Fig. [1a,b](#)). To further decrease the structural complexity, we deleted the two internal CTCF sites. This led to the loss of the associated loops (Extended Data Fig. [1c](#)) and resulted in a simple homogeneous internal structure, as revealed by capture-C with tiled oligonucleotides spanning 2.9 Mb around the transgene (Fig. [1e](#) and Extended Data Fig. [1c](#)).

We first heterozygotously inserted a single copy (Extended Data Fig. [1e](#)) of a version of the transgene carrying the mouse *Sox2* promoter and the essential 4.8 kb region of its distal enhancer known as *Sox2* control region (SCR)^{[29,30](#)} (Extended Data Fig. [1d](#) and Methods), from which we deleted its single

CTCF site, which is not essential for transcriptional regulation at the endogenous locus¹⁷. Transgene insertion did not lead to substantial structural rearrangements within the TAD besides new moderate interactions with the CTCF sites at the 3' and 5' end of the TAD (Extended Data Fig. 1f). Mobilization of the piggyBac-SCR cassette led to random genomic reinsertions with a preference for chromosome 15 itself (Extended Data Fig. 1g). Individual experiments resulted in several tens of cell lines of which the eGFP levels were unimodally distributed (Fig. 1f), generally higher than those detected in control lines in which transcription was driven by the *Sox2* promoter alone (Fig. 1f), and remained stable over cell passages (Fig. 1g). Mean eGFP levels in single cell lines were linearly correlated with average numbers of *eGFP* mRNAs measured using single-molecule RNA fluorescence *in situ* hybridization (smRNA-FISH) (Extended Data Fig. 1h). We therefore used flow cytometry as a readout of transcriptional activity.

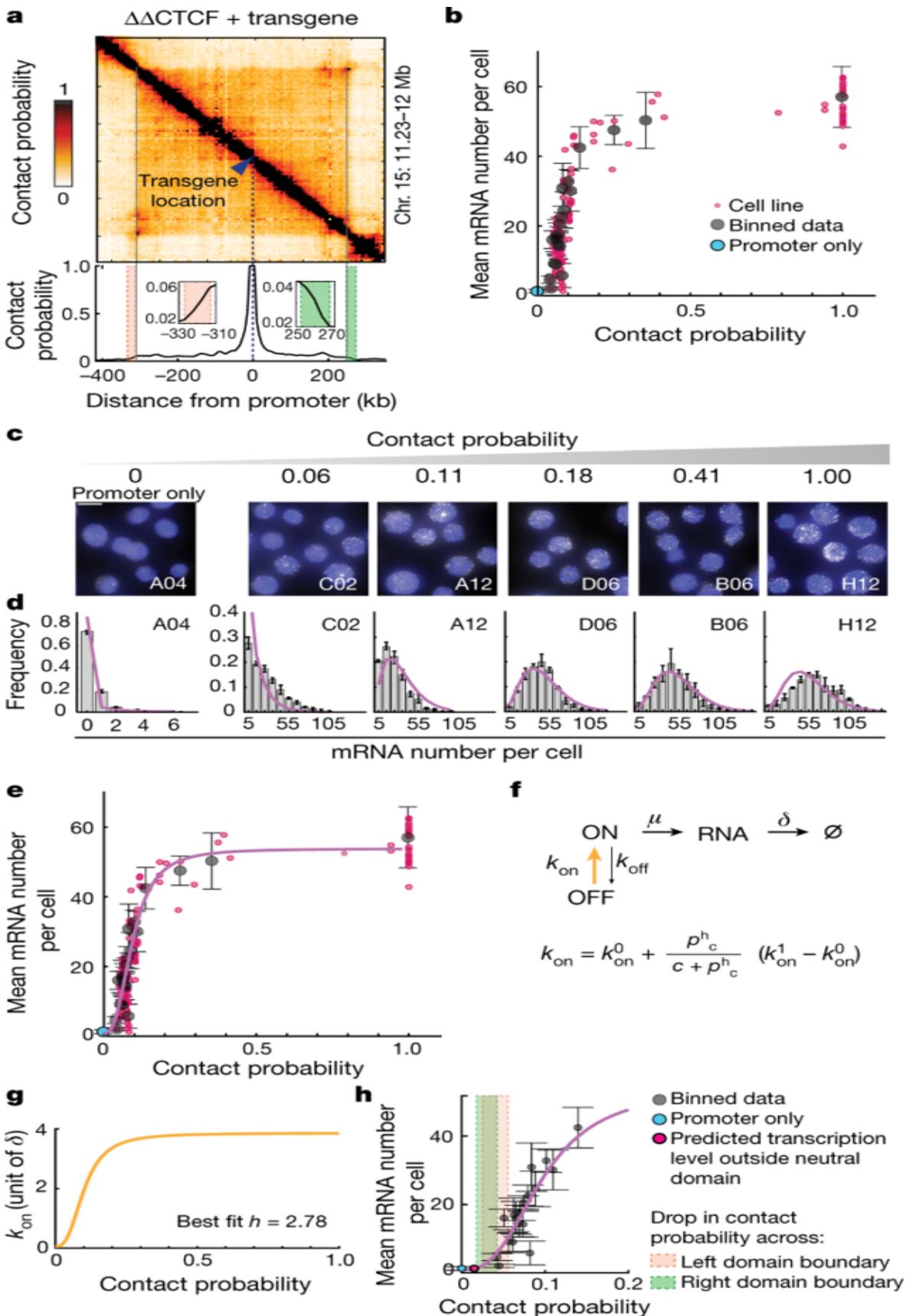
Mapping of piggyBac-SCR positions in more than 300 cell lines revealed that, although in around 15% of them the transposon had not been successfully mobilized, in 99% of those in which it had (262 out of 264), the enhancer reinserted within the initial TAD (Fig. 1h and Extended Data Fig. 1i). In the two cell lines in which the enhancer transposed outside the TAD, eGFP levels were comparable to basal transcription driven by promoter-only control cell lines (Extended Data Fig. 1j). Notably, within the TAD, expression levels decreased with increasing enhancer–promoter genomic distance (Fig. 1h). Genomic distance accounted for a tenfold dynamic range in gene expression, from around 5 to 60 mRNAs per cell on average on the basis of smRNA-FISH calibration (Extended Data Fig. 1h). Insertions downstream of the non-transcribed *Npr3* gene generated lower transcription levels (Fig. 1h), possibly because this is a predominantly repressive region. Mild positive and negative deviations from the average decay in transcription levels indeed correlated with local enrichment in active and repressive chromatin states, respectively (Extended Data Fig. 1k). Consistent with the classical notion derived from reporter assays that enhancer activity is independent of genomic orientation³¹, enhancers inserted in forward or reverse orientations generated equivalent transcription levels (Fig. 1i). Interestingly, cell-to-cell heterogeneity in eGFP levels (assessed using coefficients of variation (CVs)) showed an opposite trend to mean expression levels and increased with increasing enhancer–promoter genomic

distance (Extended Data Fig. 1l; examples of eGFP intensity distributions are provided in Extended Data Fig. 1m). Importantly, these results did not depend on the specific fluorescence gate used to define eGFP⁺ cells (Extended Data Fig. 1n, o). Together, these data show that the range of activity of the enhancer extends to the entire TAD and is delimited by its boundaries. However, transcription levels and their cell-to-cell variability quantitatively depend on enhancer–promoter genomic distance.

Enhancer contacts modulate burst frequency

We next examined the relationship between transcription levels and contact probabilities. Although reads from the wild-type allele might underemphasize changes introduced by the heterozygous insertion of the transgene, contact patterns detected in capture-C did not change substantially in individual cell lines in which the SCR was mobilized compared to the founder line before piggyBac mobilization (Extended Data Fig. 2a). Thus, the ectopic enhancer and promoter do not create prominent specific interactions, which enabled us to use capture-C data from the founder line (Methods)³² to infer contact probabilities between promoter and enhancer locations (Fig. 2a). Contact probabilities steeply decayed with increasing genomic distance from the promoter, fell considerably while approaching TAD boundaries (from 1 to around 0.05) and further dropped by a factor of around 3 across boundaries (Fig. 2a). This is consistent with previous estimations³³ confirmed using cross-linking and ligation-free methods³⁴ and is representative of the contact probabilities experienced by promoters in mES cells (Extended Data Fig. 2b, c). However, such a trend is at odds with our observation that transcription levels rather mildly decreased inside the TAD and dropped to promoter-only levels outside its boundaries (Fig. 1h and Extended Data Fig. 2d). Interestingly, plotting the mean *eGFP* mRNA numbers as a function of contact probabilities revealed a highly nonlinear relationship (Fig. 2b).

Fig. 2: The promoter on rate is a sigmoidal function of enhancer–promoter contact probabilities.



a, Capture-C (6.4 kb resolution) analysis of the founder cell line used for the experiments in Fig. 1 after converting read counts into contact probabilities (top) (Methods). Bottom, cross-section showing contact probabilities from the ectopic *Sox2* transgene. Insets: magnification of contact probability across the TAD boundaries. **b**, Mean *eGFP* mRNA numbers per cell plotted against contact probabilities between the ectopic *Sox2* promoter and SCR insertions. The red dots show individual cell lines. The black dots show the average values within equally spaced 20 kb bins \pm s.d. The number of cell lines per bin varies from 1 to 28. **c**, Representative smRNA-FISH images from cell lines in which *eGFP* transcription is driven by the *Sox2* promoter alone (left) or by the SCR located at different distances and contact probabilities (right). Scale bar, 10 μ m. **d**, Distributions of mRNA numbers per cell measured in the cell lines shown in c. The error bars show the minimum and maximum frequency. $n = 3$ technical replicates. The line shows the best fit of the phenomenological two-state model to the experimental data shown in b and d. **e**, Best fit to experimental data of b and d. Best-fit parameters are shown in Extended Data Fig. 3b. **f**, Description of the phenomenological two-state model with a variable on rate. The Hill function describes the dependency of k_{on} on contact probability (p_c). $(\{k\}_{\text{rm}}^{\text{on}})^0$ and $(\{k\}_{\text{rm}}^{\text{on}})^1$ are the minimum and maximum on rates, respectively; c and h are the Hill function critical threshold and the sensitivity parameter, respectively. \emptyset symbolizes degraded RNA. **g**, The best-fitting Hill function for k_{on} (in units of mRNA lifetime δ), corresponding to a sigmoidal curve. **h**, Close-up of e, highlighting the predicted insulation outside the TAD boundaries (red and green shaded areas). Data are presented as in b.

We sought to understand whether such a nonlinear relationship could be related to how enhancer–promoter interactions translate into transcription in individual cells. Transcription occurs in intermittent bursts³⁵ that give rise to variable mRNA numbers in single cells. smRNA-FISH analysis revealed substantial cell-to-cell variability in *eGFP* mRNA numbers in a panel of cell lines in which promoter–SCR contact probabilities ranged from zero (promoter-only control cell line) to one (Fig. 2c). Similar to eGFP protein distributions (Extended Data Fig. 2e), CVs of mRNA distributions increased with decreasing contact probabilities (Extended Data Fig. 2f). Bursty promoter behaviour can generally be described in terms of a two-state model

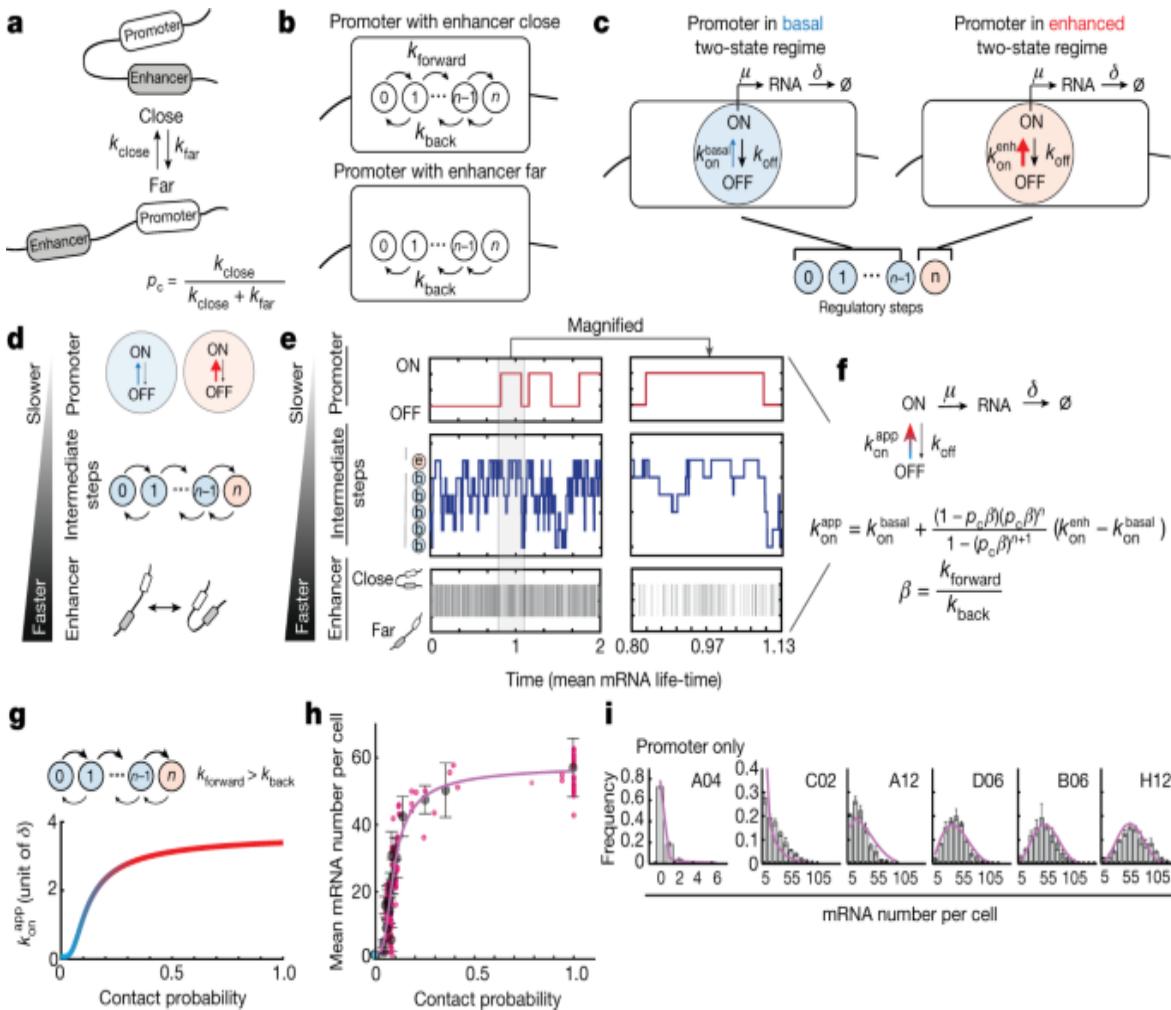
of gene expression³⁶ in which the promoter stochastically switches with rates k_{on} and k_{off} between an OFF and an ON state in which transcription can initiate with rate μ . Consistent with this notion, mRNA number distributions (Fig. 2d) and mean transcription levels (Fig. 2e) in individual cell lines could be well approximated by a phenomenological two-state model in which the ‘on’ rate k_{on} (and therefore the burst frequency) non-linearly depends on enhancer–promoter contact probability through a Hill function (Fig. 2f and [Supplementary Information](#), model description). Interestingly, the best agreement with experimental data occurred with a Hill coefficient (h) of 2.8 (95% confidence interval = 2.4–3.2; Extended Data Fig. 3a, b). This corresponds to a sigmoidal transcriptional response in which the enhancer would be no longer able to activate the promoter outside the approximately threefold drop in contact probabilities generated by TAD boundaries (Fig. 2g, h). Importantly the sigmoidal behaviour of k_{on} was not an artefact due to systematic errors in estimation of contact probabilities (Extended Data Fig. 3c), confounding effects of CTCF sites and repressive chromatin in the 3' part of the TAD, or inclusion of promoter-only cell lines in the fit (Extended Data Fig. 3d). Alternative two-state models in which ‘off’ or initiation rates depend on contact probability rather than the on rate failed to reproduce the observed decrease in CV with contact probabilities ([Supplementary Information](#), model description).

Mechanistic model of enhancer regulation

We next examined which mechanism could in principle generate such a phenomenological two-state model with sigmoidal modulation of k_{on} . Enhancer–promoter contacts are stochastic^{32,37,38} and probably dynamic³⁹ in single cells. Molecular processes that are thought to transmit regulatory information from enhancers to promoters (such as recruitment of transcription factors and coactivators, assembly of the Mediator complex⁴⁰), as well as those that are associated with promoter operation itself (such as pre-initiation complex assembly, RNA polymerase II pausing and release^{41,42}) are also stochastic and dynamic⁴³. We reasoned that the interplay between the timescales of these processes might generate nonlinear effects, as was recently hypothesized to explain promoter bursting⁴⁴. To investigate this concept in a quantitative manner, we developed a

mechanistic model describing the simple hypothesis that, in single cells, the on rate of the promoter is transiently increased after stochastic interactions with an enhancer. We assumed that enhancer–promoter interactions occur and disassemble with rates k_{close} and k_{far} , corresponding to a steady-state contact probability of $k_{\text{close}}/(k_{\text{close}} + k_{\text{far}})$ (Fig. 3a). When the enhancer is close to the promoter, it triggers one or more (n) reversible regulatory steps that transmit information to the promoter with forward and reverse rates k_{forward} and k_{back} (Fig. 3b). These steps are an abstract representation of any stochastic regulatory processes occurring at the enhancer–promoter interface. When the enhancer is far, no information is transmitted to the promoter and regulatory steps can only revert at rate k_{back} (Fig. 3b). The promoter operates in a basal two-state regime with a small on rate ($\langle k \rangle_{\langle \text{rm} \{ \text{on} \} \rangle}^{\langle \text{mathrm} \{ \text{basal} \} \rangle}$) (Fig. 3c) unless all regulatory steps have been completed. In this case, the promoter transiently enters an ‘enhanced’ two-state regime with a higher on rate ($\langle k \rangle_{\langle \text{rm} \{ \text{on} \} \rangle}^{\langle \text{mathrm} \{ \text{enh} \} \rangle}$), thus transiently increasing its transcriptional activity (Fig. 3c and [Supplementary Information](#), model description). A transient increase in promoter activity therefore requires enhancer interactions that are either long enough (Extended Data Fig. 4a) or frequent enough (Extended Data Fig. 4b) to allow the completion of the n regulatory steps.

Fig. 3: A mechanistic model of enhancer–promoter communication.



a, Stochastic promoter–enhancer interactions occur and disassemble with rates k_{close} and k_{far} . **b**, In the close state, the enhancer can trigger n reversible regulatory steps with forward and reverse rates k_{forward} and k_{back} . In the far state, regulatory steps can revert only at rate k_{back} . **c**, The promoter operates in a basal two-state regime with a small on rate ($\langle \{k\}_{\text{on}} \rangle$) unless all n regulatory steps have been completed, in which case it transiently enters an enhanced two-state regime with a higher on rate ($\langle \{k\}_{\text{on}}^{\text{enh}} \rangle$). **d**, Schematic of the parameter constraints under which the mechanistic model reduces to an apparent two-state model: $k_{\text{close,far}} \gg k_{\text{forward,back}} \gg \langle \{k\}_{\text{on}} \rangle$, k_{off} , μ . **e**, Representative single-cell dynamics of enhancer–promoter interactions, promoter regulatory steps and promoter states predicted by the mechanistic model with $n = 5$ and rates satisfying the constraint on timescales described in **d** (time unit, $1/\delta$). **f**, Mathematical derivation of the apparent rates. **g**, Relationship between contact probability and apparent on rate. **h**, Mean mRNA number per cell vs contact probability. **i**, Histograms of mRNA number per cell for various cell lines (A04, C02, A12, D06, B06, H12) with fitted curves.

Reduction of the mechanistic model to an apparent two-state model. The equation describes how the apparent on rate \(\langle k_{\text{on}} \rangle^{\text{app}}\) depends on contact probability (p_c) and other parameters of the mechanistic model. **g**, Dependency of \(\langle k_{\text{on}} \rangle^{\text{app}}\) on contact probability, illustrated for the best fitting parameters shown in **h** and **i**. **h**, Best fit of the apparent two-state model to the experimental transcriptional response shown in Fig. [2b](#). **i**, Best fit of the apparent two-state model to the experimental mRNA distributions shown in Fig. [2c](#). Best-fit parameters are shown in Extended Data Fig. [4c](#).

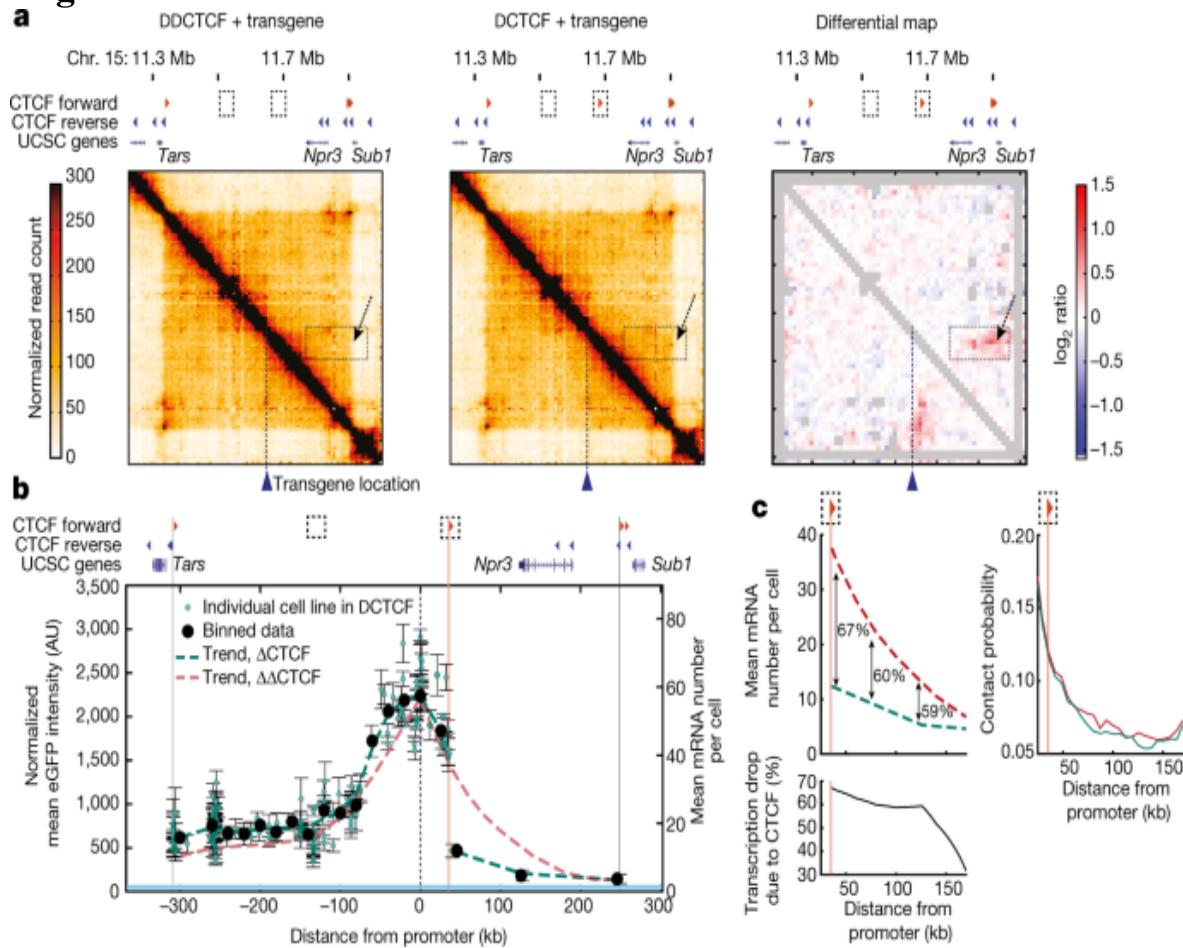
This mechanistic model does not generally reproduce the phenomenological two-state behaviour observed in Fig. [2e](#), [f](#) for the ectopic *Sox2* promoter. However, when the timescales of enhancer–promoter interactions are faster than those of intermediate regulatory steps, and both are faster than the promoter’s intrinsic bursting dynamics ($k_{\text{close,far}} \gg k_{\text{forward,back}} \gg \langle k_{\text{on}} \rangle^{\text{basal,enh}}$), k_{off} , μ) (Fig. [3d](#), [e](#)), the mechanistic model reduces to an apparent two-state model (Fig. [3f](#) and [Supplementary Information](#), model description). If forward transitions through $n > 1$ regulatory steps are favoured over backward reactions ($k_{\text{forward}} > k_{\text{back}}$), then the on rate of the apparent two-state model (\(\langle k_{\text{on}} \rangle^{\text{app}}\)) depends sigmoidally on contact probabilities (Fig. [3g](#)). This shows that, in principle, the promoter’s phenomenological two-state behaviour with sigmoidal modulation of k_{on} observed in Fig. [2e](#), [f](#) could arise from stochastic enhancer–promoter interactions being transmitted into slower promoter ON/OFF dynamics through small numbers of intermediate regulatory processes. The resulting sigmoidal transcriptional response would enable an enhancer to act efficiently even when contact probabilities rapidly decay away from the promoter (Extended Data Fig. [2d](#)), and contribute to block enhancer action when small drops in contact probabilities occur across TAD boundaries (Fig. [2h](#)). The mechanistic model also predicts that enhancer–promoter contacts should not correlate with transcription bursts (Fig. [3e](#)), as recently suggested by simultaneous imaging of *Sox2* transcription and genomic locations flanking the endogenous *Sox2* and SCR^{[20](#)}.

Finally, we verified that, when reduced to a two-state model, the mechanistic model could simultaneously fit the experimental transcriptional response to contact probabilities and smRNA-FISH distributions (Fig. 3*h, i*). Best agreement occurred with five intermediate regulatory steps (95% confidence interval = 3–7; Extended Data Fig. 4*c, d* and [Supplementary Information](#), model description) and, consistent with previous observations²⁰, promoter ON/OFF transitions that occur in the timescale of several minutes (considering that the time unit in the model is mRNA lifetime, expected to be around 1.5 h)⁴⁵ (Extended Data Fig. 4*c, d*). Regulatory processes at the interface between enhancers and promoters have been estimated to occur in the order of tens of seconds^{41,43,46}, consistent with the condition that intermediate regulatory steps should be faster than bursting kinetics (Fig. 3*f*). The requirement that enhancer–promoter interactions should be even faster (Fig. 3*f*) therefore predicts that they should occur on a timescale of seconds or less.

Enhancer strength controls insulation levels

We next set out to examine whether CTCF binding affects the observed nonlinear relationship between transcription and contact probabilities. To this aim, we repeated the enhancer mobilization assay in mES cells in which only one of the two internal CTCF sites was homozygously deleted. The remaining forward CTCF site is located 36 kb downstream of the transgene and loops onto the reverse CTCF sites at the 3' end of the domain (Fig. 4*a*). SCR mobilization in this context resulted in 172 cell lines of which the transcription levels were indistinguishable from those generated in the ‘empty’ TAD, except across the CTCF site that severely, but not completely, insulated the ectopic *Sox2* promoter from the enhancer (Fig. 4*b*). Transcription levels across the CTCF site were about 60% lower than those generated in the absence of the CTCF site (Fig. 4*c*). Strikingly, this occurred in the absence of notable changes in the promoter’s interaction probabilities with the region downstream of the CTCF site, at least in the current experimental set-up (capture-C data with 6.4 kb resolution) (Fig. 4*c*). This suggests that a single CTCF site might exert transcriptional insulation through additional mechanisms beyond simply driving physical insulation, possibly depending on site identity⁴⁷ and flanking sequences¹⁶.

Fig. 4: Insulation by a single CTCF site exceeds contact probability changes.



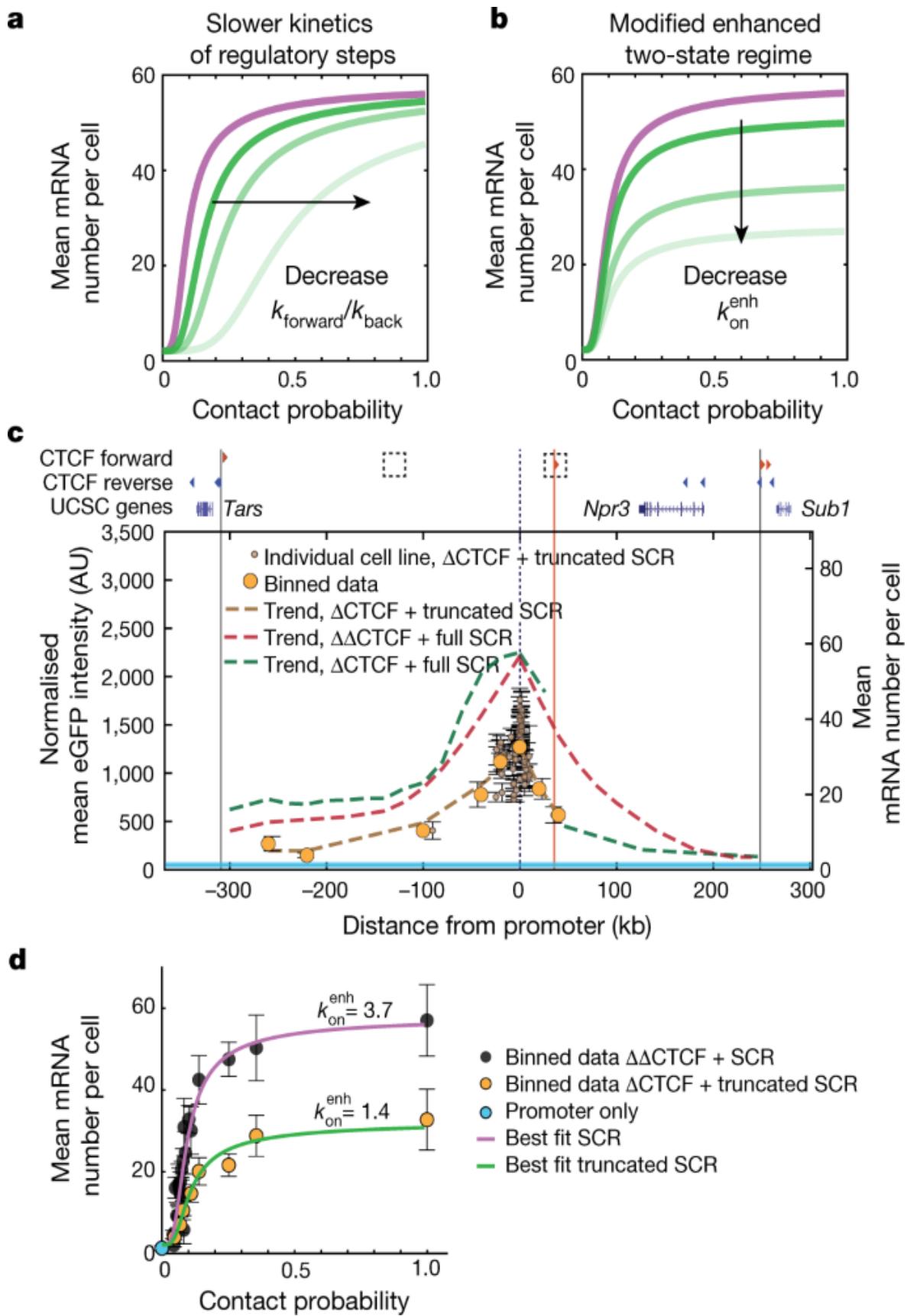
a, Capture-C (6.4 kb resolution) analysis of founder mES cell lines in the absence ($\Delta\Delta$ CTCF) or presence (single CTCF-site deletion, Δ CTCF) of a forward CTCF motif 36 kb downstream of the transgene, and the corresponding differential map. The grey pixels show ‘noisy’ interactions that did not pass quality control filters (Methods). The dotted boxes and arrows indicate the position of the CTCF site and the structural changes it generates. **b**, The normalized mean \pm s.d. eGFP levels in 172 individual eGFP⁺ cell lines following SCR mobilization in Δ CTCF mES cells (green dots); $n = 3$ measurements performed on different days. The black dots show the mean \pm s.d. values within equally spaced 20 kb bins. The green dashed line shows the spline interpolation of average eGFP values. The vertical pink line shows the position of the CTCF site at +36 kb. The red dashed line shows the trend of eGFP levels in the $\Delta\Delta$ CTCF background (compare with Fig. 1h). The blue line shows the promoter-only eGFP level as in Fig. 1h. **c**,

Magnification of spline interpolants of GFP⁺ cell lines in the absence ($\Delta\Delta$ CTCF, red dashed line) or presence (Δ CTCF, green dashed line) of the CTCF binding site at +36 kb (vertical pink line) (left). The numbers represent the percentage fold changes between trendlines. Bottom, the percentage fold changes as a function of distance from the promoter. Right, contact probabilities from the location of the ectopic *Sox2* transgene in $\Delta\Delta$ CTCF (red line) and Δ CTCF (green line) mES cells.

The SCR is a strong enhancer that accounts for most of the transcriptional activity of endogenous *Sox2*^{29,30}. We reasoned that a weaker enhancer should lead to a different transcriptional response to contact probabilities with the promoter. There are two ways in which the parameters in the model shown in Fig. 3f might change when reducing enhancer strength. The ratio between transition rates through regulatory steps k_{forward} and k_{back} (β) in Fig. 3h) might decrease, resulting in a slower transmission of regulatory information (Fig. 5a). This would generate a transcriptional response with maximal transcriptional levels that are similar to those generated by the SCR but different sensitivity to changes in contact probabilities (Fig. 5a). Alternatively (although not exclusively), the on rate in the enhanced promoter regime ($k_{\text{on}}^{\text{enh}}$) could decrease (Fig. 5b). This would conserve the shape of the transcriptional response but decrease the maximal transcription level (Fig. 5b). To test these predictions, we performed the enhancer mobilization assay using a truncated version of the SCR (Extended Data Fig. 5a). This contained only one of the two ~1.5 kb subregions that share similar transcription-factor-binding sites²⁹ and independently operate as weaker enhancers of the *Sox2* promoter in transient reporter assays²⁹ (Extended Data Fig. 5b). Mobilization of the truncated SCR in mES cells with a forward CTCF site downstream of the promoter (compare with Fig. 4a) led to 74 eGFP⁺ cell lines displaying approximately twofold lower transcription levels compared with those generated by the full-length SCR at comparable genomic distances (Fig. 5c). In contrast to the full-length SCR, the truncated enhancer was completely insulated from the promoter by the CTCF site (Fig. 5c). Thus, the level of functional insulation generated by the same CTCF site depends on the strength of the enhancer. In the region upstream of the CTCF sites, the transcriptional response generated by the truncated SCR (Fig. 5d) was in quantitative agreement with model predictions under the hypothesis that enhancer strength decreases the

on rate rather than changing the intermediate regulatory steps (Fig. 5b), and could be predicted using the full-length SCR best-fit parameters with a two-fold decreased $\langle k \rangle_{\rm on}^{\rm rm{enh}}$. This further strengthens our interpretation that enhancer strength modulates the ability of the promoter to turn on, possibly by regulating chromatin state, transcription factor binding or RNA polymerase II dynamics at the promoter^{35,44}.

Fig. 5: Enhancer strength modulates promoter on rates and determines insulation levels through a CTCF site.



a, Model predictions under the hypothesis that decreasing enhancer strength results in a slower flow of regulatory information to the promoter. **b**, Model predictions as in **a**, under the alternative hypothesis that decreasing enhancer strength modifies the enhanced on rate ($\langle k \rangle_{\rm rm{on}}^{\rm \Delta \rm{enh}}$). **c**, Normalized eGFP levels in 74 individual GFP⁺ cell lines (brown dots; the error bars show the s.d. of $n=3$ measurements performed on different days), binned data (orange dots) and data trend (brown dashed line) after mobilization of the truncated SCR in the ΔCTCF background. Trends of eGFP levels in individual GFP⁺ cell lines in which the SCR was mobilized either in the ΔΔCTCF background (red dashed line; Fig. 1h) or in the ΔCTCF background (green dashed line; Fig. 4b) are shown for comparison. Promoter-only eGFP levels (light blue) are shown as in Fig. 1h. **d**, The transcriptional response of the truncated SCR (green line) can be predicted from the best fit to the full-length SCR (purple line) with a modified enhanced on rate ($\langle k \rangle_{\rm rm{on}}^{\rm \Delta \rm{enh}}$). Data are mean \pm s.d. eGFP values were calculated within equally spaced 20 kb bins as in **c**; the number of cell lines per bin varies from 1 to 56.

In the nonlinear transcriptional response that we identified, high sensitivity in the low contact probability regime (that is, at long genomic distances) might contribute to secure insulation by TAD boundaries of even strong enhancers such as the SCR. Interestingly, in mES cells, the contact probabilities of most (~75%) active promoters with the nearest TAD boundary are comparable to those experienced by the ectopic *Sox2* promoter in our experiments (lower than 0.2) (Extended Data Fig. 5c). These promoters should therefore experience the same insulation mechanisms. The remaining promoters are closer (or adjacent) to a TAD boundary and therefore experience larger contact probabilities with the boundary, at which the transcriptional response is less sensitive (Extended Data Fig. 5d). However, interestingly, drops in contact probabilities across a boundary increase with decreasing genomic distance from the boundary itself (Extended Data Fig. 5d). This might contribute to the functional insulation of this class of promoters. Boundaries associated with clusters of CTCF sites might also benefit from the fact that insulation from CTCF sites can exceed the changes in contact probabilities that they generate (Fig. 4).

Discussion

Our study provides unbiased and systematic measurements of promoter output as a function of large numbers of enhancer positions with minimal confounding effects. The analysis of hundreds of cell lines enables us to move beyond locus-specific observations, and establishes a quantitative framework for understanding the role of chromosome structure in long-range transcriptional regulation. Our data reveal that, within a TAD, absolute transcription levels generated by an enhancer depend on its genomic distance from the promoter and are determined by a nonlinear relationship with their contact probabilities. Minimal regulatory and structural complexities introduce deviations from this behaviour and might therefore confound its detection outside a highly controlled genomic environment, notably when studying regulatory sequences in their endogenous context²³. Mathematical modelling suggests that the observed nonlinear transcriptional response involves a modulation of the promoter's burst frequency, which could arise from transient enhancer–promoter interactions being translated into slower promoter bursting dynamics in individual cells. In addition to readily explaining the absence of correlation between transcription and physical proximity in single-cell experiments, this argues that the absence of such correlation should not be interpreted as the absence of causality. Although alternative explanations cannot be ruled out (such as cooperative effects through biomolecular condensates^{21,48}), our model provides a simple explanatory framework for both population-averaged and single-cell behaviour of enhancer-driven transcription, based on a minimal set of general and realistic hypotheses. Future live-cell imaging experiments with improved spatial and temporal resolution⁴⁹ will probably enable the testing of the model's prediction that enhancer–promoter interactions should occur on a timescale of seconds or less, therefore enabling the assessment of the model's premises. Finally, our study reveals that enhancer strength is not only a determinant of absolute transcription levels, but also of the level of insulation provided by CTCF. Our data therefore imply that transcriptional insulation is not an intrinsic absolute property of TAD boundaries or CTCF interactions but, rather, a graded variable depending on enhancer strength, boundary strength and distance from a promoter.

Methods

Culture of embryonic stem cells

All cell lines are based on E14 mES cells, provided by E. Heard's laboratory. Cells were cultured on gelatin-coated culture plates in Glasgow minimum essential medium (Sigma-Aldrich, G5154) supplemented with 15% fetal calf serum (Eurobio Abcys), 1% l-glutamine (Thermo Fisher Scientific, 25030024), 1% sodium pyruvate MEM (Thermo Fisher Scientific, 11360039), 1% MEM non-essential amino acids (Thermo Fisher Scientific, 11140035) 100 µM β-mercaptoethanol, 20 U ml⁻¹ leukaemia inhibitory factor (Miltenyi Biotec, premium grade) in 8% CO₂ at 37 °C. Cells were tested for mycoplasma contamination once a month and no contamination was detected. After piggyBac-enhancer transposition, cells were cultured in standard E14 medium supplemented with 2i (1 µM MEK inhibitor PDO35901 (Axon, 1408) and 3 µM GSK3 inhibitor CHIR 99021 (Axon, 1386)).

Generation of enhancer–promoter piggyBac targeting vectors

Homology arms necessary for the knock-in, the *Sox2* promoter, the SCR and the truncated version of the SCR (Ei) were amplified from E14 mES cell genomic DNA by Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific, F549) using primers compatible with Gibson assembly cloning (NEB, E2611). The targeting vector was generated starting from the 3-SB-EF1-PBBAR-SB plasmid⁵⁰, gifted by Rob Mitra. To clone homology arms into the vector, BspEI and BclI restriction sites were introduced using Q5 Site-Directed Mutagenesis Kit (NEB, E0554). The left homology arm was cloned using Gibson assembly strategy by linearizing the vector with BspEI (NEB, R0540). The right homology arm was cloned using Gibson assembly strategy by linearizing the vector with BclI (NEB, R0160). The *Sox2* promoter was cloned by first removing the Efla promoter from the 3-SB-EF1-PBBAR-SB vector using NdeI (NEB, R0111) and SalI (NEB, R0138) and subsequently using Gibson assembly strategy. The SCR and its truncated version (truncated SCR or Ei) were cloned between the piggyBac transposon-specific inverted terminal repeat sequences (ITR) by linearizing

the vector with BamHI (NEB, R3136) and NheI (NEB, R3131). A transcriptional pause sequence from the human alpha2 globin gene and an SV40 poly(A) sequence were inserted at both 5' and 3' ends of the enhancers using Gibson assembly strategy. A selection cassette carrying the puromycin resistance gene driven by the PGK promoter and flanked by FRT sites was cloned in front of the *Sox2* promoter by linearizing the piggyBac vector with the AsiSI (NEB, R0630) restriction enzyme. A list of the primers used for cloning is provided in [Supplementary Table 1](#).

Generation of founder mES cell lines carrying the piggyBac transgene

The gRNA sequence for the knock-in of the piggyBac transgene on chromosome 15 was designed using the online tool (https://eu.idtdna.com/site/order/designtool/index/CRISPR_SEQUENCE) and purchased from Microsynth AG. gRNA sequence was cloned into the PX459 plasmid (Addgene) using the BsaI restriction site. E14 mES cell founder lines carrying the piggyBac transgene were generated using nucleofection with the Amaxa 4D-Nucleofector X-Unit and the P3 Primary Cell 4D-Nucleofector X Kit (Lonza, V4XP-3024 KT). Cells (2×10^6) were collected with accutase (Sigma-Aldrich, A6964) and resuspended in 100 μ l transfection solution (82 μ l primary solution, 18 μ l supplement, 1 μ g piggyBac targeting vector carrying the SCR, truncated SCR or promoter alone, and 1 μ g of PX459 ch15_gRNA/Cas9) and transferred into a single Nucleocuvette (Lonza). Nucleofection was performed using the protocol CG110. Transfected cells were directly seeded in prewarmed 37 °C culture in E14 standard medium. Then, 24 h after transfection, 1 μ g ml⁻¹ of puromycin (InvivoGen, ant-pr-1) was added to the medium for 3 days to select cells transfected with PX459 gRNA/Cas9 vector. Cells were then cultured in standard E14 medium for an additional 4 days. To select cells with insertion of the piggyBac targeting vector, a second pulse of puromycin was carried out by culturing cells in standard medium supplemented with 1 μ g ml⁻¹ of puromycin. After 3 days of selection, single cells were isolated by fluorescence-activated cell sorting (FACS) on 96-well plates. Sorted cells were kept for 2 days in standard E14 medium supplemented with 100 μ g μ l⁻¹ primocin (InvivoGen, ant-pm-1) and 10 μ M ROCK inhibitor (STEMCELL Technologies, Y-27632). Cells were then cultured in standard

E14 medium with $1 \mu\text{g ml}^{-1}$ of puromycin. Genomic DNA was extracted by lysing cells with lysis buffer (100 mM Tris-HCl pH 8.0, 5 mM EDTA, 0.2% SDS, 50 mM NaCl, proteinase K and RNase) and subsequent isopropanol precipitation. Individual cell lines were analysed by genotyping PCR to determine heterozygous insertion of the piggyBac donor vector. Cell lines showing the corrected genotyping pattern were selected and expanded. A list of the primers used for genotyping is provided in [Supplementary Table 1](#).

Puromycin resistance cassette removal

Cells (1×10^6) were transfected with $2 \mu\text{g}$ of a pCAG-FlpO-P2A-HygroR plasmid encoding for the flippase (Flp) recombinase using Lipofectamine 3000 (Thermo Fisher Scientific, L3000008) according to the manufacturer's instructions. Transfected cells were cultured in standard E14 medium for 7 days. Single cells were then isolated using FACS on 96-well plates.

Genomic DNA was extracted by lysing cells with lysis buffer (100 mM Tris-HCl pH 8.0, 5 mM EDTA, 0.2% SDS, 50 mM NaCl, proteinase K and RNase) and subsequent isopropanol precipitation. Individual cell lines were analysed by genotyping PCR to verify the deletion of the puromycin resistance cassette. A list of the primers used for genotyping is provided in [Supplementary Table 1](#). Cell lines showing the correct genotyping pattern were selected and expanded. Selected cell lines were processed for targeted Nanopore sequencing with Cas9-guided adapter ligation (nCATS)⁵¹ and only the ones showing unique integration of the piggyBac donor vector were used as founder lines for the enhancer mobilization experiments.

Mobilization of the piggyBac-enhancer cassette

A mouse codon-optimized version of the piggyBac transposase (PBase) was cloned in frame with the red fluorescent protein tagRFPt (Evrogen) into a pBroad3 vector (pBroad3_hyPBase_IRES_tagRFPt) using Gibson assembly cloning (NEB, E2611). Cells (2×10^5) were transfected with $0.5 \mu\text{g}$ of pBroad3_hyPBase_IRES_tagRFPt using Lipofectamine 3000 (Thermo Fisher Scientific, L3000008) according to the manufacturer's instructions. To increase the probability of enhancer transposition, typically 12 independent PBase transfections were performed at the same time in 24-well plates. Transfection efficiency as well as expression levels of

hyPBase_IRES_tagRFPt transposase within the cell population were monitored by flow cytometry analysis. Then, 7 days after transfection with PBase, individual eGFP⁺ cell lines were isolated using FACS in 96-well plates. Sorted cells were kept for 2 days in standard E14 medium supplemented with 100 µg ml⁻¹ primocin (InvivoGen, ant-pm-1) and 10 µM ROCK inhibitor (STEMCELL Technologies, Y-27632). Cells were cultured in E14 standard medium for additional 7 days and triplicated for genomic DNA extraction, flow cytometry analysis and freezing.

Sample preparation for mapping piggyBac-enhancer insertion sites in individual cell lines

Mapping of enhancer insertion sites in individual cell lines was performed using splinkerette PCR. The protocol was performed as described previously⁵² with a small number of modifications. Genomic DNA from individual eGFP⁺ cell lines was extracted from 96-well plates using the Quick-DNA Universal 96 Kit (Zymo Research, D4071) according to the manufacturer's instructions. Purified genomic DNA was digested by 0.5 µl of Bsp143I restriction enzyme (Thermo Fisher Scientific, FD0784) for 15 min at 37 °C followed by a heat-inactivation step at 65 °C for 20 min. Long (HMSpAa) and short (HMSpBb) splinkerette adapters were first resuspended with 5× NEBuffer 2 (NEB, B7002) to reach a concentration of 50 µM. Then, 50 µl of HMSpA adapter was mixed with 50 µl of HMSpBb adapter (Aa+Bb) to reach a concentration of 25 µM. The adapter mix was denatured and annealed by heating it to 95 °C for 5 min and then cooling to room temperature. Then, 25 pmol of annealed splinkerette adapters was ligated to the digested genomic DNA using 5 U of T4 DNA ligase (Thermo Fisher Scientific, EL0011) and incubating the samples for 1 h at 22 °C followed by a heat-inactivation step at 65 °C for 10 min. For splinkerette amplifications, PCR 1 was performed combining 2 µl of the splinkerette sample, 1 U of Platinum Taq polymerase (Thermo Fisher Scientific, 10966034), 0.1 µM of HMSp1 and 0.1 µM of PB5-1 (or PB3-1) primer, and splinkerette PCR 2 was performed using 2 µl of PCR 1, 1 U of Platinum Taq polymerase (Thermo Fisher Scientific, 10966034), 0.1 µM of HMSp2 and 0.1 µM of PB5-5 (or PB3-2) primer. The quality of PCR amplification was checked by agarose gel electrophoresis. Samples were sent for Sanger Sequencing (Microsynth AG) using the PB5-2 (or PB3-2) primer. A list of

the primers used for splinkerette PCRs and sequencing is provided in [Supplementary Table 1](#). Mapping of enhancer insertion sites in individual cell lines was performed as described in the ‘Mapping of piggyBac-enhancer insertion sites in individual cell lines’ section.

Flow cytometry eGFP fluorescence intensity measurements and analysis

eGFP⁺ cell lines were cultured in serum + 2i medium for 2 weeks before flow cytometry measurements. eGFP levels of individual cell lines were measured on the BD LSRII SORP flow cytometer using BD High Throughput Sampler (HTS), which enabled sample acquisition in 96-well plate format. Measurements were repeated three times for each clone. Mean eGFP fluorescence intensities were calculated for each clone using FlowJo and all three replicates were averaged.

Normalization of mean eGFP fluorescence intensities

Mean eGFP fluorescence levels of each cell line measured in flow cytometry were first corrected by subtracting the mean eGFP fluorescence intensities measured in wild-type E14 mES cells cultured in the same 96-well plate. The resulting mean intensities were then normalized by dividing them by the average mean intensities of all cell lines where the SCR was located within a 40 kb window centred at the promoter location, and multiplied by a common factor.

Sample preparation for high-throughput sequencing of piggyBac-enhancer insertion sites

Cells (5×10^5) were transfected with 2 µg of PBase using Lipofectamine 3000 (Thermo Fisher Scientific, L3000008) according to the manufacturer’s instructions. Transfection efficiency as well as expression levels of PBase within the cell population were monitored by flow cytometry analysis. Then, 5 days after transfection with PBase, genomic DNA was purified using the DNeasy Blood & Tissue Kit (Qiagen, 69504). To reduce the contribution from cells in which excision of piggyBac-enhancer did not occur, we depleted eGFP sequences using an in vitro Cas9 digestion strategy. gRNA

sequences for eGFP depletion were designed using the online tool (https://eu.idtdna.com/site/order/designtool/index/CRISPR_SEQUENCE) ([Supplementary Table 1](#)). Custom-designed Alt-R CRISPR-Cas9 crRNAs containing the gRNA sequences targeting eGFP (gRNA_1_3PRIME and gRNA_2_3PRIME), Alt-R CRISPR-Cas9 tracrRNA (IDT, 1072532) and Alt-R *Streptococcus pyogenes* Cas9 enzyme (IDT, 1081060) were purchased from IDT. In vitro cleavage of the eGFP fragment by Cas9 was performed according to the IDT protocol ‘In vitro cleavage of target DNA with ribonucleoprotein complex’. In brief, 100 µM of Alt-R CRISPR–Cas9 crRNA and 100 µM of Alt-R CRISPR–Cas9 tracrRNA were assembled by heating the duplex at 95 °C for 5 min and allowing to cool to room temperature (15–25 °C). To assemble the RNP complex, 10 µM of Alt-R guide RNA (crRNA:tracrRNA) and 10 µM of Alt-R *Sp*Cas9 enzyme were incubated at room temperature for 45 min. To perform in vitro digestion of eGFP, 300 ng of genomic DNA extracted from the pool cells transfected with the PBase was incubated for 2 h with 1 µM Cas9/RNP. After the digestion, 40 µg of proteinase K was added and the digested sample was further incubated at 56 °C for 10 min to release the DNA substrate from the Cas9 endonuclease. After purification using AMPURE beads XP (Beckman Coulter, A63881), genomic DNA was digested by 0.5 µl of Bsp143I restriction enzyme (Thermo Fisher Scientific, FD0784) for 15 min at 37 °C followed by a heat-inactivation step at 65 °C for 20 min. Annealed splinkerette adapters (Aa+Bb; 125 pmol) were then ligated to the digested genomic DNA using 30 U of T4 DNA ligase HC (Thermo Fisher Scientific, EL0013), and the samples were incubated for 1 h at 22 °C followed by a heat-inactivation step at 65 °C for 10 min. For splinkerette amplifications, 96 independent PCR 1 reactions were performed combining 100 ng of the splinkerette sample, 1 U of Platinum Taq polymerase (Thermo Fisher Scientific, 10966034), 0.1 µM of HMsP1 and 0.1 µM of PB3-1 primer, and splinkerette PCR 2 was performed using 4 µl of PCR 1 product, 1 U of Platinum Taq polymerase (Thermo Fisher Scientific, 10966034), 0.1 µM of HMsP2 and 0.1 µM of PB3-2 primer. A list of the primers used for splinkerette PCRs is provided in [Supplementary Table 1](#). Splinkerette amplicon products were processed using the NEB Ultra II kit according to the manufacturer’s protocol, using 50 ng of input material. Mapping of genome-wide insertions was performed as described in the ‘Mapping of

piggyBac-enhancer insertion sites in population-based splinkerette PCR' section.

Sample preparation for tagmentation-based mapping of PiggyBac insertions

PiggyBac integrations in pools of cells were mapped using a Tn5-transposon-based ITR mapping technique based on ref. [53](#) with minor alterations. Cells (2×10^5) were transfected with 0.5 µg of PBase using Lipofectamine 3000 (Thermo Fisher Scientific, L3000008) according to the manufacturer's instructions in 24-well plates. Eight independent transfections were performed in parallel. Transfection efficiency as well as expression levels of PBase within the cell population were monitored by flow cytometry analysis. Then, 7 days after transfection with PBase, 6 cell pools of 10,000 cells from low GFP values (gates low 1 and low 2) and 6 cell pools of 337 cells of high GFP values (gate high) were sorted in a 24-well plate. Sorted cells were kept for 2 days in standard E14 medium supplemented with 100 µg ml⁻¹ primocin (InvivoGen, ant-pm-1) and 10 µM ROCK inhibitor (StemCell Technologies, Y-27632). Cells were cultured in E14 standard medium for either 1 passage (pools from gates low 1 and low 2) or 2 passages (pools from gate high) and genomic DNA from individual pools was extracted using the Quick-DNA Miniprep Plus Kit (Zymo Research, D4069) according to the manufacturer's instructions. The Tn5 transposon was produced as described in ref. [54](#). The tagmentation reaction was performed as follows. The primers TAC0101 & TAC0102 (45 µl of 100 µM) each were mixed with 10 µl 10× Tris-EDTA (pH 8) and annealed by heating to 95 °C followed by a slow ramp down (0.1 °C s⁻¹) until 4 °C. The transposome is obtained by combining the adapters (1 µl of 1:2 diluted adapters) and the Tn5 transposon (1.5 µl of 2.7 mg ml⁻¹ stock) in 18.7 µl Tn5 dilution buffer (20 mM HEPES, 500 mM NaCl, 25% glycerol) and incubating the mix for 1 h at 37 °C. The tagmentation was performed by mixing 100 ng of genomic DNA with 1 µl of assembled transposome, 4 µl 5× TAPS-PEG buffer (50 mM TAPS-NAOH, 25 mM MgCl₂, 8% (v/v) PEG8000) in a final volume of 20 µl. The reaction was incubated at 55 °C for 10 min and quenched with 0.2% SDS afterwards. For the best mapping results, both sides of the PiggyBac transposon were processed to obtain 5'

ITR- and 3' ITR-specific libraries. First, we enriched our target region by linear amplification PCR with 3' ITR-specific (TAC0006) and 5' ITR-specific (TAC0099) primers. The PCR mix was 3 µl of tagmented DNA, 1 µl of 1 µM enrichment primer, 2 µl dNTPs (10 mM), 4 µl 5× Phusion HF Buffer (NEB), 0.25 µl Phusion HS Flex polymerase ($2 \text{ U } \mu\text{l}^{-1}$, NEB), in a final volume of 20 µl and amplified as follows: 30 s at 98 °C; 45 cycles of 10 s at 98 °C, 20 s at 62 °C and 30 s at 72 °C; then 20 s at 72 °C. PCR 1 of the library preparation was performed using TAC0161 (3' ITR) and TAC0110 (5' ITR) in combination with N5xx (Illumina, Nextera Index Kit). The PCR mix was 5 µl of enrichment PCR, 1 µl of 10 µM primers, 2 µl dNTPs (10 mM), 4 µl 5× Phusion HF Buffer and 0.25 µl Phusion HS Flex polymerase (NEB), in a final volume of 25 µl and amplified as follows: 30 s at 98 °C; 3 cycles of 10 s at 98 °C, 20 s at 62 °C and 30 s at 72 °C; and 8 cycles of 10 s at 98 °C, 50 s at 72 °C. In PCR 2 the N7xx (Illumina, Nextera Index Kit) adapters were added to the PiggyBac specific locations as follows. PCR was performed with TAC0103 (both ITRs) and N7xx. The PCR mix was 2 µl of PCR1, 1 µl of 10 µM primers, 2 µl dNTPs (10 mM), 4 µl 5× Phusion HF Buffer and 0.25 µl Phusion polymerase (Thermo Fisher Scientific), in a final volume of 22 µl and amplified as follows: 30 s at 98 °C; 10 cycles of 10 s at 98 °C, 20 s at 63 °C and 30 s at 72 °C. Then, 5 µl of library was checked on a 1% agarose gel and different samples were pooled according to smear intensity. Finally, the library was purified by bead purification using CleanPCR (CleanNA) beads at a ratio 1:0.8 sample:beads. The final library was sequenced using the Illumina MiSeq (150 bp, paired-end) system. Mapping of genome-wide insertions was performed as described in the ‘Mapping of piggyBac-enhancer insertion sites by tagmentation’ section.

Deletion of genomic regions containing CTCF-binding sites

gRNA sequences for depletion of the genomic regions containing the CTCF-binding sites were designed using the online tool (https://eu.idtdna.com/site/order/designtool/index/CRISPR_SEQUENCE) and purchased from Microsynth AG ([Supplementary Table 1](#)). gRNA sequences were cloned into the PX459 plasmid (Addgene) using the BsaI restriction site. To remove the first forward CTCF-binding site (chromosome 15: 11520474–11520491), 3×10^5 cells were transfected with 0.5 µg of

PX459 CTCF_KO_gRNA3/Cas9 and 1 µg of PX459 CTCF_KO_gRNA10/Cas9 plasmids using Lipofectamine 2000 (Thermo Fisher Scientific, 11668019) according to the manufacturer's instructions. To remove the second forward CTCF-binding sites (chromosome 15: 11683162–11683179), 1×10^6 cells were transfected with 1 µg of PX459 gRNA2_CTCF_KO/Cas9 and 1 µg of PX459 gRNA6_CTCF_KO/Cas9 plasmids using Lipofectamine 2000 (Thermo Fisher Scientific, 11668019) according to the manufacturer's instructions. Then, 24 h after transfection, 1 µg ml⁻¹ of puromycin was added to the medium for 3 days. Cells were then cultured in standard E14 medium for an additional 4 days. To select cell lines with homozygous deletion, single cells were isolated by FACS on 96-well plate. Sorted cells were kept for 2 days in E14 standard medium supplemented with 100 µg ml⁻¹ primocin (InvivoGen, ant-pm-1) and 10 µM ROCK inhibitor (STEMCELL Technologies, Y-27632). Cells were then cultured in standard E14 medium. Genomic DNA was extracted by lysing cells with lysis buffer (100 mM Tris-HCl pH 8.0, 5 mM EDTA, 0.2% SDS, 50 mM NaCl, proteinase K and RNase) and subsequent isopropanol precipitation. Individual cell lines were analysed by genotyping PCR to determine homozygous deletion of the genomic regions containing the CTCF-binding sites. Cell lines showing the corrected genotyping pattern were selected and expanded. A list of the primers used for genotyping is provided in [Supplementary Table 1](#).

smRNA-FISH

Cells were collected with accutase (Sigma-Aldrich, A6964) and adsorbed on poly-l-lysine (Sigma-Aldrich, P8920) precoated coverslips. Cells were then fixed with 3% PFA (EMS, 15710) in PBS for 10 min at room temperature, washed with PBS and kept in 70% ethanol at -20 °C. After at least 24 h incubation in 70% ethanol, the coverslips were incubated for 10 min with freshly prepared wash buffer composed of 10% formamide (Millipore Sigma, S4117) in 2× SSC (Sigma-Aldrich, S6639). The coverslips were hybridized overnight (around 16 h) at 37 °C in freshly prepared hybridization buffer composed of 10% formamide, 10% dextran sulfate (Sigma-Aldrich, D6001) in 2× SSC and containing 125 nM of RNA-FISH probe sets against Sox2 labelled with Quasar 670 (Stellaris) and against eGFP labelled with Quasar 570 (Stellaris). After hybridization, the

coverslips were washed twice with wash buffer prewarmed to 37 °C for 30 min at 37 °C with shaking, followed by 5 min incubation with 500 ng ml⁻¹ DAPI solution (Sigma-Aldrich, D9564) in PBS (Sigma-Aldrich, D8537). The coverslips were then washed twice in PBS and mounted on slides with Prolong Gold medium (Invitrogen, P36934) and cured at room temperature for 24 h. The coverslips were then sealed and imaged within 24 h.

RNA-FISH image acquisition

Images were acquired on a Zeiss Axion Observer Z1 microscope equipped with 100 mW 561 nm and 100 mW 642 nm HR diode solid-state lasers, an Andor iXon 885 EMCCD camera, and an α Plan-Fluar $\times 100/1.45$ NA oil-immersion objective. Quasar 570 signal was collected with the DsRed ET filter set (AHF Analysentechnik, F46-005), Quasar 670 with Cy5 HC mFISH filter set (AHF Analysentechnik, F36-760) and DAPI with the Sp. Aqua HC-mFISH filter set (AHF Analysentechnik, F36-710). The typical exposure time for RNA-FISH probes was set to around 300–500 ms with 15–20 EM gain and 100% laser intensity. DAPI signal was typically imaged with an exposure time of 20 ms with EM gain 3 and 50% laser intensity. The pixel size of the images was 0.080 \times 0.080 μm with a z-step of 0.25 μm for around 55–70 z-planes.

Image processing and quantification of mRNA numbers

Raw images were processed in KNIME, python and Fiji to extract the numbers of RNAs per cell. The KNIME workflow described below is based on a previously published workflow⁵⁵. z-stacks were first projected to a maximal projection for each fluorescence channel. Individual cells were then segmented using the DAPI channel using Gaussian convolution ($\sigma = 3$), followed by filtering using global threshold with Otsu filter, watershed and connected component analysis for nuclei segmentation. Cytoplasmic areas were then estimated with seeded watershed. Cells with nuclei partially outside the frame of view were automatically excluded. Cells containing obvious artifacts, wrongly segmented or not fully captured in xyz dimensions were manually excluded from the final analysis. Spot detection is based on the Laplacian of Gaussian method implemented in TrackMate⁵⁶. For the channels containing RNA-FISH probes signal, RNAs spots were detected

after background subtraction (rolling ball radius 20–25 pixels) by selecting spot size 0.2 μ m and threshold for spot detection based on visual inspection of multiple representative images. Spot detection is based on the Laplacian of Gaussian method from TrackMate. Subpixel localization of RNA spots was detected for RNA channels and a list of spots per cell for each experimental condition and replicate was generated. Spots in each channel were then aggregated by cell in python to extract the number of RNAs per cell.

Enhancer reporter assays

To generate vectors for the enhancer reporter assay, the *Sox2* promoter, SCR and the truncated versions of the SCR (Ei and Eii) were amplified from E14 mES cell genomic DNA with Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific, F549) using primers compatible with Gibson assembly strategy. The *Sox2* promoter was cloned into the 3-SB-EF1-PBBAR-SB vector as described above. The SCR and the truncated versions Ei and Eii were cloned in front of the *Sox2* promoter by linearizing the vector with AgeI (NEB, R3552) and subsequently using Gibson assembly cloning. A transcriptional pause sequence from the human α 2-globin gene and an SV40 poly(A) sequence was inserted at both the 5' and 3' ends of the enhancers. To test enhancers activity, 3×10^5 cells were co-transfected with 0.5 μ g of the different versions piggyBac vectors and 0.5 μ g of pBroad3_hyPBase_IRES_tagRFPt using Lipofectamine 2000 (Thermo Fisher Scientific, 11668019) according to the manufacturer's instructions. As a control, only 0.5 μ g of the piggyBac vector carrying the *Sox2* promoter was transfected. 24 h after transfection, cells were collected and analysed by flow cytometry.

Capture-C sample preparation

Cells (20×10^6) were cross-linked with 1% formaldehyde (EMS, 15710) for 10 min at room temperature and quenched with glycine (final concentration, 0.125 M). Cells were lysed in 1 M Tris-HCl pH 8.0, 5 M NaCl and 10% NP40 and complete protease inhibitor (Sigma-Aldrich, 11836170001) and enzymatically digested using 1,000 U of MboI (NEB, R0147). Digested chromatin was then ligated at 16 °C with 10,000 U of T4 DNA ligase (NEB,

M0202) in ligase buffer supplemented with 10% Triton X-100 (Sigma-Aldrich, T8787) and 240 µg of BSA (NEB, B9000). Ligated samples were de-cross-linked with 400 µg proteinase K (Macherey Nagel, 740506) at 65 °C and phenol–chloroform purified. 3C library preparation and target enrichment using a custom-designed collection of 6,979 biotinylated RNA ‘baits’ targeting single MboI restriction fragments chromosome 15: 10283500–13195800 (mm9) ([Supplementary Table 2](#); Agilent Technologies; designed as in ref. ⁵⁷) were performed according to the SureSelectXT Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library protocol. The only exceptions were the use of 9 µg of 3C input material (instead of 3 µg) and shearing of DNA using Covaris sonication with the following settings: duty factor: 10%; peak incident power: 175; cycles per burst: 200; treatment time: 480 s; bath temperature: 4 °C to 8 °C).

Targeted nCATS analysis

gRNA sequences targeting specific genomic regions of chromosome 15 external to the homology arms of the transgene were designed using the online tool (https://eu.idtdna.com/site/order/designtool/index/CRISPR_SEQUENCE) ([Supplementary Table 1](#)). Custom-designed Alt-R CRISPR–Cas9 crRNAs (5 crRNAs targeting the region upstream and 5 crRNAs targeting the region downstream the integrated transgene), Alt-R CRISPR–Cas9 tracrRNA (IDT, 1072532) and Alt-R *Sp*Cas9 enzyme (IDT, 1081060) were purchased from IDT. Sample preparation and Cas9 enrichment were performed according to a previously described protocol⁵¹ with a few modifications. Genomic DNA from mES cell founder lines was extracted using the Gentra Puregene Cell Kit (Qiagen, 158745) according to the manufacturer’s instructions. The quality of the high molecular mass DNA was checked using the TapeStation (Agilent) system. Typically, 5 µg of high molecular mass DNA was processed for incubation using shrimp alkaline phosphatase (rSAP; NEB, M0371) for 30 min at 37 °C followed by 5 min at 65 °C to dephosphorylate DNA-free ends. For Cas9 enrichment of the target region, all ten Alt-R CRISPR–Cas9 crRNAs were first pooled at an equimolar amount (100 µM) and subsequently incubated with 100 µM of Alt-R CRISPR–Cas9 tracrRNA at 95 °C for 5 min to assemble the Alt-R guide RNA duplex (crRNA:tracrRNA). To assemble the RNP complex, 4 pmol of Alt-R *Sp*Cas9

enzyme was incubated with 8 pmol Alt-R guide RNA (crRNA:tracrRNA) at room temperature for 20 min. In vitro digestion and A-tailing of the DNA were performed by adding 10 µl of the RNP complex, 10 mM of dATP (NEB, N0440) and 5 U of Taq Polymerase (NEB, M0267) and incubating the samples for 30 min at 37 °C followed by 5 min at 72 °C. Adapter ligation for Nanopore sequencing was performed using the Ligation Sequencing Kit (Nanopore, SQK-CAS109) according to the manufacturer's instructions. After purification with AMPure PB beads (Witec, 100-265-900), the samples were loaded into the MiION system, selecting the SQK-CAS109 protocol.

Nanopore sequencing analysis

To map Nanopore sequencing reads, we first built a custom genome consisting of the transgene sequence flanked by ~10 kb mouse genomic sequence upstream and downstream of the target integration site. The custom genome can be found at GitHub (https://github.com/zhanyinx/Zuin_Roth_2021/blob/main/Nanopore/cassette/cassette.fa). Reads were mapped to the custom genome using minimap2 (v.2.17-r941) with the '-x map-ont' parameter. Nanopore sequencing analysis has been implemented using Snakemake workflow (v.3.13.3). Reads were visualized using IGV (v.2.9.4). The full workflow can be found at GitHub (https://github.com/zhanyinx/Zuin_Roth_2021).

RNA-sequencing sample preparation and analysis

Mouse embryonic stem cells were collected with accutase (5 min, 37 °C) and counted. Cells (3×10^5) were lysed with 300 µl TRIzol reagent. RNA was extracted using the Direct-Zol RNA extraction kit from Zymo. Library preparation was performed after Illumina TruSeq Stranded mRNA-seq according to the manufacturer protocol. Reads were mapped to the *Mus musculus* genome (build mm9) using STAR⁵⁸, using the following options: --outSJfilterReads Unique --outFilterType BySJout --outFilterMultimapNmax 10 --alignSJoverhangMin 6 --alignSJDBoverhangMin 2 --outFilterMismatchNoverLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --outSAMstrandField intronMotif --outFilterIntronMotifs RemoveNoncanonicalUnannotated --

outSAMtype BAM SortedByCoordinate --seedSearchStartLmax 50 --twopassMode basic. Gene expression was quantified using qCount from QuasR package⁵⁹ using the ‘TxDb.Mmusculus.UCSC.mm9.knownGene’ database for gene annotation (Bioconductor package: Carlson M and Maintainer BP. TxDb.Mmusculus.UCSC.mm9.knownGene: Annotation package for TxDb object(s); R package v.3.2.2). Active promoters were defined as genes with $\log_2[\text{RPKM} + 0.1]$ higher than 1.5.

Capture-C analysis

Capture-C data were analysed using HiC-Pro⁶⁰ (v.2.11.4); the parameters can be found at GitHub (https://github.com/zhanyinx/Zuin_Roth_2021). In brief, read pairs were mapped to the mouse genome (build mm9). Chimeric reads were recovered after recognition of the ligation site. Only unique valid pairs mapping to the target regions were used to build contact maps. Iterative correction⁶¹ was then applied to the binned data. The target regions can be found at GitHub (https://github.com/zhanyinx/Zuin_Roth_2021). For SCR_ΔΔCTCF, SCR_ΔCTCF and the derived clonal lines, data from replicate one were used to make the quantification and plots throughout the manuscript.

Differential capture-C maps

To evaluate the structural perturbation induced by the insertion of the transgene and the mobilization of the enhancer (ectopic sequences), we accounted for differences in genomic distances due to the presence of the ectopic sequence. In the founder cell line (for example, SCR_ΔΔCTCF), insertion of the transgene modifies the genomic distance between loci upstream and downstream the insertion site. To account for these differences, we generated distance-normalized capture-C maps in which each entry corresponds to the interaction normalized to the corrected genomic distance between the interacting bins. Outliers (defined using the interquartile rule) or bins with no reported interactions from capture-C were treated as noise and filtered out. Singletons, defined as the top 0.1 percentile of Z-score, were also filtered out. The Z-score is defined as $(\text{obs} - \text{exp})/\text{stdev}$, where obs is the capture-C signal for a given interaction and exp and stdev are the genome-wide average and standard deviation, respectively, of

capture-C signals at the genomic distance separating the two loci. We next calculated the ratios between distance normalized and noise-filtered capture-C maps. A bilinear smoothing with a window of 2 bins was applied to the ratio maps to evaluate the structural perturbation induced by the insertion of the ectopic sequence.

Chromatin state calling with ChromHMM

Chromatin states were called using ChromHMM²⁸ with four states. The list of histone modification datasets used is provided in [Supplementary Table 3](#). States with enrichment in H3K9me3 and H3K27me3 were merged, therefore resulting in three chromatin states: active (enriched in H3K27ac, H3K36me3, H3K4me1 and H3K9ac), repressive (enriched in H3K9me3 and H3K27me3) and neutral (no enrichment).

Mapping of piggyBac-enhancer insertion sites in population-based splinkerette PCR

To identify true-positive enhancer re-insertion sites, we first filtered out reads containing eGFP fragments. We then retained only read pairs for which one side mapped to the ITR sequence and the other side mapped to the splinkerette adapter sequence. We mapped separately the ITR/splinkerette sides of the read pair to the mouse genome (build mm9) using BWA mem⁶² with the default parameters. Only integration sites that had more than 20 reads from both ITR and splinkerette sides were retained.

Mapping of piggyBac-enhancer insertion sites in individual cell lines

To map the enhancer position in individual cell lines, Sanger sequencing (Microsynth) without the adapter sequences were filtered out. The first 24 bp of each read after the adapter was then mapped to the mouse genome (mm9) using vmatchPattern (Biostrings v.2.58.0). The script used to map Sanger sequencing can be found at GitHub (https://github.com/zhanyinx/Zuin_Roth_2021).

Mapping of piggyBac-enhancer insertion sites by tagmentation

Before aligning paired-end sequencing reads, reads were filtered using an adaptation of cutadapt⁶³, processing each read pair in multiple steps. Sequence patterns originating from Tn5 and each ITR were removed. The paired-end reads coming from both ITRs were treated the same. First, the presence of the unique part of the 5' ITR and 3' ITR sequence was detected at the start of the second read of the pair and, if present, this sequence was trimmed. Next, the sequence up to and including the TTAA site that was found on both the 5'ITR and 3'ITR was trimmed off. This sequence only partly contained the respective primers used for each ITR, and was used to filter reads that contained the sequence expected for a correct PCR product starting at the transposon. The sequence up to, but not including, the TTAA was removed. Next, all of the other sequence patterns coming from either Tn5 or the ITR were removed from the 5' end of the first read in the pair and the 3' end of both reads.

After filtering and trimming the reads, the reads were aligned to a reference genome with an *in silico* insertion of the split-GFP construct, but with a single TTAA motif instead of the PiggyBac transposon. This was done by aligning the homology arms found in the plasmid against mm10 reference genome. The complete sequence on the reference matching both arms was replaced by the plasmid sequence inserted.

Alignment was performed using Bowtie2 with the fragment length set to a minimum of 0 bp and maximum of 2,000 bp and the very-sensitive option was used. After reads were aligned to the genome, sambamba⁶⁴ was used to remove duplicates and samtools⁶⁵ was used to filter out read pairs that were not properly paired. We then designated, for each read pair, the position of the first 4 nucleotides of the second read as a putative insertion site. To calculate the fraction of reads originating from the non-mobilized position, the number of read pairs that overlapped the non-mobilized position (the TTAA replacing the PiggyBac of the *in silico* insert) was divided over the total number of reads originating from putative insertion sites supported by at least one read pair with a mapping quality higher than 2. Confident insertions were identified as those with at least one read for both 5' and 3' ITR.

Calibration of the mean number of mRNAs per cell with smRNA-FISH

A linear model was used to predict the average number of eGFP mRNAs on the basis of the mean eGFP intensity. The model was fitted on 7 data points corresponding to the average number of eGFP mRNAs obtained using single-molecule RNA fluorescence *in situ* and the mean eGFP intensity obtained by flow cytometry (Extended Data Fig. 1h; $R^2 = 0.9749$, $P < 0.0001$, *t*-test).

Mathematical model and parameter fitting

The phenomenological two-state model (Fig. 2) and the apparent two-state model deduced from the mechanistic enhancer–promoter model (Fig. 3) were both fitted simultaneously to the mean eGFP levels measured in individual cell lines and to the distributions of RNA numbers measured by smRNA-FISH in six cell lines where the SCR was located at different distances from the promoter. The mean number of mRNAs was calculated analytically and the steady-state distribution of the number of mRNA per cell was approximated numerically ([Supplementary Information](#), model description). The parameters for the phenomenological two-state model are the minimum on rate $\langle k \rangle_{\rm on}^0$, the minimum on rate $\langle k \rangle_{\rm on}^1$, the off rate $k_{\rm off}$, the initiation rate μ and the constant c and Hill exponent h , which together control the nonlinear dependency of $k_{\rm on}$ on contact probability. The parameters for the apparent two-state model are the basal on rate $\langle k \rangle_{\rm on}^{\rm basal}$, the enhanced on rate $\langle k \rangle_{\rm on}^{\rm enh}$, the off rate $k_{\rm off}$, the initiation rate μ , the ratio between the forward and backward rates of the regulatory steps β and the number of regulatory steps n . All of these parameters were considered to be free in the fitting procedure. The apparent two-state model was also fitted to the binned mean number of mRNA molecules inferred from the eGFP⁺ cell lines with the truncated version of the SCR (Fig. 4). In this case, three versions of the apparent two-state model were fitted to the data using log-transformed likelihood ratios. The parameter β (version 1) or $\langle k \rangle_{\rm on}^{\rm enh}$ (model 2) or both (model 3) were

considered to be free parameters, whereas the other parameters were fixed to the best fit values obtained for the full-length SCR dataset. Using log-transformed likelihood ratios, the fit of the three versions was compared to the fit of the model for which all of the parameters were considered to be free. The mathematical description of the enhancer–promoter communication model, the derivation of the apparent two-state model, and the fitting procedures are explained in detail in the [Supplementary Information](#) (model description).

Reporting summary

Further information on research design is available in the [Nature Research Reporting Summary](#) linked to this paper.

Data availability

All capture-C, RNA-seq, Oxford Nanopore, tagmentation and population-based splinkerette PCR sequencing fastq files generated in this study have been uploaded to the Gene Expression Omnibus (GEO) under accession number [GSE172257](#). The following public databases were used:

BSgenome.Mmusculus.UCSC.mm9

(<https://bioconductor.org/packages/release/data/annotation/html/BSgenome.Mmusculus.UCSC.mm9.html>), TxDb.Mmusculus.UCSC.mm9.knownGene (<https://bioconductor.org/packages/release/data/annotation/html/TxDb.Mmu sculus.UCSC.mm9.knownGene.html>).

Code availability

Custom codes generated in this study are available at GitHub (https://github.com/zhanyinx/Zuin_Roth_2021 (cHiC, Nanopore, Insertion mapping); https://github.com/gregroth/Zuin_Roth_2021 (mathematical model); and https://github.com/vansteensellab/tagmap_hopping/tree/giorgetti (tagmentation-based mapping of PiggyBac insertions)).

References

1. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
2. Dowen, J. M. et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374–387 (2014).
3. Lupiáñez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
4. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, aad9024 (2016).
5. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* **167**, 1170–1187 (2016).
6. McCord, R. P., Kaplan, N. & Giorgetti, L. Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function. *Mol. Cell* **77**, 688–708 (2020).
7. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
8. Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016).
9. Ghavi-Helm, Y. et al. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.* **51**, 1272–1282 (2019).
10. Despang, A. et al. Functional dissection of the *Sox9–Kcnj2* locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).
11. Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944 (2017).

12. Rao, S. S. P. et al. Cohesin loss eliminates all loop domains. *Cell* **171**, 305–320 (2017).
13. Galupa, R. et al. A conserved noncoding locus regulates random monoallelic *Xist* expression across a topological boundary. *Mol. Cell* **77**, 352–367 (2020).
14. Guo, Y. et al. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* **162**, 900–910 (2015).
15. Hanssen, L. L. P. et al. Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nat. Cell Biol.* **19**, 952–961 (2017).
16. Huang, H. et al. CTCF mediates dosage and sequence-context-dependent transcriptional insulation through formation of local chromatin domains. *Nat. Genet.* **53**, 1064–1074 (2021).
17. de Wit, E. et al. CTCF binding polarity determines chromatin looping. *Mol. Cell* **60**, 676–684 (2015).
18. Rodríguez-Carballo, E. et al. Chromatin topology and the timing of enhancer function at the *HoxD* locus. *Proc. Natl Acad. Sci. USA* **117**, 31231–31241 (2020).
19. Benabdallah, N. S. et al. Decreased enhancer–promoter proximity accompanying enhancer activation. *Mol. Cell* **76**, 473–484 (2019).
20. Alexander, J. M. et al. Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. *eLife* **8**, e41769 (2019).
21. Xiao, J. Y., Hafner, A. & Boettiger, A. N. How subtle changes in 3D structure can create large changes in transcription. *eLife* **10**, e64320 (2021).
22. Fukaya, T., Lim, B. & Levine, M. Enhancer control of transcriptional bursting. *Cell* **166**, 358–368 (2016).

23. Fulco, C. P. et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
24. Anderson, E., Devenney, P. S., Hill, R. E. & Lettice, L. A. Mapping the Shh long-range regulatory domain. *Development* **141**, 3934–3943 (2014).
25. Symmons, O. et al. Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* **24**, 390–400 (2014).
26. Symmons, O. et al. The Shh topological domain facilitates the action of remote enhancers by reducing the effects of genomic distances. *Dev. Cell* **39**, 529–543 (2016).
27. Li, M. A. et al. The piggyBac transposon displays local and distant reintegration preferences and can cause mutations at noncanonical integration sites. *Mol. Cell. Biol.* **33**, 1317–1330 (2013).
28. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
29. Zhou, H. Y. et al. A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev.* **28**, 2699–2711 (2014).
30. Li, Y. et al. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS ONE* **9**, e114485 (2014).
31. Schaffner, W. Enhancers, enhancers—from their discovery to today’s universe of transcription enhancers. *Biol. Chem.* **396**, 311–327 (2015).
32. Giorgetti, L. et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**, 950–963 (2014).
33. Dekker, J. & Mirny, L. The 3D genome as moderator of chromosomal communication. *Cell* **164**, 1110–1121 (2016).

34. Redolfi, J. et al. DamC reveals principles of chromatin folding in vivo without crosslinking and ligation. *Nat. Struct. Mol. Biol.* **26**, 471–480 (2019).
35. Coulon, A., Chow, C. C., Singer, R. H. & Larson, D. R. Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nat. Rev. Genet.* **14**, 572–584 (2013).
36. Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).
37. Amano, T. et al. Chromosomal dynamics at the *Shh* locus: limb bud-specific differential regulation of competence and active transcription. *Dev. Cell* **16**, 47–57 (2009).
38. Bintu, B. et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, eaau1783 (2018).
39. Chen, H. et al. Dynamic interplay between enhancer–promoter topology and gene activity. *Nat. Genet.* **50**, 1296–1303 (2018).
40. Vernimmen, D. & Bickmore, W. A. The hierarchy of transcriptional activation: from enhancer to promoter. *Trends Genet.* **31**, 696–708 (2015).
41. Nguyen, V. Q. et al. Spatiotemporal coordination of transcription preinitiation complex assembly in live cells. *Mol. Cell* **81**, 3560–3575 (2021).
42. Krebs, A. R. et al. Genome-wide single-molecule footprinting reveals high RNA polymerase II turnover at paused promoters. *Mol. Cell* **67**, 411–422 (2017).
43. Cho, W.-K. et al. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412–415 (2018).
44. Lammers, N. C., Kim, Y. J., Zhao, J. & Garcia, H. G. A matter of time: using dynamics and theory to uncover mechanisms of transcriptional bursting. *Curr. Opin. Cell Biol.* **67**, 147–157 (2020).

45. Horvathova, I. et al. The dynamics of mRNA turnover revealed by single-molecule imaging in single cells. *Mol. Cell* **68**, 615–625 (2017).
46. Chen, J. et al. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell* **156**, 1274–1285 (2014).
47. Anania, C. et al. In vivo dissection of a clustered-CTCF domain boundary reveals developmental principles of regulatory insulation. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.04.14.439779> (2021).
48. Bialek, W., Gregor, T. & Tkačik, G. Action at a distance in transcriptional regulation. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.1912.08579> (2019).
49. Brandão, H. B., Gabriele, M. & Hansen, A. S. Tracking and interpreting long-range chromatin interactions with super-resolution live-cell imaging. *Curr. Opin. Cell Biol.* **70**, 18–26 (2021).
50. Qi, Z. et al. An optimized, broadly applicable piggyBac transposon induction system. *Nucleic Acids Res.* **45**, e55 (2017).
51. Gilpatrick, T. et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* **38**, 433–438 (2020).
52. Uren, A. G. et al. A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nat. Protoc.* **4**, 789–798 (2009).
53. Stern, D. L. Tagmentation-based mapping (TagMap) of mobile DNA genomic insertion sites. Preprint at *bioRxiv* <https://doi.org/10.1101/037762> (2017).
54. Schep, R. et al. Impact of chromatin context on Cas9-induced DNA double-strand break repair pathway balance. *Mol. Cell* **81**, 2216–2230 (2021).
55. Voigt, F., Eglinger, J. & Chao, J. A. in *RNA Detection: Methods and Protocols* (ed. Gaspar, I.) 373–384 (Springer, 2018).

56. Tinevez, J.-Y. et al. TrackMate: an open and extensible platform for single-particle tracking. *Methods* **115**, 80–90 (2017).
57. Schoenfelder, S. et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).
58. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
59. Gaidatzis, D., Lerch, A., Hahne, F. & Stadler, M. B. QuasR: quantification and annotation of short reads in R. *Bioinformatics* **31**, 1130–1132 (2015).
60. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
61. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
62. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.1303.3997> (2013).
63. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
64. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
65. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
66. Zhan, Y. et al. Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Res.* **27**, 479–490 (2017).

Acknowledgements

We thank R. Mitra for sharing the piggyBac-splitGFP vector; A. Boettiger and J.Y. Xiao for discussions on modelling; M. Michalski and S. Andrews for capture-C primer design; L. Gelman and J. Eglinger for help with microscopy and image analysis; and G. Natoli, D. Schübeler, H. Grosshans and E. Nora for discussions and comments on the manuscript. J.Z. was supported by a Marie Skłodowska-Curie grant (748091 ‘3DQuant’).

Research in the Giorgetti laboratory is funded by the Novartis Foundation, the European Research Council (ERC) (759366 'BioMeTre') and Marie Skłodowska-Curie Innovative Training Networks (813327 'ChromDesign' and 813282 'PEP-NET') under the European Union's Horizon 2020 research and innovation program, and the Swiss National Science Foundation (310030_192642). Research in the Meister laboratory is supported by the Swiss National Science Foundation (IZCOZ0_189884/31003A_176226 to P. Meister.). Research in the van Steensel laboratory is supported by ERC Advanced Grant 694466 'GoCADiSC'.

Author information

Author notes

1. These authors contributed equally: Jessica Zuin, Gregory Roth

Affiliations

1. Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland

Jessica Zuin, Gregory Roth, Yinxiu Zhan, Julie Cramard, Josef Redolfi, Ewa Piskadlo, Pia Mach, Mariya Kryzhanovska, Gergely Tihanyi, Hubertus Kohler, Sebastien Smallwood & Luca Giorgetti

2. University of Basel, Basel, Switzerland

Pia Mach & Gergely Tihanyi

3. Division of Gene Regulation and Oncode Institute, Netherlands Cancer Institute, Amsterdam, The Netherlands

Mathias Eder, Christ Leemans & Bas van Steensel

4. University of Bern, Bern, Switzerland

Peter Meister

Contributions

L.G. and J.Z. conceived and designed the study. J.Z., J.C., E.P., M.K. and G.T. performed the experiments. G.R. wrote and analysed the mathematical model. H.K. provided assistance with flow cytometry. J.C., P. Meister. and S.S. contributed to setting up nCATS. S.S. also provided assistance with high-throughput sequence experiments. J.Z., J.C., J.R. and P. Mach. generated cell lines. M.E., C.L. and B.v.S. assisted with tagmentation-based mapping of insertions. G.R. and Y.Z. analysed the data, except for flow cytometry and single-clone insertion mapping (J.Z.) and smRNA-FISH (E.P.). L.G. wrote the paper with G.R., J.Z. and Y.Z., and input from all of the authors.

Corresponding author

Correspondence to [Luca Giorgetti](#).

Ethics declarations

Competing interests

The authors declare no competing interests.

Peer review

Peer review information

Nature thanks Mikhail Spivakov and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Extended data figures and tables

Extended Data Fig. 1 Enhancer action is modulated by genomic distance from the target promoter and constrained by TAD boundaries.

a. Top: capture-C contact map at 6.4 kb resolution in wild-type (WT) mES cells in a 2.6 Mb region centred around the neutral TAD on chromosome 15 we used for the experiments. Vertical grey lines: TAD boundaries. Bottom: genomic datasets and ChromHMM analysis showing that the chosen TAD is devoid of active and repressive chromatin states, with the exception of 80 kb at the 3b at t which is enriched in repressive chromatin states. b. Close-up view of panel a, highlighting the presence of CTCF-mediated chromatin loops (dotted boxes) in WT mES cells. c. capture-C contact map at 6.4 kb resolution for the same region as panel b in the cell line with double CTCF site deletions. CTCF deletions lead to loss of CTCF-mediated chromatin loops (dotted boxes). d. Top: UCSC snapshot of the endogenous *Sox2* locus and *Sox2* control region (SCR). Bottom: close-up views showing the regions of the *Sox2* promoter, the SCR region found in ref. ²⁹ and the SCR used in the transgene construct. e. IGV snapshot showing nanopore sequencing reads mapped to a modified mouse genome including the transgene integration. Reads spanning from genomic DNA upstream the left homology arm to genomic DNA downstream the right homology arm confirmed single insertion of the transgene. f. capture-C maps at 6.4 kb resolution of the mES cell line with double CTCF sites deletion (left) and the founder mES cell line with transgene insertion (centre). Right: differential contact map. Grey pixels correspond to 'noisy'

interactions that did not satisfy our quality control filters (see Methods). Transgene insertion induces new mild interactions with CTCF sites at the 3. and 5a extremities of the TAD (arrows). g. Barplot showing the fraction of piggyBac-SCR reinsertions genome-wide determined by Illumina sequencing of splinkerette PCR products from a pool of cells after PBase expression. See Methods for a detailed description of the protocol. h. Top: Representative smRNA-FISH image and flow cytometry profiles over different passages in a cell line where the SCR was mobilized in the immediate vicinity of the ectopic *Sox2* promoter. Scale bar, 10 μ m. Bottom: Linear relationship between the mean eGFP intensity and the average number of eGFP mRNAs measured using smRNA-FISH for seven single cell lines ($R^2=0.9749$), ($p < 0.0001$, t-test). Error bars on the x-axis: standard deviation of three measurements performed on different days, as in Fig. 1h. Error bars on the y-axis: standard deviation of three technical replicates. i. Normalized mean eGFP intensities levels in individual eGFP+ cell lines are plotted as a function of the genomic position of the SCR in individual eGFP+ lines. Data from 127 individual cell lines (light red dots) from a single experiment are presented as mean +/- standard deviation (n=3 measurements performed in different days, as in Fig. 1g). Average eGFP values calculated within equally spaced 20 kb bins (black dots) are shown. Mean mRNA numbers per cell were inferred from eGFP counts using calibration with smRNA-FISH, see Extended Data Fig. 1h. Shaded light blue area indicates the interval between mean +/- standard deviation of eGFP levels in three promoter-only cell lines. j. Same plot as Fig. 1h showing the only two SCR insertions we detected outside the TAD boundaries (brown dot) and on another chromosome (yellow dot). k. Left: Log10 average eGFP expression (from Fig. 1h) as a function of log10 absolute genomic distance between transgene position and SCR reinsertion. Points are colour-coded as in panel A (chromHMM active, neutral, and repressive states). Black line denotes linear regression. Black circles denote SCR reinsertions within the Npr3 gene body. Right: deviations of eGFP expression levels from the linear regression correlate with chromatin states called using ChromHMM (n: active = 16; neutral = 83; Npr3 = 17; repressive = 7). Reinsertion of SCR within active or repressive regions respectively increases or decreases enhancer activity compared to neutral regions. Box plot: centre line denotes the median; boxes denote lower and upper quartiles (Q1 and Q3, respectively); whiskers denote 1.5x the

interquartile region (IQR) below Q1 and above Q3; points denote outliers. l. Coefficients of variation (CV) of eGFP levels measured by flow cytometry plotted against SCR insertion locations in eGFP+ cell lines (light red dots). Data are presented as mean +/- standard deviation ($n = 3$ measurements in different days). Shaded light blue area indicates the interval between mean +/- standard deviation of eGFP level CVs in three promoter-only cell lines. m. Representative eGFP distributions (normalized to mean eGFP level) in clones with increasing absolute genomic distance (1.7 kb, 42.4 kb, 112.5 kb, and 259.43 kb) between the mobilized enhancer and the ectopic *Sox2* promoter. Vertical line indicates normalized mean eGFP levels. n. FACS plot showing standard (top) and less stringent (bottom) gates on eGFP levels used for single cells sort and insertion analysis of corresponding clonal cell lines. o. Left: FACS plot showing the gates used to sort pools of cells for tagmentation-based mapping of PiggyBac-enhancer insertions. For gates “low 1” and “low 2”, six pools of 10000 cells were sorted while for gate “high”, six pools of 337 cells were sorted. Gate “high” corresponds to the standard gate used to isolate eGFP positive cell lines for the mobilization experiments. Centre: Barplot showing the fraction of sequencing reads mapping to non-mobilized enhancer cassette determined by tagmentation-based mapping from the different pools sorted in gates “low 1”, “low 2” and “high”. See Methods for a detailed description of the protocol. Right: Numbers and genomic locations of confident insertion sites (identified as those with at least one read for both 5'oth 5 mapping from the different pools sorted in gates “low 1”, “low 2” and “high”).

Extended Data Fig. 2 Analysis of chromosome structure around the transgenic locus and genome-wide in mES cells.

a. Top: capture-C maps (6.4 kb resolution) of four cell lines where the SCR (black arrow) has been reinserted at different distances from the promoter (blue arrow). Bottom: differential contact map between individual cell lines and the founder line. Grey pixels: correspond to ‘noisy’ interactions that did not satisfy quality control filters (see Methods). Right: barplot showing the change in average interaction probabilities between the SCR reinsertion and the cassette, calculated using a square of 5 bins (6.4 kb resolution) centred at the cassette SCR reinsertion interaction. b. Left: example of Hi-C heatmap in mES cells at 6.4 kb resolution. Centre: scheme depicting how the

probability of interaction between a promoter and the region immediately before the nearest TAD boundary (P_{in} , 12.8 kb i.e. two 6.4 kb bins before the boundary called using CaTCH⁶⁶) and after the nearest TAD boundary (P_{out}) are calculated. Right: distribution of contact probability between all active promoters in mES cellss and the closest inner TAD boundary (P_{in}) ($n = 9655$). Box plot description as in Extended Data Fig. [1k](#). c. Box plots showing the distribution of contact probability changes within the TAD and across the closest TADs boundary for all active promoters in mES cells ($n = 9655$) whose contact probability outside the TAD is higher than 0.001 ($n = 834$). Box plot description as in Extended Data Fig. [1k](#); outliers not shown. d. Contact probabilities of the founder line from the location of the ectopic *Sox2* transgene (black line) and normalized averaged mean number of mRNAs per cell (highest value = 1) generated in individual eGFP+ lines by the SCR mobilization are plotted as a function of its genomic position (dashed red line). The average is calculated within equally spaced 20 kb bins as in Fig. [1h](#) (black dots). e. Coefficients of variation (CV) of eGFP levels measured by flow cytometry plotted against contact probabilities between the ectopic *Sox2* promoter and the locations of SCR insertions. Data are presented as mean values +/- standard deviation ($n = 3$ measurements in different days). Shaded light blue area indicates the interval between mean +/- standard deviation of eGFP level CVs in three promoter-only cell lines. f. Coefficients of variation (CV) of mRNA number per cell measured by smRNA-FISH plotted against contact probabilities between the ectopic *Sox2* promoter and the locations of SCR in the cell the lines shown in Fig. [2c](#), d. Data are presented as mean values +/- standard deviation ($n = 3$ technical replicates).

Extended Data Fig. 3 Phenomenological two-state model fitting and robustness analysis.

- a. Parameter values and 95% confidence intervals for the best fitting phenomenological two-state model. The rates are in the unit of RNA decay rate ((δ)). b. Profile likelihood functions for all the parameters of the phenomenological two-state model. The red dashed line shows the threshold used to calculate the 95% confidence intervals (see [Supplementary Model description](#) for more details). c. Best fit of the phenomenological two-state model under different perturbations of the contact probabilities. Panels with

blue curves show the best fit transcriptional responses when the scaling exponent of the contact probabilities was artificially set to 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. The scaling exponent of the original contact probabilities is 0.77. Panels with orange curves show the best fit transcriptional responses when contact probabilities were artificially increased by a factor $1/x$ with $x = 0.1, \dots, 0.9$ with step of 0.1. Data are presented as average eGFP values calculated within equally spaced 20 kb bins +/- standard deviation ($n =$ number of cell lines per bin), as in Fig. 1h. d. Left: Normalized mean eGFP intensities in individual eGFP+ cell lines are plotted as a function of the genomic position of the SCR. Data from 135 individual cell lines (light red dots) are presented as mean +/- standard deviation ($n = 3$ measurements performed on different days, as in panel g). Shaded grey area indicates the genomic regions that were excluded from the fit shown in the right panel. Right: Best fit of the phenomenological two-state model in the absence of the promoter-only control cell line and the cell lines with insertions that landed beyond the first CTCF site at the 3' of the TAD (region highlighted in the left panel). Data are presented as average eGFP values calculated within equally spaced 20 kb bins +/- standard deviation ($n =$ number of cell lines per bin). e. Profile likelihood function for the Hill coefficient for the fit described in panel d.

Extended Data Fig. 4 Fit of the mechanistic enhancer–promoter model and robustness analysis.

- a. Schematic description of the dynamics of the mechanistic model (here with two regulatory steps ($n=2$) for illustration). This case illustrates a scenario where, the enhancer–promoter interaction is long enough to allow the completion of the 2 regulatory steps and transiently increases the promoter activity. b. In an alternative scenario, the interactions are shorter but frequent enough to allow the completion of the 2 regulatory steps and transiently increase the promoter activity. c. Parameter values and 95% confidence intervals for the best fitting apparent two-state model. The rates are in the unit of RNA decay rate ((δ)). d. Profile likelihood functions for all the parameters of the apparent two-state model. Red dashed lines show the threshold used to calculate the 95% confidence intervals (see [Supplementary Model description](#) for more details).

Extended Data Fig. 5 Dependence of transcription levels and insulation on enhancer strength.

a. Top: UCSC genome browser snapshot of the endogenous *Sox2* locus and *Sox2* control region (SCR). Bottom: close-up view showing the SCR (black) identified in ref. ²⁹ and the enhancer regions used in the transient reporter assays shown in panel b. Full-length enhancer is in red (same as in Fig. 1); truncated versions are in brown (Ei) and orange (Eii). Experiments in Fig. 5 were performed with Ei. b. Flow cytometry analysis of mES cells transiently transfected with PBase-RFP and different versions of split eGFP plasmids carry either no enhancer, or the full-length SCR (red, see panel a), or the first (brown-Ei) or second (orange-Eii) SCR subregions in front of the *Sox2* promoter. Transcription levels generated upon co-transfection with PBase are higher in the presence of the full-length SCR compared to truncated versions. Numbers in each quadrant represent the % of cells either negative or RFP, GFP and RFP-GFP positive. c. Top: distribution of contact probabilities between all active promoters in mES cells and the nearest inner TAD boundaries, calculated as in Extended Data Fig. 2b. Bottom panel: Model prediction for the mean eGFP mRNA numbers per cell plotted against contact probabilities shown as a comparison (same as Fig. 2e). Shaded areas correspond to promoters with contact probability with the closest TAD boundary below 0.2. d. Left panel: scheme of how the probabilities of interaction between promoter and the region before (P_{in}) and after the TAD boundary (P_{out}) are calculated, same criteria as in Extended Data Fig. 2b. Central panel: promoters with higher contact probabilities with TAD boundaries experience stronger drops of contact probability across boundaries. Right panel: promoters closer to TAD boundaries experience a stronger drop of contact probability across boundaries.

Supplementary information

Supplementary Information

Model description and gating strategy.

Reporting Summary

Supplementary Table 1

Lists of primers used in the study.

Peer Review File

Supplementary Table 2

List of oligonucleotides used in capture-C experiments.

Supplementary Table 3

List of datasets used to call chromatin states using ChromHMM.

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Reprints and Permissions

About this article

Cite this article

Zuin, J., Roth, G., Zhan, Y. *et al.* Nonlinear control of transcription through enhancer–promoter interactions. *Nature* **604**, 571–577 (2022).
<https://doi.org/10.1038/s41586-022-04570-y>

- Received: 22 April 2021
- Accepted: 21 February 2022
- Published: 13 April 2022
- Issue Date: 21 April 2022
- DOI: <https://doi.org/10.1038/s41586-022-04570-y>

Share this article

Anyone you share the following link with will be able to read this content:

[Get shareable link](#)

Sorry, a shareable link is not currently available for this article.

[Copy to clipboard](#)

Provided by the Springer Nature SharedIt content-sharing initiative

This article was downloaded by **calibre** from <https://www.nature.com/articles/s41586-022-04570-y>.

Amendments & Corrections

- [**Author Correction: A graph placement methodology for fast chip design**](#) [31 March 2022]

Author Correction •

[Download PDF](#)

- Author Correction
- [Published: 31 March 2022](#)

Author Correction: A graph placement methodology for fast chip design

- [Azalia Mirhoseini](#) ORCID: orcid.org/0000-0002-2440-0944¹ na1,
- [Anna Goldie](#) ORCID: orcid.org/0000-0002-4887-6293^{1,3} na1,
- [Mustafa Yazgan](#)²,
- [Joe Wenjie Jiang](#)¹,
- [Ebrahim Songhori](#)¹,
- [Shen Wang](#)¹,
- [Young-Joon Lee](#)²,
- [Eric Johnson](#)¹,
- [Omkar Pathak](#)²,
- [Azade Nazi](#)¹,
- [Jiwoo Pak](#)²,
- [Andy Tong](#)²,
- [Kavya Srinivasa](#)²,
- [William Hang](#)³,
- [Emre Tuncer](#)²,
- [Quoc V. Le](#)¹,
- [James Laudon](#) ORCID: orcid.org/0000-0003-1061-1780¹,
- [Richard Ho](#)²,
- [Roger Carpenter](#)² &
- [Jeff Dean](#)¹

[Nature](#) volume **604**, page E24 (2022)

- 566 Accesses
- 2 Altmetric
- [Metrics details](#)

Subjects

- [Computational science](#)
- [Electrical and electronic engineering](#)

The [Original Article](#) was published on 09 June 2021

Correction to: *Nature* <https://doi.org/10.1038/s41586-021-03544-w>
Published online 9 June 2021

In this Article, the ‘Code availability’ section has been updated from: “The code used to generate these data is available from the corresponding authors upon reasonable request” to: “The code used to generate these data is available in the following GitHub repository: https://github.com/google-research/circuit_training”. The original Article has been corrected online

Author information

Author notes

1. These authors contributed equally: Azalia Mirhoseini, Anna Goldie

Affiliations

1. Google Research, Brain Team, Google, Mountain View, CA, USA

Azalia Mirhoseini, Anna Goldie, Joe Wenjie Jiang, Ebrahim Songhori, Shen Wang, Eric Johnson, Azade Nazi, Quoc V. Le, James Laudon & Jeff Dean

2. Google Chip Implementation and Infrastructure (CI2) Team, Google, Sunnyvale, CA, USA

Mustafa Yazgan, Young-Joon Lee, Omkar Pathak, Jiwoo Pak, Andy Tong, Kavya Srinivasa, Emre Tuncer, Richard Ho & Roger Carpenter

3. Computer Science Department, Stanford University, Stanford, CA, USA

Anna Goldie & William Hang

Corresponding authors

Correspondence to [Azalia Mirhoseini](#) or [Anna Goldie](#).

Rights and permissions

[Reprints and Permissions](#)

About this article

Cite this article

Mirhoseini, A., Goldie, A., Yazgan, M. *et al.* Author Correction: A graph placement methodology for fast chip design. *Nature* **604**, E24 (2022). <https://doi.org/10.1038/s41586-022-04657-6>

- Published: 31 March 2022
- Issue Date: 21 April 2022
- DOI: <https://doi.org/10.1038/s41586-022-04657-6>

Share this article

Anyone you share the following link with will be able to read this content:

[Get shareable link](#)

Sorry, a shareable link is not currently available for this article.

[Copy to clipboard](#)

Provided by the Springer Nature SharedIt content-sharing initiative

This article was downloaded by **calibre** from <https://www.nature.com/articles/s41586-022-04657-6>

| [Section menu](#) | [Main menu](#) |

Focal Point