A smaller, faster, cheaper route to fusion energy *p. 854*

Sport, sex, and testosterone go to court *p. 858*

Making C–N bonds from nitro groups *pp. 863 & 886*

Science

$10
22 MAY 2015
sciencemag.org

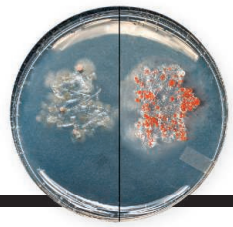
AAAS

A world of plankton

Marine biodiversity surveyed by the schooner *Tara*

pp. 865 & 873

CONTENTS



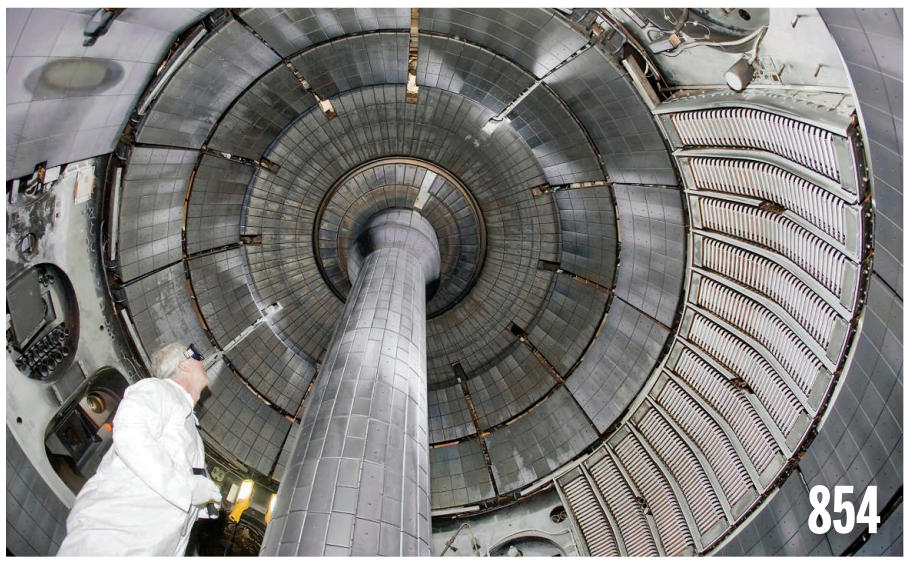
22 MAY 2015 • VOLUME 348 • ISSUE 6237

NEWS

IN BRIEF

IN DEPTH

FEATURES



INSIGHTS

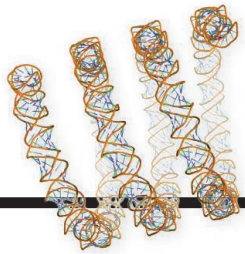
PERSPECTIVES

BOOKS *ET AL.*

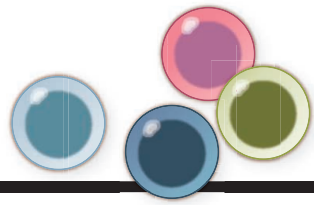
LETTERS

PHOTO: ELLE STARKMAN/PPPL OFFICE OF COMMUNICATIONS

Published by AAAS

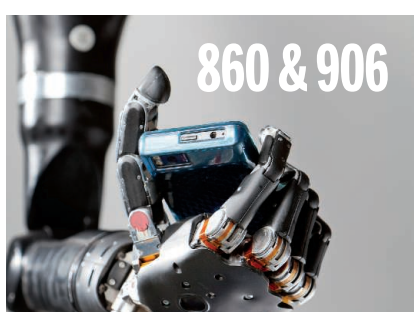


914
A-form DNA can stand the heat



879
Innate lymphoid cells have diverse functions

RESEARCH



Various species of plankton. Using the research vessel *Tara*, the *Tara* Oceans expedition sampled plankton ecosystems around the world and collected associated oceanographic data. The wet part of the expedition was accompanied by an integrative analysis on land, incorporating DNA sequencing and bioinformatics, microscopic imaging, and network analysis. See pages 841, 865, 870, and 873. *Photos: C. Sardet/CNRS/*Tara *Oceans/Plankton Chronicles/*Plankton: Wonders of the Drifting World

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices.

Domestic individual membership and subscription (51 issues): $153 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $1282; Foreign postage extra: Mexico, Caribbean (surface mail) $55; other countries (air assist delivery) $85. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #1254 88122. Publications Mail Agreement Number 1069624. **Printed in the U.S.A.**
Change of address: Allow 4 weeks, giving old and new addresses and 8-digit account number. **Postmaster:** Send change of address to AAAS, P.O. Box 96178, Washington, DC 20090–6178. **Single-copy sales:** $10.00 current issue, $15.00 back issue prepaid includes surface postage; bulk rates on request.

Published by AAAS

Science
AAAS
1200 New York Avenue, NW, Washington, DC 20005
Bateman House, 82-88 Hills Road, Cambridge, UK CB2 1LQ

SUBSCRIPTION SERVICES For change of address, missing issues, new orders and renewals, and payment questions: 866-434-AAAS (2227) or 202-326-6417, FAX 202-842-1065. Mailing addresses: AAAS, P.O. Box 96178, Washington, DC 20090-6178 or AAAS Member Services, 1200 New York Avenue, NW, Washington, DC 20005

INSTITUTIONAL SITE LICENSES 202-326-6755 **REPRINTS:** Author Inquiries 800-635-7181 **COMMERCIAL INQUIRIES** 803-359-4578 **PERMISSIONS** 202-326-6765, permissions@aaas.org **AAAS Member Services** 202-326-6417 or http://membercentral.aaas.org/discounts

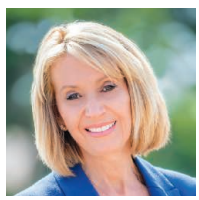
Science serves as a forum for discussion of important issues related to the advancement of science by publishing material on which a consensus has been reached as well as including the presentation of minority of conflicting points of view. Accordingly, all articles published in *Science*—including editorials, news and comment, and books reviews—are signed and reflect the individual views of the authors and not official points of view adopted by AAAS or the institutions with which the authors are affiliated.

INFORMATION FOR AUTHORS See pages 678 and 679 of the 6 February 2015 issue or access www.sciencemag.org/about/authors

Published by AAAS

EDITORIAL

Oceans and Earth's habitability



Marcia McNutt
Editor-in-Chief
Science *Journals*

On 8 June, the United Nations Educational, Scientific and Cultural Organization (UNESCO) celebrates World Oceans Day, a fitting occasion to remind ourselves of the essential role of the oceans in making Earth a habitable planet. We have had an official day of celebration for the oceans only since December 2008. In contrast, Earth Day has been celebrated every year since 1970. Conceived by U.S. Senator Gaylord Nelson in the aftermath of the 1969 Santa Barbara oil spill, Earth Day became a focus for the growing environmental movement (it became an international event in 1990) and the catalyst that led to the Clean Air, Clean Water, and Endangered Species Acts in the United States. Imagine what might be accomplished if World Oceans Day could similarly inspire actions for improving the state of the oceans worldwide.

"...World Oceans Day could... inspire actions for improving the state of the oceans worldwide."

Many environmental crises play out in the ocean in slow motion and are not currently addressed by the protections that are in place. For example, oceans absorb about 90% of the heat building up from the release of excess greenhouse gases. The system of Argo profiling floats indicates that the heat content of the upper 2000 meters of the ocean has increased by about 8×10^{22} joules over the past 10 years. The yearly increase in heat to the ocean is roughly equivalent to 100 times the average annual energy consumption of the United States (100 quadrillion BTU = 10^{20} joules). We have so much to learn about the microbiota in the upper ocean (see the *Tara* Oceans special section on p. 873), and the effect that this added heat will have on them is entirely unknown. It is likely to have deleterious impacts on fisheries already stressed from overharvesting. And yet, if it were not for the large amount of heat that the oceans absorb, the amount of global warming we would otherwise experience would be truly intolerable.

It is not just excess heat that the oceans absorb. As CO_2 is released to the atmosphere from the burning of fossil fuels, about a quarter is absorbed by the ocean, lowering its pH. Since the start of the Industrial Revolution, ocean acidity has increased by 30%, with negative repercussions for many organisms, including those that build their shells from calcium carbonate minerals. Such organisms are essential links in marine food webs and the foundation for very profitable fisheries. As the oceans become more saturated with CO_2, their ability to mitigate the buildup of CO_2 in the atmosphere by absorbing it will decrease, and greenhouse warming will accelerate.

The oceans help to moderate climate, keeping tropical latitudes cool and temperate latitudes warm through major circulation systems that transport large amounts of equatorial heat poleward. The ongoing warming could change ocean circulation in complex ways, a problem worth addressing at the UN Framework Convention on Climate Change (COP21) in December. World Oceans Day's focus on the ocean's role in the climate system will expand global awareness just ahead of this summit.

When scientists search for extraterrestrial worlds that might be habitable, they look for water and signs of an ocean. I find it ironic that in the most recent budget for the National Aeronautics and Space Administration, the U.S. Congress is willing to explore these distant worlds but slashes funding to monitor Earth, the one planet we know is suitable for life as we know it.

With every other breath you take this 8 June, take a moment to thank the ocean for supplying half of your oxygen and for all the other ways in which it makes Earth a habitable planet. It is time to start valuing the ocean and stop using it as a dump for waste heat, CO_2, sewage, pollutants, and other trash.

– Marcia McNutt

10.1126/science.aac5755

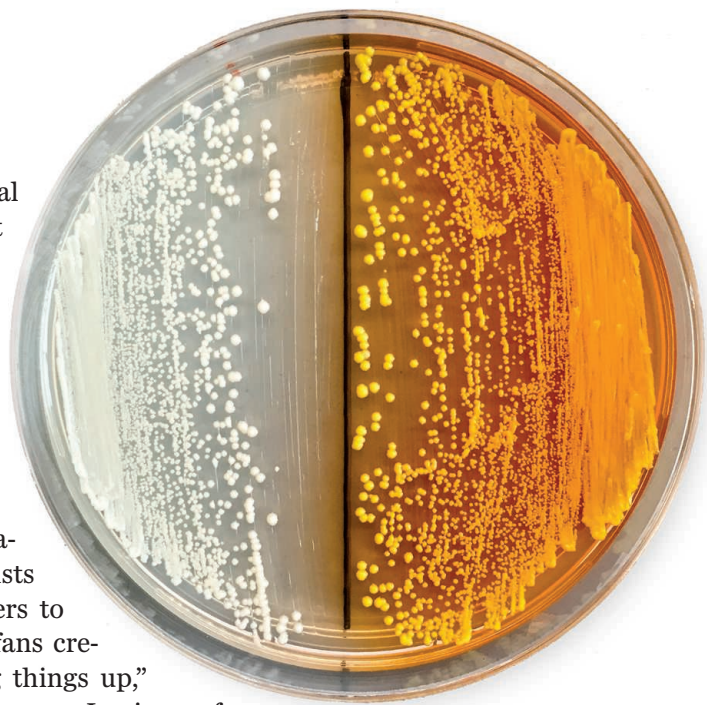
NEWS

70% Target set by Representative John Culberson (R–TX), chair of a panel that sets the National Science Foundation's budget, for spending on so-called core disciplines—excluding the geo and social sciences. Currently, those areas get 65%.

IN BRIEF

A rival for opium poppies?

Researchers are closing in on a long-standing goal of engineering a suite of genes into yeast that would allow the microbes to synthesize morphine, codeine, and other medicines harvested from opium poppies for thousands of years. In a paper published online this week in *Nature Chemical Biology*, scientists reported inserting an enzyme from sugar beets into yeast that carries out one of the few remaining steps needed to enable microbes to synthesize opiates. The work could lead to the cheap, easy production of widely used medicines with new capabilities and fewer side effects. But policy specialists worry that the new strains could allow narcotics dealers to convert sugar to morphine or heroin as easily as beer fans create homebrews. "There really is potential for screwing things up," says Kenneth Oye, a biotech policy expert at the Massachusetts Institute of Technology in Cambridge. In a *Nature* commentary this week, Oye and colleagues proposed regulations such as asking gene synthesis companies not to distribute genes needed to produce illicit compounds and engineering morphine-producing yeast strains with traceable genetic watermarks. http://scim.ag/yeastopiates



An enzyme engineered into yeast allows researchers to see which microbes are making L-Dopa (yellow), a key step in the pathway to making opiates.

PHOTO: WILLIAM DELOACHE

AROUND THE WORLD

New avenue for fusion research

WASHINGTON, D.C. | The Advanced Research Projects Agency-Energy, the Department of Energy's agency for blue-skies energy research, announced a slew of projects on 14 May that it hopes will break the logjam in fusion research, which has been trying to replicate the power source of the sun for more than 60 years without success. The Accelerating Low-cost Plasma Heating and Assembly program seeks a middle way between the two primary approaches: high-density laser fusion and low-density magnetic fusion. Under the program, nine projects will share $30 million to investigate whether plasma jets, ion beams, current pulses, high-pressure gas, and pneumatic pistons may be able to achieve the temperatures and pressures necessary to get hydrogen ions to fuse together, releasing energy. A lack of funds has hindered U.S. government labs from investigating such approaches, prompting a new breed of startups (*Science*, 25 July 2014, p. 370).

German scientists push for GM

BERLIN | Those who oppose genetically modified (GM) food usually advocate for labels on it, and those who support it usually see no need. But this week, a group of German scientists joined other GM proponents to launch a campaign to require labeling of food, feed, drugs, textiles, chemicals, and other products produced with the help of GM organisms. The petition to the German parliament is actually a gamble: The groups hope the new law will show Germans how widespread such products already are and that there is nothing to be afraid of. The petition also calls on the government to advocate for a similar law at the E.U. level. The text has the backing of several prominent scientists, including Nobel Prize winner Christiane Nüsslein-Volhard, as well as some politicians. If it receives more than 50,000 signatures in the next 4 weeks, the German parliament has to consider the proposal. http://scim.ag/_GMlabel

Tackling embryo gene editing

WASHINGTON, D.C. | Responding to an uproar over attempts to genetically modify human embryos, the U.S. National Academies is launching an international

initiative to discuss this ethically fraught area. Although genetically modifying the human germ line—eggs, sperm, or embryos—to create a baby has long been considered taboo, new gene-editing technologies such as CRISPR have heightened concerns that genetically modified babies are on the horizon. Whether even basic research in this area should move forward is hotly debated, particularly following an April report by a Chinese team describing its editing experiment on defective human embryos. A fall meeting by the National Academy of Sciences and National Academy of Medicine is intended to set the stage for a committee to begin working out guidelines.

NEWSMAKERS

Three Q's

Agriculturalist **Cary Fowler** was executive director of the Global Crop Diversity Trust from 2005 to 2012, helping create the Svalbard Global Seed Vault in Norway. This month, *Seeds of Time*, a documentary that chronicles Fowler's efforts to protect the genetic diversity of the world's crops, opens in New York and Los Angeles. Fowler discussed his work with *Science*.
http://scim.ag/FowlerQA

Q: When will Svalbard be complete?
A: There isn't an endgame. We have samples of 864,000 distinct crop populations. I guess we have upwards of 1.5 million samples around the world that could go in Svalbard. [So] you might be tempted to say we're more than halfway there, but that's not the way to look at it. It's not a numbers game. It's a diversity game.

Q: How would you rate the overall security of crop diversity today?
A: I'd rate the diversity that's in Svalbard at a 10 [safe as can be]. We've really put an end to extinction. For the genetic material that's not in Svalbard, the number is much lower. That depends on its location. It could be anywhere from 1 to 6 or 7.

Q: How do you hope scientists use Svalbard in the future?
A: I hope they never use Svalbard. It's an insurance policy. I do worry that while we have really big collections for the top 15 major crops, we're deficient in the rest. That doesn't bode well in an era of climate change, where we need to use that diversity to adapt our crops.

PHOTOS: (LEFT TO RIGHT) *SEEDS OF TIME*; © AGE FOTOSTOCK/ALAMY

Great tits know when caterpillars will be most plentiful.

Trees set birds' hatching schedule

While most expecting moms never quite know when they will give birth, great tits (*Parus major*) have their timing nailed down. Their eggs hatch right when nearby oak trees—those within 50 meters—produce leaves, says ornithologist Ben Sheldon of the University of Oxford in the United Kingdom. That leafing out triggers a 2-week explosion in the abundance of the winter moth caterpillars that munch on the leaves—and great tit parents depend on that caterpillar bonanza to feed their chicks. Researchers have shown that the birds' reproductive timing is shifting with global climate change. But for individual birds, the cues are local: They set their mating schedule according to when the trees they are likely to visit leaf out, Sheldon and his colleagues report in the July issue of *American Naturalist*, based on 45 years' worth of data on great tits living near the university. The researchers don't know what the birds are looking for, but some trees always leaf out early; others later, the researchers showed. And "the birds match that local effect," he says.

FINDINGS

Unraveling a day care–cancer link

Scientists have long noticed that children who went to day care early in life are less likely to develop the most common childhood cancer: acute lymphoblastic leukemia (ALL). Now, a study that unravels the molecular mechanism driving ALL may explain why early exposure to routine infections might boost the immune system and ultimately help protect against the disease. The immune system's B cells reprogram their DNA to recognize different infections through a sequence of enzymes. Researchers suspected that, in children with a genetic abnormality linked to ALL, repeated infections later in childhood could trigger unregulated mutations in the B cells, causing leukemia. The team took mouse B cells with the genetic flaw and subjected them to repeated "infections"—exposure to a molecule that triggers an immune response. All 14 mice injected with those B cells got leukemia and died, the team reported online this week in *Nature Immunology*.
http://scim.ag/daycarecanc

MICROBIOLOGY

Leaf bacteria fertilize trees, researchers claim

Free-living nitrogen fixers defy textbooks and could boost crop production

By **Elizabeth Pennisi**,
in Yosemite National Park, California

The fastest growing trees outside the tropics are poplars. Tall and slender, they can reach 30 meters in less than a decade despite the seemingly inhospitable ground they favor—burned areas and sandy riverbanks, for example. Sharon Doty says the credit goes to microbes in their leaves and other tissues. While the poplar's leaf cells are busy converting sunlight to energy, she says, bacteria between those cells are transforming nitrogen from the air into a form the tree needs to sustain this rapid growth.

That's a radical notion, because nitrogen fixation is generally thought to happen primarily in bacteria-rich nodules on the roots of legumes and a few other plants, and not in the treetops. "We are completely fighting dogma," says Doty, a plant microbiologist at the University of Washington, Seattle.

Earlier this month at the Fifth Annual Yosemite Symbiosis Workshop here, Doty bolstered her case. She reported the first direct evidence that poplars do get nitrogen from certain microbes, and she got support from Carolin Frank, an environmental microbiologist at the University of California (UC), Merced, who studies a different tree that thrives on poor soil. Frank reported that nitrogen fixation may also occur in the needles of limber pines, which grow on stony, high-elevation slopes in western North America.

Frank and Doty suspect that nitrogen-fixing leaf bacteria may be widespread, and, if transferred to crops, could help boost yields on marginal soil. Doty has found that a number of crops grow better when inoculated with the bacteria, and at the Yosemite meeting she reported the latest to benefit: rice. Other plant biologists, although far from convinced, are paying attention. "If there's an unrecognized set of nitrogen fixers in a wide number of [tree] species, that's a big deal," says Douglas Cook, a plant and microbial biologist at UC Davis.

The belief that significant nitrogen fixation takes place only in those bacteria-filled root nodules has been under strain since the 1990s, when researchers discovered nitrogen fixing in sugarcane, which doesn't have nodules. Since then, investigators have reported clues that bacteria called endophytes, which live inside plant tissues, provide nitrogen to their hosts. But, Cook contends, "the proper studies haven't been done, and they are not trivial."

He and others argue that the nitrogenase enzyme key to the process is too sensitive to oxygen to work in leaves. And even if the microbes are processing nitrogen in the air, "that doesn't mean that they are actually providing a host with any benefit," says Sharon Long, a researcher who studies nitrogen fixation at Stanford University in Palo Alto, California.

Doty has tried to answer all of those objections. She first began to suspect that nitrogen fixation might take place outside root nodules about 15 years ago, when she discovered her poplar cell cultures were full of bacteria related to known nitrogen fixers. She put the bacteria on media that lacked nitrogen, yet some thrived—apparently getting their own nitrogen from the air.

She's since documented that dozens of bacterial strains from poplar promote growth not only of poplar but also of rye, turfgrass, maize, cottonwood, tomato, and now, she reported, rice. Her greenhouse experiments show that rice seedlings dipped for 4 hours in a broth containing poplar endophytes wind up with the microbes throughout the plant body and grow taller, have more biomass, and sprout more tillers—which produce heads of grain—than untreated rice.

If Doty is right, a dose of the bacteria could be a boon to farmers. "Nitrogen is a huge constraint, particularly for farms in Africa," says Katherine Kahn, a plant biologist and program officer at the Bill & Melinda Gates Foundation in Seattle. Current remedies fall short: Fertilizer is costly and environment-damaging, adding nitrogen-fixing bacteria to the soil doesn't work well, and equipping crops with the genes needed to form nodules or to fix nitrogen them-

PHOTO: JEFFRY B. MITTON

Growing in harsh conditions, limber pines may get help from nitrogen-fixing bacteria in their needles.

selves is still a distant dream.

Skeptics note that some of the leaf-dwelling bacteria Doty has isolated make plant hormones, which could increase growth. But because Doty did these experiments in artificial soil lacking nitrogen, she argues that nitrogen supplied by the bacteria must be driving the growth. At the meeting, Doty's former technician, Andrew Sher, reported what she considers the strongest evidence yet. Sher put cuttings from wild poplars into flasks and exposed them to a heavier form of nitrogen than exists in air. Afterward, the same isotope turned up in the plant tissues, evidence that the bacteria had captured it and converted it to a usable nutrient, Doty says.

Frank converged on the same conclusion from a different starting point: a 2012 discovery that 30% to 80% of the microbes in limber pine needles were related to known nitrogen-fixing species. It struck her that these bacteria might explain a puzzle. In forests, foliage and soil contain more nitrogen than they should, given the known sources. Nitrogen-rich bedrock can explain some of the extra, but about 25% remains unaccounted for, says Benjamin Houlton, a global ecologist at UC Davis who specializes in the nitrogen cycle. "When you add up the numbers you come up short," he says. If nitrogen fixers were at work in leaves and needles, they might balance the books, Frank thought.

Still, she was initially skeptical. "I'd had a lot of doubt, lying awake at night," she recalls. But at the meeting, she described putting a limber pine twig with needles into a jar and replacing some of the vessel's air with acetylene. As microbes fix nitrogen, their nitrogenase enzymes convert acetylene into ethylene. The presence of ethylene at the end of the experiment told Frank that nitrogen-fixing microbes were at work, far from any root nodule.

Others are now cautiously embracing the idea. "There's a change in attitude, not from skepticism to believing but from skepticism to cautious questioning," says Gerald Tuskan, a plant geneticist at Oak Ridge National Laboratory in Tennessee. Tuskan and his colleagues have isolated about 3000 microbes from poplar, many of which are equipped with nitrogenase. Some sequester themselves in biofilms with oxygen-limited compartments, where nitrogenase could function even in the leaf's oxygen-rich environment.

Bit by bit, the case for treetop nitrogen fixation is building, Frank says. "I think we are converting people slowly, including ourselves." ■

IMAGE: ESA

EARTH SCIENCE

Alarm over a sinking delta

Rise and Fall project seeks ways to slow land subsidence in Vietnam's populous Mekong delta

By **Charlie Schmidt**, *in Soc Trang, Vietnam*

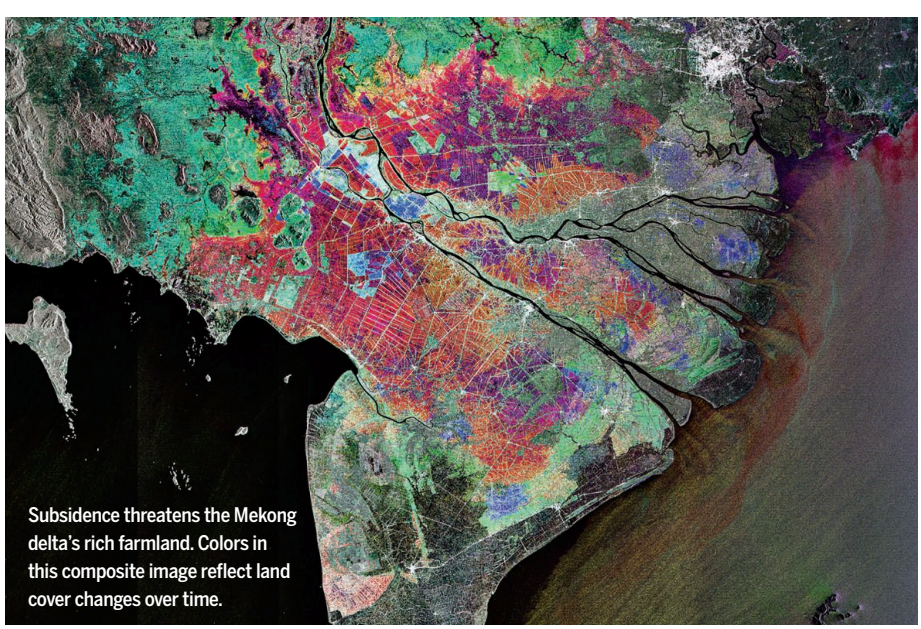
Leaning over a pond carved into the soft soils of the Mekong River delta, Ngwyen Khuong strains to lift a net of flapping shrimp. "We can harvest 4000 kilograms from a pond like this every 3 months," he says. But Khuong's booming shrimp business may be undermining the very land it occupies. Shrimp farmers in the delta are pumping prodigious amounts of ground water into their brackish ponds, causing water tables to drop, overlying sediments to compact, and the land to subside. The trends could expose the world's third largest river delta—home to some 20 million people—to flooding and other threats. "We face big problems if we have subsidence on one side and rising seas on the other," says environmental scientist Nguyen Hieu Trung of Vietnam's Can Tho University.

An alliance of Vietnamese and Dutch scientists is now trying to get ahead of the problem. They met here recently to launch the Rise and Fall project, a $1 million, 5-year effort to better understand what's driving Mekong delta subsidence and develop strategies to reverse it. "We know virtually nothing about what's beneath our feet," said geographer Philip Minderhoud, a co-leader of the project and doctoral candidate at Utrecht University in the Netherlands, during the 11 March gathering. "In many places the rates, causes, and future implications of subsidence remain an open question."

Although researchers have documented subsidence in other large river deltas, they only recently published the first hard evidence that the Mekong delta, which covers some 55,000 square kilometers and sits about 2 meters above sea level, is sinking. Ground- and satellite-based instruments have clocked average subsidence rates of 1 to 4.7 centimeters per year, a group led by hydrogeologist Laura Erban of Stanford University in Palo Alto, California, reported last year in *Environmental Research Letters*. In Ca Mau, a province on the delta's southern tip, the sinking reaches nearly 5 cm annually.

Among the culprits: levees that prevent sediment from spilling out of rivers and into the delta, and some 1 million wells drilled since the 1980s for drinking and agriculture. If groundwater depletion continues at present rates, researchers estimate, the delta could sink by nearly a meter by midcentury. Ca Mau alone has more than 100,000 wells, which have caused water tables to fall some 5 meters and allowed seawater to creep inland, making the well water increasingly salty.

At the meeting, the researchers began sharing what they know. The next step in the project, which is primarily funded by the Dutch Science Foundation, will be extensive



Subsidence threatens the Mekong delta's rich farmland. Colors in this composite image reflect land cover changes over time.

fieldwork, say project leaders Minderhoud and Pham Van Hung, director of the Center for Water Resources Technology for the South of Vietnam in Ho Chi Minh City. Geologists, for example, will map layers of sand, clay, and peat, which compact in different ways. Such data will be fed into modeling tools that will help researchers and policymakers understand how water use, development, and sea level rise could affect the fate of the delta.

One sensitive question is exactly how much of the subsidence is due to groundwater extraction—a main driver of delta economic growth. "People just say, 'Ground water is causing this,' but we have no data to prove it," says Bui Tran Vuong, the deputy director of the Division of Water Resources, Planning, and Investigation for South Vietnam in Ho Chi Minh City. Other factors are likely at play, says geologist Esther Stouthamer of Utrecht University. Urban infrastructure can squash poorly drained soils, and intruding salt water can weaken the chemical bonds between soil grains, making soils more likely to compress. Still, Stouthamer says, "ground water is probably the main driver" of subsidence.

In other nations, government efforts to limit groundwater use or switch to surface supplies have slowed or halted subsidence, but can require intrusive regulation and expensive infrastructure. Another option is to pump water back into the ground to raise the surface, a process called recharge. But the pumping tends to require a lot of energy, the water can escape through unseen cracks, and roads and buildings can "buckle as the land rises," says James Syvitski, an oceanographer at the University of Colorado, Boulder.

Syvitski is similarly skeptical of scenarios that envision the delta becoming an Asian version of Holland: a lowland protected from the sea by tall dikes. "Doing that for the Mekong coastline is cost-prohibitive," he believes. Others disagree. "Life on the future delta will be lived below sea level," predicts historian David Biggs, a Vietnam specialist at the University of California, Riverside. "But to make it work on the scale that we see in Holland will require a lot of education and democratic participation."

In the meantime, the delta confronts existential threats from abroad. Nations upstream along the Mekong are building dams expected to reduce the flow of sediments that build the delta, and sea level is rising. Still, many researchers are optimistic that such change can be managed. Projects like Rise and Fall are coming none too soon, Syvitski believes. "The Mekong delta," he says, "is at a tipping point." ■

Charlie Schmidt is a freelance writer in Portland, Maine.

BIOMEDICAL RESEARCH

Canadian registry to track thousands of pot smokers

Data could answer questions about safety, efficacy, and dosage

By **Lizzie Wade**, *in Montreal, Canada*

When a healthy looking man in his 70s walked into a sickle cell clinic in Kingston, Mark Ware sat up and took notice. A newly minted doctor, Ware saw many patients in chronic pain who often died young. The elderly Rastafarian seemed unscathed by the disease. "I asked him, 'What's your secret?' " says Ware, recalling an encounter that took place 15 years ago. "He leaned over, fixed me with his eyes, and said, 'Study the herb.' "

Ware is now doing so on a grand scale. A pain management researcher at McGill University Health Centre here, the native Jamaican directs the Quebec Cannabis Registry, a new, one-of-a-kind database that aims to gather information on every patient prescribed marijuana in the province over the next 10 years—thousands in all. By collecting data on symptoms, dosage, improvement, and side effects, the registry, launched on 11 May and funded by a grant from the nonprofit Canadian Consortium for the Investigation of Cannabinoids, aims to fill gaps in knowledge about the efficacy and safety of medical marijuana. It's a "wonderful step in the right direction" for "legitimizing some of the medical uses of cannabis," says Raul Gonzalez, a psychologist at Florida International University in Miami who studies the cognitive effects of cannabis use in HIV/AIDS patients.

Most drugs go through years of rigorous clinical trials before they are prescribed. That's not the case for marijuana. Even as more and more states and countries legalize pot for medical purposes, clinical trials of smoked cannabis remain rare. "Decisions [about medical marijuana] are being made at the ballot box instead of in the laboratories," Gonzalez says.

Scientists anticipate a trove of data on Canada's medical marijuana use.

PHOTO: © BLAIR GABLE/REUTERS/CORBIS

Few doubt that the drug can relieve certain symptoms. It eases neuropathic pain, reduces spasticity in people with multiple sclerosis, and improves appetite and weight gain in chemotherapy patients and those with wasting conditions, according to psychiatrist Igor Grant, director of the Center for Medicinal Cannabis Research at the University of California, San Diego. However, doctors have almost no guidance on recommended dosages or possible side effects. "If we knew what we were prescribing more accurately, we'd be a lot more willing to work with it," says Barbara Koppel, a neurologist at the Metropolitan Hospital Center in New York City.

Amassing and analyzing a large volume of patient data could answer long-standing questions, Ware says. Canada could have done this sooner: In the first 15 years of its medical marijuana program, 40,000 people were authorized to smoke the plant. But "we didn't learn anything from that process—about who they were, why they used it, how they used it, how much—nothing," Ware says. "We don't want to be in the same position 10 years from now." Through 2025, the Quebec registry aims to collect anonymous data from 3000 patients, each of whom will be tracked for 4 years to probe for rare side effects.

Large clinical trials would help bring medical marijuana out of the shadows. "Without well-controlled empirical studies, we're still going to be left scratching our heads about whether [medical marijuana] really works," Gonzalez says. Funding them is a challenge. Drug companies show scant interest in dried, smoked cannabis, Ware says, because it "may not have long-term payback." In the meantime, collecting vital data from users can't wait, he says. Marijuana "is part of our society now," Ware says, "and we need to have a means of talking to our patients about it." ■

This Romanian cave yielded a modern human with Neandertal blood.

HUMAN EVOLUTION

Ancient DNA pinpoints Paleolithic liaison in Europe

Romanian fossil was the great-great-great-grandson of a Neandertal—but an evolutionary dead end

By **Ann Gibbons**

To Erik Trinkaus, the jaw of the oldest modern human found in Europe has always looked strange. Its huge wisdom teeth and hefty, buttressed lower jaw reminded him of Neandertals, and he argued that this fossil, 37,000 to 42,000 years old, was the product of generations of mixing between modern humans and our extinct cousins. "It wasn't a popular idea," admits Trinkaus, a paleoanthropologist at Washington University in St. Louis. Other paleoanthropologists insisted that the young man whose remains were found in 2002 in Peştera cu Oase cave in Romania was just a chunky example of our own species.

Now, 15 years later, Trinkaus has been vindicated by ancient DNA. The young Oase man inherited as much as one-tenth of his DNA from a Neandertal ancestor, and that ancestor lived only 200 years or so previously, according to a talk this month at Cold Spring Harbor Laboratory in New York. "One of Oase's ancestors—its great-great-great-grandparent—is Neandertal," reported Qiaomei Fu, a geneticist at the Chinese Academy of Sciences–Max Planck Society Joint Laboratory for Human Evolution in Beijing and a postdoc in the lab of population geneticist David Reich at Harvard Medical School. The finding is "important as the first direct evidence of a very recent admixture event in Europe," says population geneticist Laurent Excoffier of the University of Bern.

Europe just after the arrival of modern humans has long seemed a likely setting for such close encounters, given that Neandertals and modern humans overlapped there about 45,000 to 39,000 years ago. But until now, ancient DNA pointed to a different time and place for such a liaison. By sequencing the genomes of fossil Neandertals and comparing them with today's human genomes, paleogeneticists had found that living Europeans and Asians—but not Africans—have inherited just 1% to 4% of their DNA from Neandertals. DNA from fossils of two modern humans from what is now Russia also suggested that their Neandertal heritage was faint (see http://scim.ag/RussDNA). So researchers proposed that modern humans and Neandertals had rare and relatively early encounters, perhaps in the Middle East, when moderns swept out of Africa 60,000 to 50,000 years ago.

This robust jawbone is partly Neandertal.

The DNA from Oase 1, a lower jaw without a skull, complicates that picture, Fu reported at the Biology of Genomes meeting. Working in a team led by paleogeneticist Svante Pääbo of the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, she and her colleagues captured 2.2 million base pairs of the fossil's DNA. Then, they sequenced 78,055 locations where the genomes of Neandertals and modern humans are known to differ. They found that the Oase man had far more Neandertal DNA—composing 4.8% to 11.3% of his genome—than either the ancient modern humans from Russia or living Europeans and Asians, Fu said.

What's more, the young man had inherited the Neandertal DNA in "large chunks," including several segments more than 50 million base pairs long; one chunk spanned half the length of chromosome 12. Those unbroken stretches of Neandertal DNA suggest that the interbreeding must have been just four to six generations back. If the mixing had been more ancient, the long DNA segments would have been broken up by the reshuffling of chromosomes that takes place every generation. "This is quite amazing," Fu said in her talk. "We're quite excited about that."

If modern humans and Neandertals had several successful matings, why do living humans' genomes record only the earlier event? An answer emerged when Fu traced how this Oase man and other early modern fossils link to later peoples. One of the early modern fossils from Russia, a 36,000- to 39,000-year-old arm bone known as Kostenki 14, is genetically similar to present-day Europeans. In contrast, the DNA of the Oase fossil, although it is from Europe, more closely resembles ancient Asians than Kostenki or living Europeans, Fu reported. The team concluded that the Oase man himself was an evolutionary dead end, who did not pass his DNA along to living Europeans. Members of the team have declined to comment further, because their report is in press.

Fu's was "an impressive talk," says population geneticist Andrew Clark of Cornell University, and suggests that "there were many interbreeding events." But the recent mixing surprises some. "I thought that interbreeding would be a lot less likely at 40,000 years when there were so few Neandertals left," says Chris Stringer of the Natural History Museum in London. "I have to admit Erik has been proved right."

Trinkaus wasn't surprised when he saw a copy of the manuscript. "It confirms things a bunch of us have been saying for a long time." ■

PHOTOS: (TOP TO BOTTOM) OASE PROJECT; EMIL RACOVIŢĂ/MIRCEA GHERASE

Research commissioner Carlos Moedas, flanked by Nobelists Paul Nurse (*left*) and Jules Hoffmann (*right*), at last week's announcement.

EUROPE

E.U. commission promises to listen to scientists

Panel of seven top scientists to act as watchdog of new advice system

By **Tania Rabesandratana**

The European Commission extended an olive branch to the scientific community on 13 May. Surrounded by six Nobel laureates, commission President Jean-Claude Juncker announced his long-awaited plan to restructure the commission's scientific advice process—and tried to reassure scientists that policymakers in Brussels will take their views seriously. Under the commission's new Science Advice Mechanism, a high-level group of seven scientists will channel the input of national academies and learned societies to give the commission the best scientific advice.

The announcement ends months of suspense. When Juncker took office last November, he didn't renew the position of chief scientific adviser (CSA), then held by Scottish biologist Anne Glover. But he didn't offer an alternative, either—which some scientists, especially in the United Kingdom, took as a sign of disregard for science (*Science*, 21 November 2014, p. 904). Although last week's announcement provided critics with some reassurance, many details remain to be worked out, including how the high-level group will operate effectively. "[C]ommittees in general are at risk of being conservative, reaching conclusions that no one member stands behind and consensus that doesn't really exist," the British group Sense About Science wrote in a lukewarm reaction.

In relying on a collective rather than a single person, the new structure is more suited to the commission's culture of consensus, says Jerzy Langer, a physicist and former deputy science minister in Poland who's familiar with the intricacies of E.U. policymaking. While CSAs are a fixture in the United Kingdom and the United States, most European countries have never had them, Langer points out. "The commission by definition is a collective body, which must consult member states. The CSA is alien," he says. Glover expressed "strong opinions"—for instance emphasizing the safety of genetically modified crops—and that was "uncomfortable for the commission," adds Sofie Vanthournout, head of the Brussels office of the European Academies Science Advisory Council (EASAC).

Unlike Glover, the new group won't be employed by the commission and thus will be independent, research commissioner Carlos Moedas said last week. It will also have better support: The commission will assign about 25 people in Brussels to run the new advice mechanism. Robert-Jan Smits, the commission's director-general for research, was reported as saying last week that the group is not expected to provide direct advice but rather to act as a "watchdog" to ensure that the commission draws on adequate evidence. "National academies are ideally placed to provide such advice," a research representative for the commission says, "but the idea is to cast the web as widely as possible and engage the broader scientific community when needed."

The commission will put €6 million on the table next year to help EASAC and four other European networks—representing 90 academies and learned societies—work together. According to the commission's draft call for proposals, seen by *Science*, academies should use that money to "animate public debate," produce joint events and policy papers, and set up a "working mechanism" to provide advice efficiently and fast. That will not be easy, Vanthournout says: Developing interdisciplinary, pan-European recommendations means aligning a host of national procedures for peer reviews and endorsements, she says. "We've never really done it because it [takes] extra resources."

Langer says the commission's insistence on involving academies is mostly a show of "courtesy." Academy members are eminent scientists, but "they are often over 80 years old; they are not decision-makers," he says. And "in contrast to the United States, the academic scene in Europe is extremely dispersed" across countries and disciplines; gathering input from it will be a lengthy, convoluted affair. That patchwork makes the new high-level group a "recipe for future problems," wrote Roger Pielke of the Center for Science and Technology Policy Research at the University of Colorado, Boulder, on his blog last week.

The advisory panel's members will be recruited by a three-strong "identification committee." Corporate Europe Observatory, an organization that had called for the CSA's abolition because it deemed the role opaque and vulnerable to industry influence, recommends that panel members should be only "active scientists" with recent peer-reviewed publications. Regardless of its composition, "the committee will have more means and confidence than [Glover] had," Vanthournout says—but getting the details of the system right will be complex. "We will have to learn by doing it." ■

PHOTO: EUROPEAN COMMISSION

REGENERATIVE MEDICINE

'Rejuvenating' protein doubted

Factor reported to explain how young blood restores muscle has opposite effect in another lab

By **Jocelyn Kaiser**

It was a mind-boggling observation. Hook up the circulatory systems of a young mouse and an old one, and the elderly animal seems to be rejuvenated. Since 2005, a handful of research labs have been hotly pursuing the molecules responsible for this effect, first found in the 1950s, hoping to harness them to slow or reverse aging in people. One in particular stood out: a protein found in young blood known as GDF11. In several high-profile papers, two last year in *Science*, a Harvard University team reported that the protein declines in older animals, and that replacing it rebuilds muscles, the brain, and the heart. But work described this week by a team at the Novartis Institutes for BioMedical Research in Cambridge, Massachusetts, challenges GDF11's rejuvenating powers.

The Novartis group does not question that young blood renews old mice. But they say the Harvard group's explanation is wrong. Their paper, in *Cell Metabolism*, casts doubt on the assays used in the earlier research and suggests that GDF11 actually inhibits muscle regeneration. "The whole premise is incorrect," says Michael Rudnicki of the Ottawa Hospital Research Institute, who co-wrote a commentary accompanying the paper. Others are more cautious, but agree that the new work undermines part of the original GDF11 claim. "GDF11 does not go down with age," says Thomas Rando, a biologist at Stanford University in Palo Alto, California.

Harvard stem cell biologist Amy Wagers, who led much of the original work, says the Novartis data on GDF11 levels are not persuasive. "We remain convinced that at least one form of GDF11 declines in blood with age and that maintaining GDF11 levels in an appropriate physiological range is essential for muscle health," she says.

Wagers began exploring the many ways in which joining the circulatory systems of mice—a procedure known as parabiosis—affects aging as a postdoc working with Rando and others (*Science*, 12 September 2014, p. 1234). In 2013, her group, with cardiologist Richard Lee's lab at Brigham and Women's Hospital in Boston, reported in *Cell* that levels of GDF11 in the blood fell as mice aged and that, like young-old parabiosis, restoring GDF11 through injections partially reversed age-related thickening of the heart. In *Science* last year, she and collaborators, including Lee and Harvard neuroscientist Lee Rubin, reported that GDF11 also nourished blood vessel and neuron growth in old mice's brains, improving the animals' sense of smell. In a second *Science* paper, Wagers and Lee reported that GDF11 spurred healing from a muscle injury in older mice. Aged mice receiving GDF11 did better on strength and running tests.

Some experts were flummoxed by the muscle paper, because GDF11 is a close cousin of myostatin, a well-studied protein that controls muscle growth. Animals and people lacking myostatin develop huge, bulging muscles; too much of it hinders muscle regeneration. How, then, could a very similar protein have the opposite effect?

Among the skeptics was David Glass of the Novartis center, who helped develop a myostatin-blocking drug for muscular atrophy. When his group tested GDF11 levels in rats with both of the assays Wagers had used, a proteomics assay and a commercial antibody, they could not distinguish between GDF11 and myostatin. Using more specific tests, they found that GDF11 levels actually trend upward with age in rat and human blood and that GDF11 mRNA levels rise in rat muscle with age.

The Novartis group also tested GDF11's effects on muscle regeneration. When they treated a young mouse with GDF11 and damaged its leg muscle with snake venom toxin, a common experiment, regeneration was impaired. "The bottom line is that [GDF11] seems to be harmful to muscle," Glass says.

Wagers sticks by her data, noting that her group's *Science* paper also found a drop in GDF11 with age using a different antibody that distinguished GDF11 from myostatin. And she says the Glass team's injury experiment cannot be compared to hers because they used young animals and a dose of GDF11 three times higher. (Glass did this in part because he did not see any effect in old mice at the dose Wagers used.) The signaling pathway in which GDF11 lies "is notoriously dose-sensitive," and low and high doses can have opposite effects, she says. Moreover, she says, the Novartis team's muscle regeneration test was not comparable to hers—the Harvard team made the injury by freezing tissue, which is less likely than a toxin to kill muscle stem cells needed for regeneration.

Wagers says new data from her group will show that "there is a very compelling biological explanation for the apparent discrepancies." One of her collaborators, Lee, agrees. But Rubin is more cautious: "Obviously, this report has to be taken seriously." Although the Novartis result does not challenge a second claimed benefit of GDF11, to the brain, "we're designing a series of experiments to convince ourselves that what we see in the brain is real," says Rubin, who led that study.

Others say that even if the new finding is correct, it may not contradict at least some of the benefits Wagers and others reported, says molecular biologist Se-Jin Lee of Johns Hopkins University in Baltimore, Maryland, who studies myostatin. He notes that GDF11's effects in the body are likely complex. "There's still a lot to be sorted out." ■

Young blood conundrum

The GDF11 protein may not explain how linking the blood of old and young mice renews tissues.

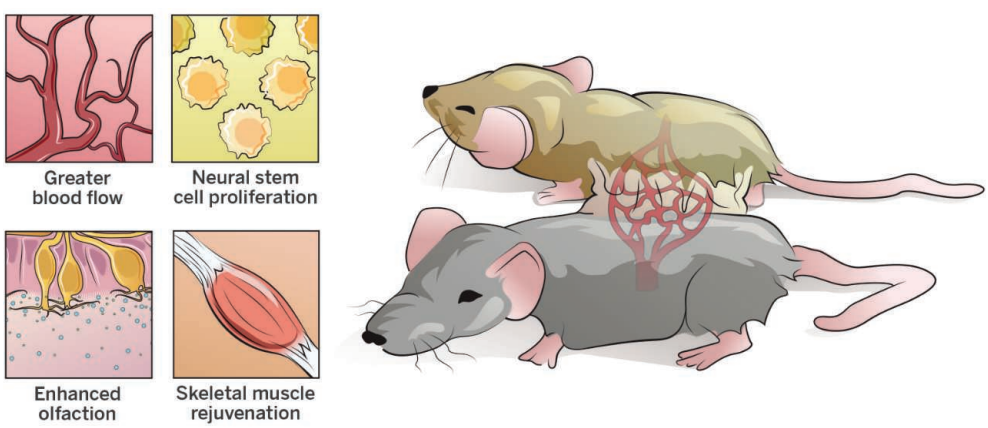


ILLUSTRATION: V. ALTOUNIAN/*SCIENCE*

FEATURES

THE DRUG PUSH

As fears of drug-resistant bacteria loom, governments try to coax companies back to the field

By **Kelly Servick**

ILLUSTRATION: DAVIDE BONAZZI

This past January, microbiologists Kim Lewis and Slava Epstein reported the discovery of teixobactin, a compound that in lab dishes kills several antibiotic-resistant strains of bacteria. Media outlets heralded the discovery, announced in *Nature*, as a new solution to the growing problem of antimicrobial resistance. A White House press release mentioned teixobactin, which Lewis and Epstein, both of Northeastern University in Boston, had isolated from soil bacteria, as the "kind of innovative research" it aims to promote with a $1.2 billion antibiotics budget initiative. And Lewis and Epstein were repeatedly asked: "When will this be in the clinic?"

Now, after years of encouraging wild, hard-to-culture microbes to fill a lab dish so he could harvest their chemical weapons, Lewis and the company he co-founded, NovoBiotic Pharmaceuticals, must engage in a different kind of coaxing. "In order to go into the clinic, we either need major investment or a big pharma partnership," he says. Someone has to bankroll studies that can turn their natural compound or a derivative of it into something that is soluble, potent, and likely to be safe—ready to try out in people.

"Those kinds of funding are really hard to come by in academia, not just for antibiotics," says June Lee, director of early translational research at the University of California, San Francisco's Clinical and Translational Science Institute, "[but] in antibiotics, you're less likely to find partners who are willing to invest that early on. … There just isn't a lot of money going into antibiotics."

That may seem counterintuitive, given recent projections of what will happen if harmful microbes continue to evolve resistance to our current drugs. A particularly ominous review commissioned by U.K. Prime Minister David Cameron lays out a worst-case scenario of 10 million deaths per year due to antimicrobial resistance by 2050. But the economics are stacked against new antibiotics. They must compete with a variety of cheap generic ones that (for now) still work for most infections. The short course of a typical antibiotic treatment makes it harder for drugmakers to turn a profit. And because using an antibiotic increases the selective pressure on bacteria to evolve resistance, doctors typically reserve newly approved treatments for the few cases where everything else has failed.

Today, only a handful of large pharmaceutical companies are willing to play those odds, and a slew of startups and academics are competing for the attention of skeptical investors. "For years, we've been starving the whole bacteria side of R&D," says Kevin Outterson, a health law professor at Boston University. As a result, "lots of ideas, both good and bad, just don't get followed up."

Recently, though, those watching the field are heralding new signs of life. Their excitement centers mostly on signals from industry superpowers. In December, pharmaceutical giant Merck announced that it would pay $8.4 billion to acquire Cubist Pharmaceuticals, a company focused on de-

veloping drugs for serious infections. A few other large firms, including Roche and Actavis (soon to be Allergan), are also building up their antibiotics programs.

Meanwhile, the United States and the European Union are discussing policies to make antibiotic development more attractive to companies. A U.K. report released last week calls for the founding of a global organization that would make multibillion-dollar lump-sum payments to firms that manage to introduce a new drug. Governments are also taking a more direct role in funding and overseeing antibiotic projects than ever before, fearing that resistant infections are evolving faster than our knowledge of how to kill them. "For the past 7 decades, we've known that this is a problem," Outterson says. "The ability to act and the willingness to act, I think, are strongest now."

BUT LEWIS AND OTHER RESEARCHERS with potential new antibiotics face an industry still deeply skeptical that developing such drugs can be profitable. That caution is fueled by recent scientific and financial disappointments.

In the late 1990s, researchers hoped that the growing field of genomics, combined with the screening of much larger chemical libraries, would help identify new antibiotics. Many companies tried sequencing bacterial DNA, then searching their libraries for compounds that could inhibit the products of key bacterial genes. But they came up empty-handed. Part of the problem was that these libraries excluded natural products isolated from plants and soil, which had been rich sources of antibiotics in the past, but are harder to work with and more expensive to manufacture, says Gail Cassell, a visiting scholar at Harvard Medical School in Boston who was vice president of infectious diseases at Eli Lilly when the company dropped its antibiotics research in 2002.

> ***"For years, we've been starving the whole bacteria side of R&D."***
>
> **Kevin Outterson**, Boston University

In place of antibiotics, industry pursued highly profitable drugs for chronic conditions such as heart disease and high blood pressure. As Cassell puts it, "This was the age of the blockbuster." Meanwhile, the U.S. Food and Drug Administration (FDA), spurred in part by safety issues with the already-approved antibiotic telithromycin, moved to tighten the requirements for any new antibiotic to win approval. Roche, Sanofi, Pfizer, Johnson & Johnson, Bristol-Myers Squibb, and Wyeth all joined Lilly in abandoning the field.

Today, the signs of a turnaround are ambiguous. "You'll see in a lot of newspaper articles … that pharma may be getting back in," says Alan Carr, a biotechnology analyst at Needham & Co. in New York City. "That's still somewhat questionable." AstraZeneca, which had continued to develop antibiotics after many companies bailed, announced earlier this year that it will spin out its anti-infectives projects into a separate company. And shortly after Merck's eye-catching purchase of Cubist, the pharma giant revealed that it would lay off 120 of the biotech's researchers and close its early-stage research and development arm.

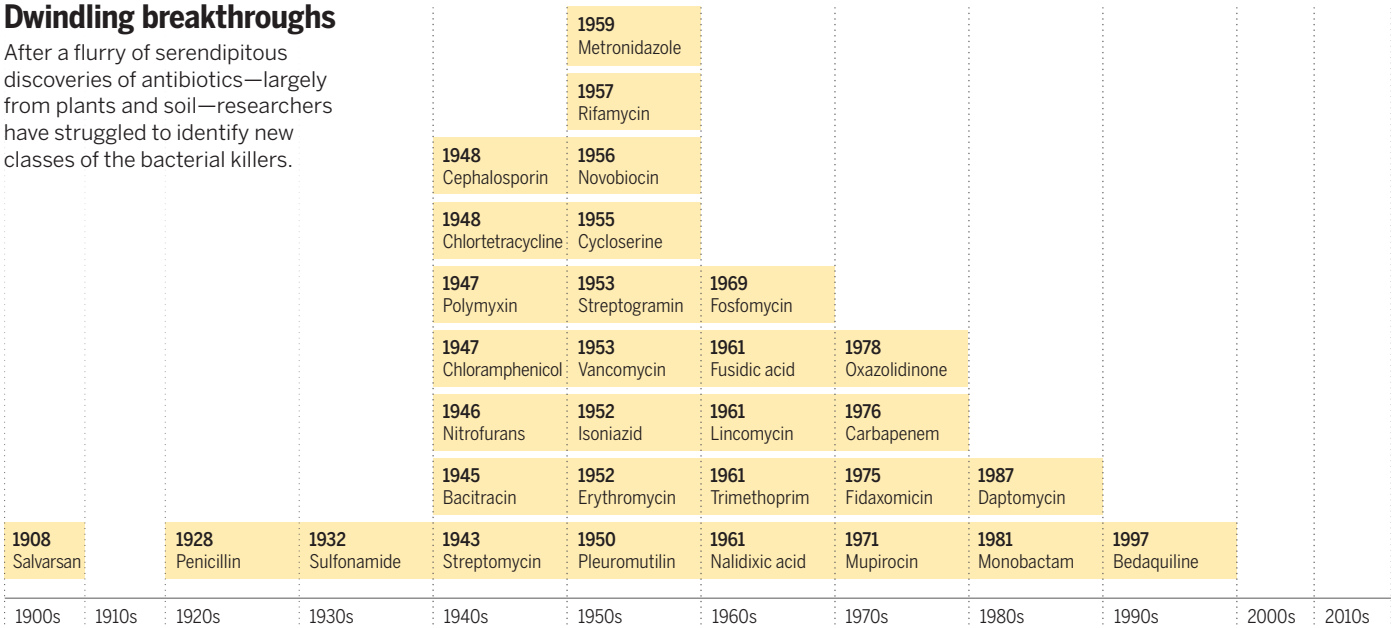
"There's a big fascination with these large pharmaceutical companies, but they are not the drivers of innovation," says Ramanan Laxminarayan, an economist who directs the Center for Disease Dynamics, Economics & Policy in Washington, D.C. "The model is, let the little guys come up with it, and then the big guys can eat them." Many hope that renewed involvement by big pharma in antibiotic development will bring in deeper pockets to fund trials, broader drug development expertise, and more influence with policymakers and regulators. But Laxminarayan is adamant that smaller companies can—and should—bring drugs to market themselves.

One company making a go of it is Tetraphase Pharmaceuticals, a spinout of Andrew Myers's Harvard University chemistry lab. Instead of hunting for new antibiotics in nature, Myers builds known natural antibiotic compounds from scratch, using cheap industrial chemicals. The approach lets him tweak their structures to thwart resistance to the original drugs. "The power of the approach is indisputable," he says.

In 2005, he and colleagues hit on a synthetic route to making a class of broad-spectrum antibiotics known as tetracyclines, the first of which was isolated from a soil bacterium in 1948. These compounds act on Gram-negative bacteria—microbes

Dwindling breakthroughs

After a flurry of serendipitous discoveries of antibiotics—largely from plants and soil—researchers have struggled to identify new classes of the bacterial killers.



with hard-to-penetrate outer membranes that are increasingly becoming resistant to available treatments. With a potentially valuable new antibiotic in hand, he faced a decision: Form a company to develop the drug, or license his discovery to a large pharma. Despite interest from "what conventional wisdom would call an outstanding suitor," Myers says he was reluctant to sign away the work to a large pharma, which might jettison the project if it hit a snag. Instead, he founded Tetraphase. (The suitor, he says, abandoned the field of antibiotics soon thereafter.)

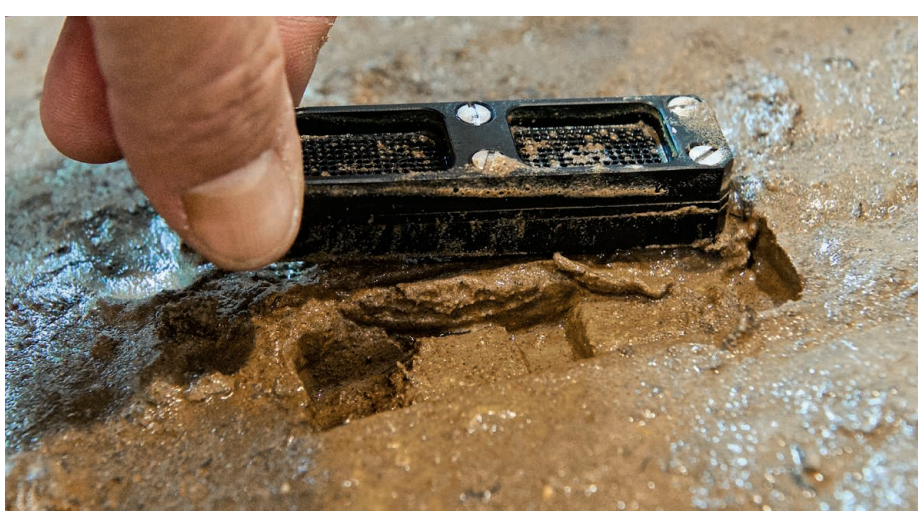
Now that Tetraphase has brought one of the new tetracycline variations into phase III clinical trials, Myers wants to repeat the feat with another class of antibiotics called macrolides. But this time around, despite a flood of investor money going into many other biomedical sectors, he found it harder to drum up enthusiasm from venture capitalists. "After the third meeting, we had one VC turn to us and say, 'Well, you know, antibiotics aren't valued in the marketplace.' " Another, whom Myers describes as "a fairly famous young wunderkind," wasn't as patient. "We had just made introductions and then he began to ridicule us, [saying] 'You antibiotics people don't even think about how to make money.' "

Yet Myers eventually did pull together a consortium of investors, and Macrolide Pharmaceuticals launched in March with $22 million in funding. Its support comes from a young venture firm called Gurnet Point Capital and, somewhat ironically, the investment arms of three large pharmas: SR One, the venture capital outfit for GlaxoSmithKline; Novartis Venture Fund; and Roche Ventures.

Roche in particular seems to have renewed its commitment to antibiotic development. The company was among the first big players to flee the field, in 1999, but in 2013, the company began shopping around for promising antibiotic projects to beef up its programs addressing "future unmet medical needs," says Janet Hammond, head of infectious diseases discovery in Roche's Pharma Research & Early Development team in Basel, Switzerland. And unlike large companies that focus exclusively on late-stage projects, Roche plans to license preclinical antibiotic projects and develop them in-house.

A small firm called Spero Therapeutics caught Hammond's eye, and last year, Roche generated media buzz when it chipped in an undisclosed amount for the company's preclinical research, in exchange for the option to buy it down the road. Hammond says that Spero has "a completely novel approach" that "com[es] at bacteria from an unexpected angle."

Microbiologist Laurence Rahme, whose lab at Harvard Medical School produced the idea behind Spero, is tickled when people call it "innovative" or "novel." She is pursuing compounds that interfere with how bacteria signal one another to produce virulence factors—molecules that help them attack and colonize a host. For years, she says, "nobody has been paying attention." But Rahme attracted the attention of her school's portfolio managers in 2012, and they connected her with Ankit Mahadevia, an entrepreneur at the venture capital firm



Soil chambers that can grow previously uncultivable soil bacteria have revealed new potential antibiotics.

CREDITS: (DATA) SILVER *ET AL*., *CLIN MICROBIOL REV*. JAN 2011; (PHOTO) SLAVA EPSTEIN

Atlas Venture in Cambridge, Massachusetts, who would become the CEO of Spero.

To Mahadevia, antibiotic development is a delicate bud ready to burst into bloom, nurtured by what he calls "a regulatory renaissance." A key reason he decided to build Spero was a 2012 U.S. law known as Generating Antibiotic Incentives Now, which gives drugs designated as "qualified infectious disease products" a faster review process at FDA and an additional 5 years of marketing exclusivity once they are approved. Another idea under discussion in the House of Representatives would allow FDA to approve drugs for rare, life-threatening infections based on smaller clinical trials than normally required for an antibiotic meant for the masses.

Such changes could be a particular boon to companies such as Spero, which is now focused on making a narrow-spectrum drug to treat *Pseudomonas aeruginosa*—a major cause of hospital-acquired blood and lung infections that is particularly common in lungs of people with cystic fibrosis. Narrow-spectrum antibiotics offer a company a smaller pool of patients, meaning it's harder to recruit for large clinical trials and harder to make back the cost of development before a company's marketing exclusivity period runs out. But they are appealing from a scientific perspective because they are less likely to exert selective pressure on other microbes, fostering the spread of resistance genes. Hammond says the regulatory changes under way make it "feasible to contemplate" developing a drug aimed at a single, high-priority pathogen.

But to make a business case for antibiotics, companies will also need confidence that they will be able to charge more for antibiotics than they have in the past, say many in the industry. A course of doxycycline, a commonly prescribed broad-spectrum tetracycline, averages less than $20 in the United States. A newer antibiotic called daptomycin, which is among the most expensive on the market, can cost as much as $1800. Meanwhile, Sovaldi, a new treatment for the hepatitis C virus, runs $84,000 a course.

"The other renaissance coming is going to be the reimbursement renaissance," Mahadevia declares. House lawmakers are now considering a bill that would increase levels of Medicare reimbursement for newer antibiotics. And a new E.U.-funded consortium involving industry and government is mulling a more dramatic step: separating a drugmaker's revenue from the number of pills prescribed. Governments would reward the maker of a drug for "the mere existence of it in the pharmacy, ready to go, not expired, because … when you need it, you need it right away," explains John Rex, senior vice president of infection for global medicine development at AstraZeneca in Waltham, Massachusetts, who helped set up the consortium.

That "de-linkage" approach last week gained the support of a U.K. government-appointed commission, chaired by former Goldman Sachs economist Jim O'Neill. The commission's new report on how to refill the antibiotics pipeline suggests a "single global body," whose member countries would pay a company between $2 billion and $3 billion for the rights to sell a new antibiotic and carefully manage its supply.

These discussions are still in the early stages, and many are skeptical that the approach would gain support in the United States. But even the conversations are enough to inspire confidence in some. "We see the tea leaves turning," Mahadevia says. "There's some folks that are waiting on the sidelines until we get an appropriate reimbursement picture. We see it. It's not explicit yet, but we're hoping and planning that it will be."

GOVERNMENTS ARE ALSO trying to supply a final element missing from the antibiotics field: drug development experience. With the departure of big pharma, "all the expertise that they had before they got out is long gone," says David Shlaes, a retired consultant specializing in antibiotic discovery and development, based in Stonington, Connecticut. Many smaller companies and academic labs don't have the knowledge or resources to optimize potential drugs for clinical trials, he says.

One U.S. program aims to inject drug development knowledge—along with a large chunk of cash—into new antibiotic projects. The Broad Spectrum Antimicrobials Program at the U.S. Biomedical Advanced Research and Development Authority (BARDA) hands out 5-year contracts of $50 million to $85 million for clinical-stage research and offers the recipients access to its team of drug development experts. In an unusual move, it made a $200 million agreement with GlaxoSmithKline in 2013 to fund a broad portfolio of projects, some still in preclinical stages. BARDA hopes to expand the model, provided that the president's proposed antibiotics budget initiative is funded, says program head Joseph Larsen, who is based in Arlington, Virginia. Other companies dipping a toe into antibiotic research have expressed interest in a similar partnership, he adds.

Additional initiatives are also trying to kick-start antibiotic development projects. The European Gram-negative Antibacterial Engine (ENABLE), another arm of the partnership between the European Union and the European pharmaceutical industry, has assembled a team of 32 companies and academic institutions and given them €85 million to bring one drug candidate for Gram-negative infections through a phase I clinical trial by 2019. "We're essentially a virtual pharmaceutical company," says Diarmaid Hughes, a microbiologist at Uppsala University in Sweden, which manages ENABLE. Small companies and academic labs can submit drug candidates, and if the experts are interested, ENABLE will pay for and help manage their development.

Hughes concedes that asking taxpayers to fund drug development by companies is bound to draw some critics. "At one level, it maybe smacks of desperation, you know—if companies won't do it themselves, let's pay them to do it." But he also sees long-term value in exposing academics to the realities of drug development. Several of the applicants, he says, have had strong chemistry background, but no knowledge of microbiology or how to test for resistance to the drugs they're developing.

Like the small cohort of investors and companies taking a risk on antibiotics, Hughes and his ENABLE colleagues are playing the long game, hoping the market will be friendlier by the time their projects reach expensive clinical trials. If not, Hughes says the work will at least help create a stockpile of potential antibiotics. "If a project is killed for economic reasons, it can just be left in a freezer," he says. "You could imagine this like discovering an oil field in the deep ocean. It may not be economical to develop it, but if the price of oil goes up, you know where it is." ■

Antibiotic stats

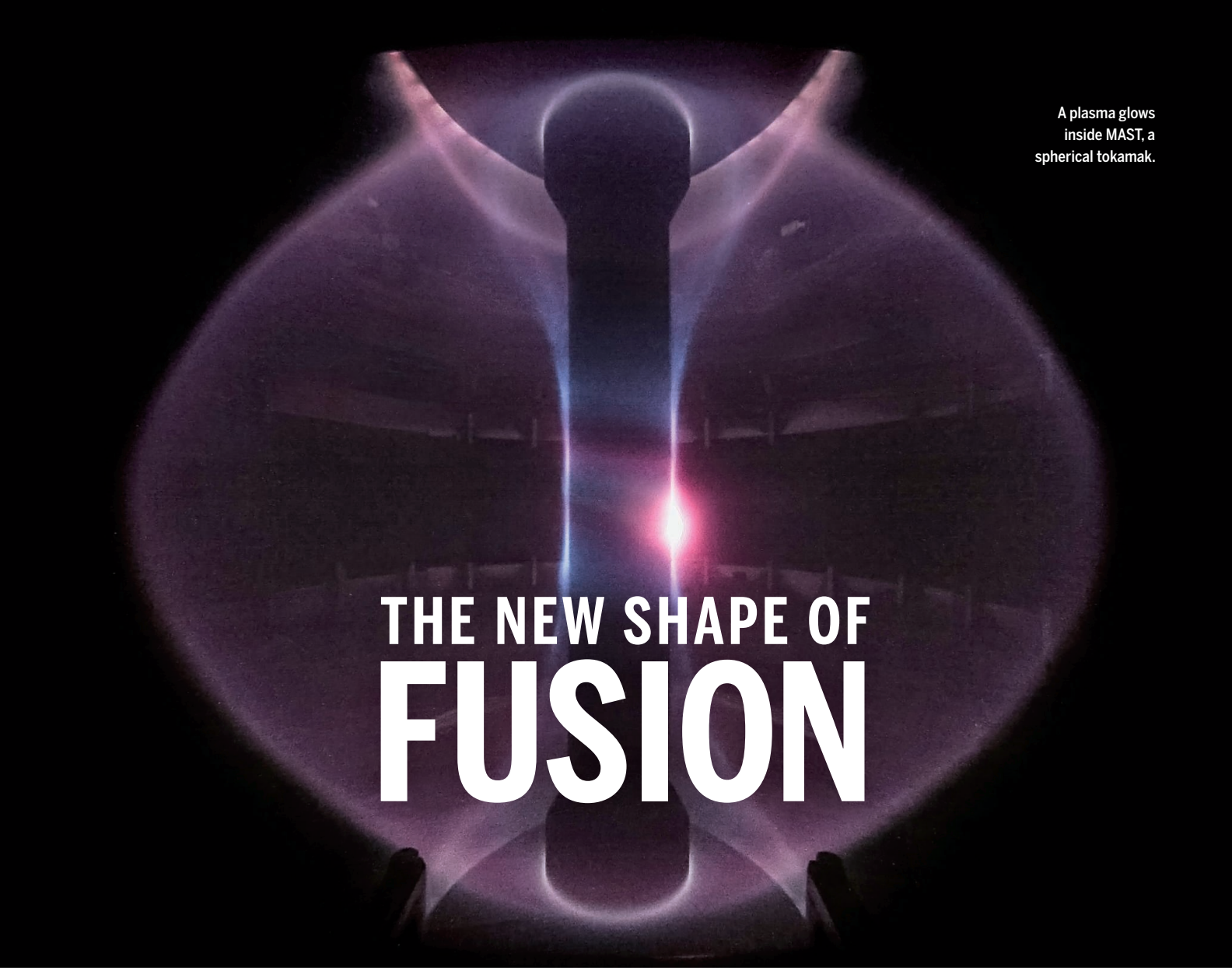
23 thousand
Annual U.S. deaths from antibiotic-resistant infections

2.05 million
Annual U.S. illnesses caused by antibiotic resistance

41
Estimated number of antibiotics now under development

23
Estimated years, from start of development, to turn a profit on an antibiotic

DATA: CDC/PEW CHARITABLE TRUSTS/U.K. *REVIEW ON ANTIMICROBIAL RESISTANCE*



A plasma glows inside MAST, a spherical tokamak.

THE NEW SHAPE OF FUSION

After decades of slow progress with doughnut-shaped reactors, magnetic fusion labs are gambling on a redesign

By **Daniel Clery**

ITER, the international fusion reactor being built in France, will stand 10 stories tall, weigh three times as much as the Eiffel Tower, and cost its seven international partners $18 billion or more. The result of decades of planning, ITER will not produce fusion energy until 2027 at the earliest. And it will be decades before an ITER-like plant pumps electricity into the grid. Surely there is a quicker and cheaper route to fusion energy.

Fusion enthusiasts have a slew of schemes for achieving the starlike temperatures or crushing pressures needed to get hydrogen nuclei to come together in an energy-spawning union. Some are mainstream, such as lasers, some unorthodox (*Science*, 25 July 2014, p. 370). Yet the doughnut-shaped vessels called tokamaks, designed to cage a superheated plasma using magnetic fields, remain the leading fusion strategy and are the basis of ITER. Even among tokamaks, however, a nimbler alternative has emerged: a spherical tokamak.

Imagine the doughnut shape of a conventional tokamak plumped up into a shape more like a cored apple. That simple change, say the idea's advocates, could open the way to a fusion power plant that would match ITER's promise, without the massive scale. "The aim is to make tokamaks smaller, cheaper, and faster—to reduce the eventual cost of electricity," says Ian Chapman, head of tokamak science at the Culham Centre for Fusion Energy in Abingdon, U.K.

Culham is one of two labs about to give these portly tokamaks a major test. The world's two front-rank machines—the National Spherical Torus Experiment (NSTX) at the Princeton Plasma Physics Laboratory in New Jersey and the Mega Amp Spherical Tokamak (MAST) in Culham—are both being upgraded with stronger magnets and more powerful heating systems. Soon they will

PHOTO: © CCFE

switch on and heat hydrogen to temperatures much closer to those needed for generating fusion energy. If they perform well, then the next major tokamak to be built—a machine that would run in parallel with ITER and test technology for commercial reactors—will likely be a spherical tokamak.

A small company spun off from Culham is even making a long-shot bet that it can have a spherical tokamak reactor capable of generating more energy than it consumes—one of ITER's goals—up and running within the decade. If it succeeds, spherical tokamaks could change the shape of fusion's future. "It's going to be exciting," says Howard Wilson, director of the York Plasma Institute at the University of York in the United Kingdom. "Spherical tokamaks are the new kids on the block. But there are still important questions we're trying to get to the bottom of."

TOKAMAKS ARE AN INGENIOUS WAY to cage one of the most unruly substances humans have ever grappled with: plasma hot enough to sustain fusion. To get nuclei to slam together and fuse, fusion reactors must reach temperatures 10 times hotter than the core of the sun, about 150 million degrees Celsius. The result is a tenuous ionized gas that would vaporize any material it touches—and yet must be held in place long enough for fusion to generate useful amounts of energy.

Tokamaks attempt this seemingly impossible task using magnets, which can hold and manipulate plasma because it is made of charged particles. A complex set of electromagnets encircle the doughnut-shaped vessel, some horizontal and some vertical, while one tightly wound coil of wire, called a solenoid, runs down the doughnut hole. Their combined magnetic field squeezes the plasma toward the center of the tube and drives it around the ring while also twisting in a slow corkscrew motion.

But plasma is not easy to master. Confining it is like trying to squeeze a balloon with your hands: It likes to bulge out between your fingers. The hotter a plasma gets, the more the magnetically confined gas bulges and wriggles and tries to escape. Much of the past 60 years of fusion research has focused on how to control plasma.

Generating and maintaining enough heat for fusion has been another challenge. Friction generated as the plasma surges around the tokamak supplies some of the heat, but modern tokamaks also beam in microwaves and high-energy particles. As fast as the heat is supplied, it bleeds away, as the hottest, fastest moving particles in the turbulent plasma swirl away from the hot core toward the cooler edge. "Any confinement system is going to be slightly leaky and will lose particles," Wilson says.

Studies of tokamaks of different sizes and configurations have always pointed to the same message: To contain a plasma and keep it hot, bigger is better. In a bigger volume, hot particles have to travel farther to escape. Today's biggest tokamak, the 8-meter-wide Joint European Torus (JET) at Culham, set a record for fusion energy in 1997, generating 16 megawatts for a few seconds. (That was still slightly less than the heating power pumped into the plasma.) For most of the fusion community, ITER is the logical next step. It is expected to be the first machine to achieve energy gain—more fusion energy out than heating power in.

In the 1980s, a team at Oak Ridge National Laboratory in Tennessee explored how a simple shape change could affect tokamak performance. They focused on the aspect ratio—the radius of the whole tokamak compared to the radius of the vacuum tube. (A Hula-Hoop has a very high aspect ratio, a bagel a lower one.) Their calculations suggested that making the aspect ratio very low, so that the tokamak was essentially a sphere with narrow hole through the middle, could have many advantages.

Near a spherical tokamak's central hole, the Oak Ridge researchers predicted, particles would enjoy unusual stability. Instead of corkscrewing lazily around the tube as in a conventional tokamak, the magnetic field lines wind tightly around the central column, holding particles there for extended periods before they return to the outside surface. The D-shaped cross section of the plasma would also help suppress turbulence, improving energy confinement. And they reckoned that the new shape would use magnetic fields more efficiently—achieving more plasma pressure for a given magnetic pressure, a ratio known as beta. Higher beta means more bang for your magnetic buck. "The general idea of spherical tokamaks was to produce electricity on a smaller scale, and more cheaply," Culham's Chapman says.

But such a design posed a practical problem. The narrow central hole in a spherical tokamak didn't leave enough room for the equipment that needs to fit there: part of each vertical magnet plus the central solenoid. In 1984, Martin Peng of Oak Ridge came up with an elegant, space-saving solution: replace the multitude of vertical ring magnets with C-shaped rings that share a single conductor down the center of the reactor (see graphic).

U.S. fusion funding was in short supply at that time, so Oak Ridge could not build a spherical machine to test Peng's design. A few labs overseas converted some small devices designed for other purposes into spherical tokamaks, but the first true example was built at the Culham lab in 1990. "It was put together on a shoestring with parts from other machines," Chapman says. Known as the Small Tight Aspect Ratio Tokamak (START), the device soon achieved a beta of 40%, more than three times that of any conventional tokamak. It also bested traditional machines in terms of stability. "It smashed the world record at the time," Chapman says. "People got more interested." Other labs rushed to build small spherical tokamaks, some in countries not known for their fusion research, including Australia, Brazil, Egypt, Kazakhstan, Pakistan, and Turkey.

The next question, Chapman says, was "can we build a bigger machine and get similar performance?" Princeton and Culham's

A ball of fire

Changing the shape of a fusion reactor from the traditional doughnut to an apple improves plasma stability and heat retention but requires redesigning the magnets that hold the plasma in place.

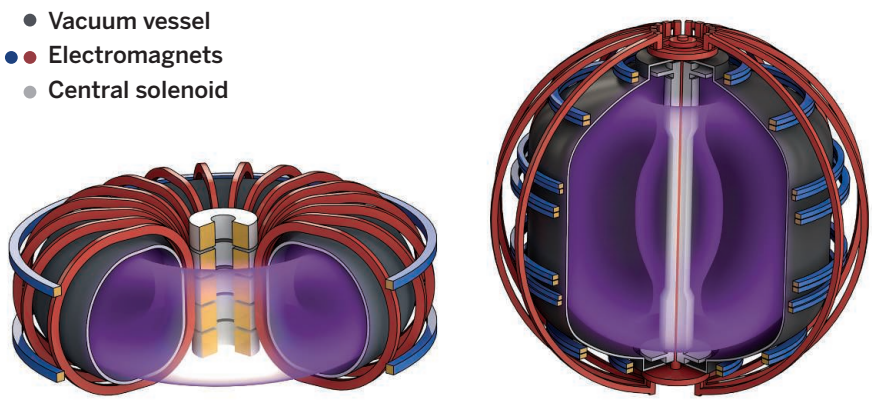


ILLUSTRATION: JAMES PROVOST

machines were meant to answer that question. Completed in 1999, NSTX and MAST both hold plasmas about 3 meters across, roughly three times bigger than START's but a third the size of JET's. The performance of the pair showed that START wasn't a one-off: again they achieved a beta of about 40%, reduced instabilities, and good confinement.

Now, both machines are moving to the next stage: more heating power to make a hotter plasma and stronger magnets to hold it in place. MAST is now in pieces, the empty vacuum vessel looking like a giant tin can adorned with portholes, while its £30 million worth of new magnets, pumps, power supplies, and heating systems are prepared. At Princeton, technicians are putting the finishing touches to a similar $94 million upgrade of NSTX's magnets and neutral beam heating. Like most experimental tokamaks, the two machines are not aiming to produce lots of energy, just learning how to control and confine plasma under fusionlike conditions. "It's a big step," Chapman says. "NSTX-U will have really high injected power in a small plasma volume. Can you control that plasma? This is a necessary step before you could make a spherical tokamak power plant."

The upgraded machines will each have a different emphasis. NSTX-U, with the greater heating power, will focus on controlling instabilities and improving confinement when it restarts this summer. "If we can get reasonable beta values, [NSTX-U] will reach plasma [properties] similar to conventional tokamaks," says NSTX chief Masayuki Ono. MAST-Upgrade, due to fire up in 2017, will address a different problem: capturing the fusion energy that would build up in a full-scale plant.

Fusion reactions generate most of their energy in the form of high-energy neutrons, which, being neutral, are immune to magnetic fields and can shoot straight out of the reactor. In a future power plant, a neutron-absorbing material will capture them, converting their energy to heat that will drive a steam turbine and generate electricity. But 20% of the reaction energy heats the plasma directly and must somehow be tapped. Modern tokamaks remove heat by shaping the magnetic field into a kind of exhaust pipe, called a divertor, which siphons off some of the outermost layer of plasma and pipes it away. But fusion heat will build up even faster in a spherical tokamak because of its compact size. MAST-Upgrade has a flexible magnet system so that researchers can try out various divertor designs, looking for one that can cope with the heat.

Researchers know from experience that when a tokamak steps up in size or power, plasma can start misbehaving in new ways. "We need MAST and NSTX to make sure there are no surprises at low aspect ratio," says Dennis Whyte, director of the Plasma Science and Fusion Center at the Massachusetts Institute of Technology in Cambridge. Once NSTX and MAST have shown what they are capable of, Wilson says, "we can pin down what a [power-producing] spherical tokamak will look like. If confinement is good, we can make a very compact machine, around MAST size."

VIDEO

Fusion scientists explain the new shape at http://scim.ag/fusionshape.

BUT GENERATING ELECTRICITY isn't the only potential goal. The fusion community will soon have to build a reactor to test how components for a future power plant would hold up under years of bombardment by high-energy neutrons. That's the goal of a proposed machine known in Europe as the Component Test Facility (CTF), which could run stably around the clock, generating as much heat from fusion as it consumes. A CTF is "absolutely necessary," Chapman says. "It's very important to test materials to make reactors out of." The design of CTF hasn't been settled, but spherical tokamak proponents argue their design offers an efficient route to such a testbed—one that "would be relatively compact and cheap to build and run," Ono says.

Engineers lift out MAST's vacuum vessel for modifications during the £30 million upgrade.

PHOTO: © CCFE

With ITER construction consuming much of the world's fusion budget, that promise won't be tested anytime soon. But one company hopes to go from a standing start to a small power-producing spherical tokamak in a decade. In 2009, a couple of researchers from Culham created a spinoff company—Tokamak Solutions—to build small spherical tokamaks as neutron sources for research. Later, one of the company's suppliers showed them a new multilayered conducting tape, made with the high-temperature superconductor yttrium-barium-copper-oxide, that promised a major performance boost.

Lacking electrical resistance, superconductors can be wound into electromagnets that produce much stronger fields than conventional copper magnets. ITER will use low-temperature superconductors for its magnets, but they require massive and expensive cooling. High-temperature materials are cheaper to use but were thought to be unable to withstand the strong magnetic fields around a tokamak—until the new superconducting tape came along. The company changed direction, was renamed Tokamak Energy, and is now testing a first-generation superconducting spherical tokamak no taller than a person.

Superconductors allow a tokamak to confine a plasma for longer. Whereas NSTX and MAST can run for only a few seconds, the team at Tokamak Energy this year ran their machine—albeit at low temperature and pressure—for more than 15 minutes. In the coming months, they will attempt a 24-hour pulse—smashing the tokamak record of slightly over 5 hours.

Next year, the company will put together a slightly larger machine able to produce twice the magnetic field of NSTX-U. The next step—investors permitting—will be a machine slightly smaller than Princeton's but with three times the magnetic field. Company CEO David Kingham thinks that will be enough to beat ITER to the prize: a net gain of energy. "We want to get fusion gain in 5 years. That's the challenge," he says.

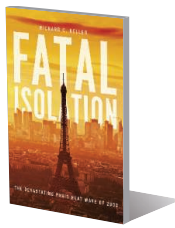
"It's a high-risk approach," Wilson says. "They're buying their lottery ticket. If they win, it'll be great. If they don't, they'll likely disappear. Even if it doesn't work, we'll learn from it; it will accelerate the fusion program."

It's a spirit familiar to everyone trying to reshape the future of fusion. ■

INSIGHTS

Cancer mutations in normal skin *p. 867*

The sociology of a natural disaster *p. 869*



PERSPECTIVES

An athlete competes in the women's triple-jump final in the 2012 Olympic games in London, UK.

PHOTO: ALEXANDER HASSENSTEIN/GETTY IMAGES

SCIENCE AND SOCIETY

Debating a testosterone "sex gap"

Policies unfairly exclude some women athletes from competition

By **Katrina Karkazis**[1]
and **Rebecca Jordan-Young**[2]

Sexual dimorphism of testosterone (T) in elite athletes was at the center of a recent case at the "Supreme Court of Sport," the Court of Arbitration for Sport in Switzerland, after teenage Indian sprinter Dutee Chand challenged a sports policy regulating competition eligibility of women with naturally high T. The idea of a "sex gap" in T is a cornerstone of this policy (*1*). Policymakers infer that men's higher T is the "one factor [that] makes a decisive difference" between men's and women's athletic performances (*2*)—so that women with naturally high T may unfairly enjoy a "massive androgenic advantage" over other women athletes (*2*). We report on an emerging scientific debate about whether the sex gap in T applies to elite athletes.

POLICY

In 2011 and 2012, respectively, the International Association of Athletics Federations (IAAF) and the International Olympic Committee (IOC) adopted controversial policies that regulate levels of natural T in women athletes (*1*, *3*). The IAAF policy sets a ceiling for women of 10 nmol/liter in blood, which it identifies as "within the normal male range," whereas the IOC policy targets levels "within the male range" (*1*, *3*). Women with high natural T, according to the IAAF, have an unfair advantage over women with lower natural levels (*1*). Unless they are androgen-resistant, women must lower their T in order to continue competing (*1*), which would require surgery or antiandrogens (*4*).

Appealing her exclusion under the IAAF policy, Chand told the Indian Express "At every level of my life … I have competed the way I am. I've been told the hormonal issue with me is natural so that's why we have decided [to appeal]" (*5*). Her March appeal was the first formal challenge to the policy; a decision is forthcoming.

The T policy is the latest attempt to use a biological marker to draw a bright line between women and men for sex-segregated sports. Decades of sex testing of all women athletes relied on biomarkers such as chro-

mosomes. Officials dropped blanket testing in the 1990s, acknowledging that sex is irreducibly complex and that there is no scientific criterion for separating all men from all women (*6*). Nevertheless, they retained an ad hoc policy for when a woman's sex was questioned, which was criticized for continuing the doomed project of sex testing and for being arbitrary (*7*).

Still determined to find a biological way to regulate who can compete as a woman, policy-makers turned to testosterone, arguing that T is both the "performance enhancing hormone" (*8*) and a sharply differentiated trait between men and women (*2*, *3*). In most studies, men's T levels are about 10 times those of women, and the highest levels seen in women are well below the lowest levels seen in healthy men. One policy-maker characterized this as "a huge no man's land" (*9*).

Recently, though, the idea of unequivocal sexual dimorphism in T levels has been challenged, at least in elite athletes. Only two large-scale studies of T in elite athletes exist, and they draw contradictory conclusions regarding a sex gap in T (*10*, *11*). In the first, data are from 446 men and 234 women across 15 highly varied Olympic events (*10*). These data were collected as part of the GH-2000 study, an IOC- and World Anti-Doping Agency–funded project aimed at developing a test to detect human growth hormone abuse. The report states, "hormone profiles from elite athletes differ from usual reference ranges" in both men and women (*10*). In fact, there was "overlap between men and women, although the mean values differ." Among women, 13.7% had T above the typical female range, and 4.7% were within the typical male range. In contrast to reference ranges, 16.5% of these elite male athletes had T below the typical male range, with 1.8% falling in the female reference range.

Not long after the GH-2000 report appeared, IAAF researchers published their own study on serum T in 849 elite women athletes in track and field from the 2011 Daegu World Championships (*11*). That study showed just 1.5% of women athletes with T above the female reference range, a sharp contrast with the 13.7% in the GH-2000 study.

DEBATING THE EVIDENCE. Three critiques of the GH-2000 report—raised by IAAF policy-makers in the Daegu study and in an IAAF-IOC rebuttal—bear on whether there is a sex gap in T (*11*, *12*). The first issue is how sera were analyzed: The GH-2000 study used immunoassay (IA); the Daegu study used mass spectrometry (MS). IA overestimates T at lower values. There is no question that MS yields more accurate T readings at lower values. The use of IA in the GH-2000 study might have resulted in some inaccurately high readings among women, but it cannot explain the fact that a considerable proportion of men had very low T levels (in fact, IA underestimation would have countered the latter pattern). So the use of IA cannot account for the finding of a male-female overlap in the GH-2000 data. The Daegu study did not report men's values, so it can only shed partial light on the question of a gap.

The second disagreement concerns when to draw serum, because T changes in response to competition. IAAF-IOC policy-makers suggest that the female-male overlap in T observed in the GH-2000 data may be an artifact of sampling within 2 hours after competition (*12*).

This criticism requires a selective reading of the evidence on the effect of competition on T levels. The IAAF-IOC critique cites a single report showing male T levels dropped and female levels were steady or modestly rose after an Ironman competition (*13*). The broader literature shows that T may rise, fall, or remain unchanged after competition, and the main factors determining the response seem to be the type and duration of competition—not the individual's sex (*14*, *15*). Intense resistance exercise and short-duration exercise are associated with increase in T, whereas endurance exercise (especially lasting greater than 3 hours) is associated with decrease in T (*14*, *15*). There are few data on endogenous T in women athletes, but the most recent review again indicates a great variety of responses to exercise are possible—including a large and lasting increase in resting T from long-term resistance training (*14*).

The timing of serum collection in the GH-2000 study makes sense in the antidoping context, because of the need to understand hormone profiles "after competition when anti-doping tests are usually made" (*10*). Doping tests are often how women with high natural T are flagged, so understanding how natural T responds to competition is important.

The tussle over timing may obscure the important point that T is dynamic. Recent research shows that, in both sexes, T dramatically responds to physical situations as well as social cues and contexts, diurnal rhythms, training, and other factors (*14*–*17*). For example, positive feedback from a coach can cause a rapid doubling of T level (*17*). The "correct" time to sample T depends on the purpose of the study, but the timing of blood draws seems unlikely to determine whether a study finds overlap in T between the sexes.

The third issue raised by the IOC–IAAF critiques of the GH-2000 study is the most fundamental: the rules for subject inclusion and exclusion. Both scientific groups agree that subjects who have doped should be excluded. Where they part is whether women with naturally high T should be excluded.

The two camps take opposite views on whether to include these women—a decision that bears directly on whether their findings support or undermine the policies. The GH-2000 study includes all women with high natural T in the sample. The Daegu study included women with high T of unknown etiology, but excluded as "confounding factors" all women whose high natural T can be traced to diverse sex development, also

Close-up of a baton before women's 4 x 400-meter relay final, 2012 Olympic games, London, UK.

PHOTO: SIMON BRUTY/SPORTS ILLUSTRATED, GETTY IMAGES

[1]Stanford University, Stanford, CA 94305, USA. [2]Barnard College, New York, NY 10027, USA.
E-mail: karkazis@stanford.edu

known as intersex (DSDI). In simple terms, some of the biological characteristics of women with DSDI would be classified as female and others as male. This challenges common ideas about sex, but it is widely recognized in medicine, law, and the social sciences that when people are born with mixed markers of sex (e.g., chromosomes, genitals, gonads), the medical standard is that gender identity is the definitive marker of sex—there is no better criterion (*18*).

What, then, is the logic that classifies women with DSDI as confounders? The Daegu report consistently pairs clinical language, such as "diagnosis" and "disorder," with hyperandrogenism for the women with DSDI, and in their rebuttal to the GH-2000 paper, IAAF-IOC policy-makers use the phrase "hyperandrogenic disorders of sex development" (*12*). This signals their judgment that women with DSDI are not healthy and, therefore, should be excluded from reference ranges. But DSDI women are not necessarily unhealthy. High T can be associated with health issues but is not, in and of itself, a health problem for women (*4*).

> ***"What looks like a [scientific] controversy ... is ultimately a social and ethical one concerning how we understand and frame human diversity."***

An a priori understanding of women with DSDI as unhealthy and, thus, outside normal variation creates a rationale for their exclusion both in reference ranges and the policies. But it is also circular: Because women with DSDI are a priori excluded when the reference ranges are created, the findings from the Daegu study—that women athletes have T levels no different from nonathlete women—reinforce their values as outsiders and justify the policy.

There is a strong scientific argument for including DSDI women in the sample. These studies aim to establish T reference ranges for elite athletes: i.e., the focus is on physiological ranges not clinical ranges. This calls for descriptive statistics, and in this case, there is no valid basis for discarding some values as outliers. In both studies, if the full range of values for women's endogenous T is included, there is an overlap in T.

CALCULATING FAIRNESS. What looks like a controversy rooted firmly in science is ultimately a social and ethical one concerning how we understand and frame human diversity. These assessments are not trivial: They shape not only the research methods and findings but also how we understand what is at stake in this policy. And this has very real consequences for people's lives.

Policy-makers, among others, claim that the problem is that women with naturally high T have unfair advantage, despite having acknowledged in their Daegu study that "there is no clear scientific evidence proving that a high level of T is a significant determinant of performance in female sports" (*11*). Others see a very different problem: Women who have lived and competed as women their whole lives suddenly find themselves having to undergo medical interventions in order to remain eligible to compete in a category to which everyone agrees they belong.

Calculating what counts as a fair and level playing field for women must take all women athletes into account, including those with naturally high T and/or DSDI. We could return to a consensus reached decades ago, where policy-makers faced these same concerns and concluded that women "who were raised as girls and classify themselves as female should not be excluded from competition as women" (*19*). In other words, ensuring that women with high endogenous T and/or DSDI "have the same rights to participation in athletics as all women" (*20*) would be a good place to start. ■

REFERENCES AND NOTES

1. IAAF, *IAAF Regulations Governing Eligibility of Females with Hyperandrogenism to Compete in Women's Competitions* (IAAF, Monaco, 2011).
2. S. Bermon, M. Ritzén, A. L. Hirschberg, T. H. Murray, *Am. J. Bioeth.* **13**, 63 (2013).
3. IOC, *IOC Regulations on Female Hyperandrogenism* (Lausanne, Switzerland, 2014).
4. R. M. Jordan-Young, P. H. Sönksen, K. Karkazis, *BMJ* **348**, g2926 (2014).
5. V. B. Padmadeo, *Indian Express*, 12 September 2014.
6. J. L. Simpson *et al.*, *JAMA* **284**, 1568 (2000).
7. L. J. Elsas *et al.*, *Genet. Med.* **2**, 249 (2000).
8. IAAF, *HA Regulations: Explanatory Notes* (IAAF, Monaco, 2011), pp. 1–4.
9. J. Macur, *New York Times*, 24 June 2012, p. SP6.
10. M. L. Healy *et al.*, *Clin. Endocrinol. (Oxf.)* **81**, 294 (2014).
11. S. Bermon *et al.*, *J. Clin. Endocrinol. Metab.* **99**, 4328 (2014).
12. M. Ritzén *et al.*, *Clin. Endocrinol. (Oxf.)* **82**, 307 (2015).
13. G. S. Ginsburg *et al.*, *Clin. Chim. Acta* **305**, 131 (2001).
14. J. L. Vingren *et al.*, *Sports Med.* **40**, 1037 (2010).
15. C. Enea *et al.*, *Sports Med.* **41**, 1 (2011).
16. S. M. van Anders, N. V. Watson, *Psychoneuroendocrinology* **31**, 715 (2006).
17. C. J. Cook, B. T. Crewther, *Horm. Behav.* **61**, 17 (2012).
18. P. A. Lee *et al.*, *Pediatrics* **118**, e488 (2006).
19. A. Ljungqvist, J. L. Simpson, *JAMA* **267**, 850 (1992).
20. Women's Sports Foundation, *Participation of Intersex Athletes in Women's Sports* (WSF, New York, 2011), pp. 1–5.

ACKNOWLEDGMENTS

Supported by NSF grants 1331115 and 1331123.

10.1126/science.aab1057

NEUROSCIENCE

Reading the mind to move the body

Decoding neural signals of intention and movement should guide the development of neural prosthetics

By **J. Andrew Pruszynski**[1] *and* **Jörn Diedrichsen**[2]

Imagine a world in which your smartphone can read your mind. Just at the moment that you decide to move your finger to delete a message, it is already gone. This sounds like science fiction, but for one human in California, this fantasy is becoming reality. On page 906 of this issue, Aflalo *et al.* (*1*) report the case of a tetraplegic individual (called "EGS") who volunteered to have his brain implanted with two small silicon chips that allow researchers to read his intentions directly from his brain activity. The chips—initially developed at the University of Utah (*2*) and now commercially available and approved for human use by the U.S. Food and Drug Administration—consist of a matrix of 96 microscopic electrodes that can record the activity of about 100 nerve cells at the same time.

The main goal of the implantation procedure was to restore EGS's ability to act in his environment. Paralyzed from the neck down, he currently relies on the help of others to perform almost all the daily actions that the vast majority of us take for granted. Using the signals from his brain and bypassing his damaged spinal cord, researchers hope to help him do these things again by allowing him to steer a robotic arm so that he can, for example, reach out, grasp a glass, and take a drink. Alternatively, the acquired signals can be used to control a cursor on a screen so that he can efficiently interact with a computer.

Previously, researchers have implanted chips into regions of the human brain that are closely related to the production of movements, such as the primary motor cortex (*3*, *4*), with the aim of reanimating a limb or controlling a prosthetic. Aflalo *et al.* have taken a different approach. They have

implanted neural recording devices in two locations of the posterior parietal cortex. From many years of basic research in monkeys, it is well established that the activity (firing patterns) of nerve cells in these areas contain a great deal of information not only about planned movements, but also more abstract concepts such as goals and intentions. For example, researchers can robustly "read out" a monkey's decision-making process as it deliberates between alternative actions—that is, look at firing patterns of neuronal activity and decode the decision that the monkey is going to make (*5*). Functional imaging of brain activity and brain lesion studies indicate that similar types of information processing occur in the human posterior parietal cortex (*6*).

Even though EGS was paralyzed more than 10 years ago, Aflalo *et al.* report that nerve cells in his posterior parietal cortex respond when he imagines making a particular movement. Indeed, the researchers were able to reliably read out where EGS intended to move by analyzing the firing patterns of about 100 nerve cells. This information was then used to steer a computer cursor or to direct a robotic arm situated beside EGS to the intended location. Aflalo *et al.* could also read out the velocity of the desired movement, and therefore determine when and how fast EGS wanted to move. The neural signals even provided information about whether EGS wanted to use his left or right hand to move to that location, lending hope to the idea that a single neural implant in the posterior parietal cortex could reanimate two limbs.

In a separate experiment, Aflalo *et al.* showed EGS the activity of a single nerve cell on a computer monitor and he was able to reliably and voluntarily modulate the activity of that nerve cell. These results extend classical work showing that monkeys could be operantly conditioned to regulate the firing rate of specific nerve cells when given similar feedback (*7*). However, Aflalo *et al.* could go further than the previous studies because they could explicitly ask their participant to tell them how he achieved these changes. EGS reported that he was often able to change the activity of these nerve cells by imagining particular motor actions. Such intentional modulation could be remarkably specific. One nerve cell, for example, would increase its activity when he imagined rotating his shoulder, and decrease its activity when he imagined touching his nose. Another nerve cell was activated when EGS imagined moving his hand to his mouth but not when he imagined touching his ear or chin.

The results of Aflalo *et al.* represent one more step toward making brain control of a robotic limb or computing device a reality. Despite the impressive series of steps taken over the past 15 years, however, these neural prosthetic devices still have a substantial way to go before becoming practical therapeutic interventions (*8*). Indeed, work is needed on many fronts, such as improving the durability of the implants, refining the isolation of single nerve cells, optimizing computational algorithms for interpreting the signals, and developing stimulation protocols to "write in" sensory signals from the prosthetic device into the brain. Of particular note is the fact that current systems run wires from within the brain to the outside world—a route for potential infection. In the long term, such systems need to become wireless and contained within the body, like modern pacemakers and cochlear implants. The results of Aflalo *et al.* do promise to deliver one of the missing pieces. The ability to decode signals that are related not only to the details of the movement but also to the patient's overall intention could improve brain control of a robot or cursor tremendously (see the figure). Ultimately, patients could have recording chips implanted in both the posterior parietal cortex and motor cortex, with the former being used to constrain the overall goal of the desired action and the latter providing fine control of the kinematic and dynamic details of the movement.

Beyond the important practical implications of these findings, the ability to record from many nerve cells in the human posterior parietal cortex opens up fascinating new avenues for basic research. For the first time, the activity of nerve cells in this area can be directly measured while simultaneously getting a verbal report about the conscious experience of the person from whom this neural activity is being gathered. This unique capacity allows Aflalo *et al.* to relate the patterns of neural activity associated with intention to the conscious experience of forming them. Such experiments should provide new insights into whether a person's future decisions can be decoded from his or her neural activity before the individual is aware of having formed them (*9*), fundamentally challenging our understanding of intentionality and free will. ■

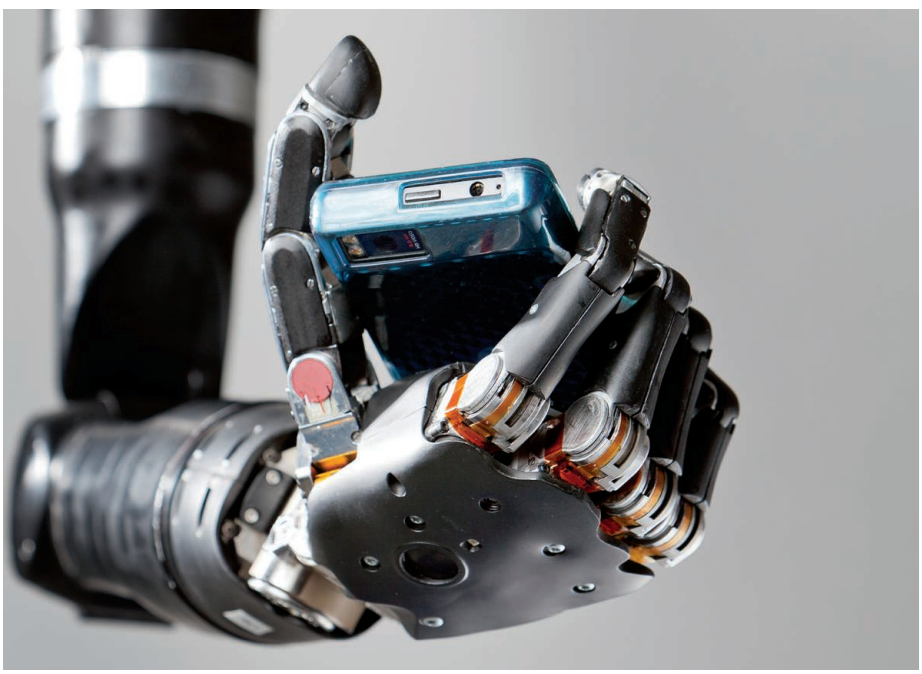


PHOTO: DARPA/JOHNS HOPKINS APPLIED PHYSICS LABORATORY

Imagine that. The ability to decode signals from neural activity in the brain related to details of movement, as well as signals related to the goals of the movement, should improve the design and operation of neural prosthetics. Patients may one day have chips of electrodes implanted in both the posterior parietal cortex and motor cortex to record neuronal activities that would then be decoded and used to control prosthetic limbs.

[1]*Department of Physiology and Pharmacology, Western University, London, Ontario N6A 3K7, Canada.* [2]*Institute of Cognitive Neuroscience, University College London, London WC1E 6BT, UK.*
E-mail: andrew.pruszynski@uwo.ca; j.diedrichsen@ucl.ac.uk

REFERENCES

1. T. Aflalo *et al.*, *Science* **348**, 906 (2015).
2. K. E. Jones, P. K. Campbell, R. A. Normann, *Ann. Biomed. Eng.* **20**, 423 (1992).
3. L. R. Hochberg *et al.*, *Nature* **442**, 164 (2006).
4. J. L. Collinger *et al.*, *Lancet* **381**, 557 (2013).
5. M. L. Platt, P. W. Glimcher, *Nature* **400**, 233 (1999).
6. J. P. Gallivan, D. A. McLean, F. W. Smith, J. C. Culham, *J. Neurosci.* **31**, 17149 (2011).
7. E. E. Fetz, *Science* **163**, 955 (1969).
8. V. Gilja *et al.*, *IEEE Trans. Biomed. Eng.* **58**, 1891 (2011).
9. B. Libet, E. W. Wright Jr., C. A. Gleason, *Electroencephalogr. Clin. Neurophysiol.* **56**, 367 (1983).

10.1126/science.aab3464

MICROBIOLOGY

Just add lanthanides

Some methanol-using bacteria may depend on lanthanide elements for carbon capture and energy generation

By **Elizabeth Skovran**
and **Norma Cecilia Martinez-Gomez**

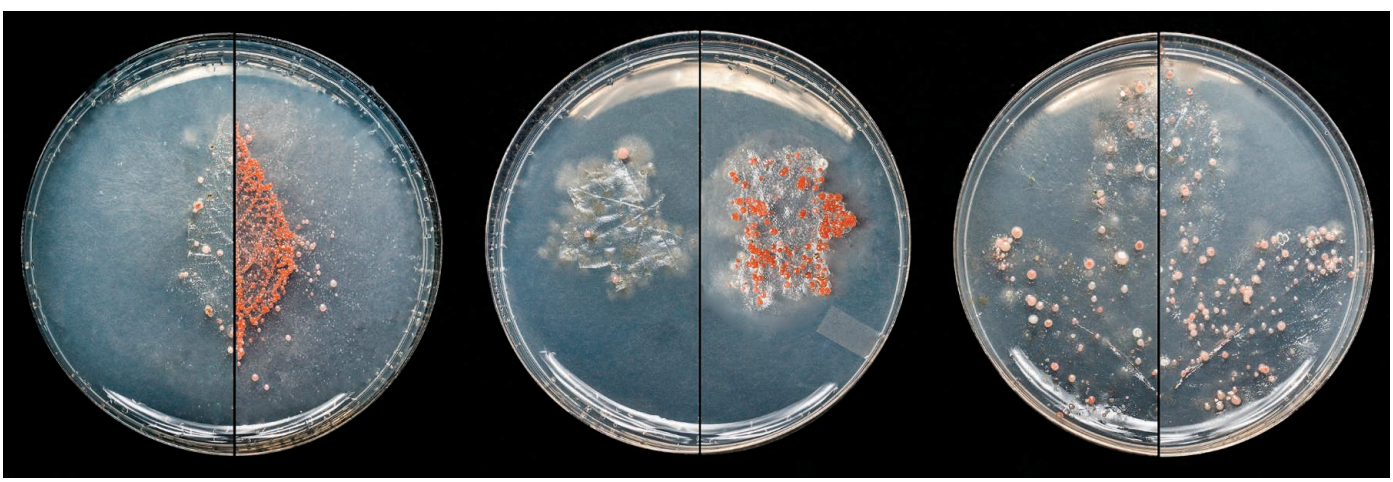
Lanthanides are used in items such as hybrid-car and smartphone batteries, magnets, and catalytic converters. Deceptively referred to as rare-earth elements (REEs), they are relatively abundant in Earth's crust but highly insoluble and scarce in pure form, requiring harsh extraction methods for purification. In China, lanthanides have been added to fertilizers and animal feed stocks to promote growth, although the growth of some crops is inhibited by lanthanides (*1*). Increased exposure to lanthanides has raised concerns that consumption of these elements in food or polluted water may have a negative impact on animals, including humans (*2*). However, recent studies have discovered that lanthanides are very important to a specialized group of bacteria that play a vital role in global carbon cycling.

In animals, lanthanides can affect bone integrity and cell signaling by displacing Ca^{2+} and can promote apoptosis in cell lines by increasing the concentrations of reactive oxygen species (*3*). Li *et al*. have shown that the biological diversity at lanthanide mining sites decreased when lanthanides were found in high concentrations (*4*), prompting researchers to isolate microorganisms that can effectively concentrate lanthanides in their cell walls for use in future bioremediation endeavors (*5*).

Although lanthanides can substitute for Ca^{2+} in some enzymes and tissues, scientists long considered the evolution of lanthanide-dependent enzymes to be implausible because of the low solubility of these elements in the environment. Recently, this belief has been challenged by the finding that REEs such as Ce^{3+} and La^{3+} are required for the activity of XoxF, a widespread but poorly characterized methanol dehydrogenase (MDH) enzyme used by some bacteria to oxidize methanol for carbon and energy (*6*, *7*).

Methylotrophic bacteria, which use single-carbon chemicals for growth, are ubiquitous in the environment. Methanol-using methylotrophs are often found on leaf surfaces, where they capture methanol released by plants during cell wall synthesis. The bacterium *Methylobacterium extorquens* AM1 serves as a model organism for studying methanol use. Its genome sequence revealed that in addition to the *mxaFI* genes encoding the well-studied pyrroloquinoline quinone (PQQ)–dependent MDH enzyme, there exist two other genes with sequence similarity to *mxaF*: *xoxF1* and *xoxF2*. These putative dehydrogenases were also predicted to require PQQ for activity. Although the PQQ-dependent MxaFI is Ca^{2+}-dependent, subsequent work showed that addition of La^{3+} to the growth medium increased the MDH activity of several *Methylobacterium* species and that the XoxF enzyme (but not the MxaFI enzyme) was present and active in cells grown with La^{3+} (*7*). The *xoxF* gene is preferentially transcribed over the *mxaFI* genes if La^{3+} is present, which suggests that *M. extorquens* AM1 can actively sense and respond to lanthanides in the environment (*8*). Transcription of the *mxaFI* genes in *M. extorquens* AM1 requires XoxF, indicating that XoxF has a regulatory as well as a catalytic role in this organism (*9*). The mechanism behind this regulation is not understood.

The contribution of lanthanide-containing XoxF enzymes to methanol oxidation in the environment has likely been vastly underestimated. Genomic DNA sequences from methylotrophic communities indicate that all PQQ-using methylotrophs have genes for the XoxF MDH; far fewer also have genes for the MxaFI MDH (*10*). Additionally, phylogenetic analyses reveal that *xoxF* sequences fall into five distinct phylogenetic groups (*10*). These diverse groups of XoxF enzymes may have different lanthanide and substrate preferences, resulting in different oxidation products. Further, XoxF homologs have been identified in nonmethylotrophs, which suggests that perhaps these organisms can use methanol as an energy source if not a carbon source (*10*). These discoveries suggest that addition of lanthanides to growth media may allow researchers to culture organisms that could not previously be grown in the laboratory. An illustration of this pos-



Living on rare earth elements. The backsides of leaves obtained from plants found on the San José State University campus grounds were pressed onto an ammonium mineral salts medium (ATCC medium: 784 AMS) that either lacked (left side) or contained (right side) 20 μM $LaCl_3$. They were then incubated at room temperature for 1 week. In testing some but not all leaves, the addition of La^{3+} allowed more methanol-using bacteria (pink) to grow. Recent work has suggested that a particular enzyme in these methanol-using bacteria requires lanthanides.

PHOTO: WILLIAM F. CARLSON IV

Department of Biological Sciences, San José State University, San José, CA 95192, USA. E-mail: elizabeth.skovran@sjsu.edu

sibility can be obtained by pressing various leaves onto methanol medium containing or lacking La^{3+} (see the figure). For some but not all leaves, addition of La^{3+} results in an increased number of pink-pigmented methylotrophs.

Are all lanthanides equal in their ability to support XoxF function? In 2014, Pol *et al.* investigated the ability of lanthanides to support growth of *Methylacidiphilum fumariolicum* SolV, a bacterium isolated from volcanic mud pots (*6*). Growth of *M. fumariolicum* SolV in the laboratory was poor unless volcanic mud pot water was added to the growth medium. Different lanthanides such as La^{3+}, Ce^{3+}, Pr^{3+}, and Nd^{3+} could substitute for the mud pot water, allowing rapid growth of the strain. Further, the catalytic properties of purified XoxF from *M. fumariolicum* SolV differed from those previously described for XoxF from *Methylobacterium* species: Methanol was oxidized to formate instead of formaldehyde, neutral pH was optimal for the reaction, and activation by ammonia was not required (*6*). XoxF crystal structure analysis and density functional theory calculations together support the hypothesis that, relative to Ca^{2+}, lanthanides are more efficient Lewis acids in the polarization of PQQ (which is necessary for substrate activation) (*6*, *11*). The recent isolation of a hybrid MDH containing two XoxF and two MxaI subunits from *Candidatus Methylomirabilis oxyfera* highlights the potential diversity of these PQQ-dependent enzymes (*12*).

Our understanding of the biological role of lanthanides is in its early stages. It is unknown how these highly insoluble elements are acquired and transported into cells. Studies on the biological roles of lanthanides may allow researchers to isolate and culture new organisms from the environment, to engineer a wide array of dehydrogenases for use in industry, to develop bioremediation strategies for cleanup of REE mining sites, and to reduce the potential for toxicity in our food and water. ■

REFERENCES

1. C. M. Rico *et al.*, *Environ. Sci. Pollut. Res. Int.* 10.1007/s11356-015-4243-y (2015).
2. F. Goecke *et al.*, *Front. Microbiol.* **6**, 2 (2015).
3. G. Pagano *et al.*, *Ecotoxicol. Environ. Saf.* **115**, 40 (2015).
4. J. Li, M. Hong, X. Yin, J. Liu, *J. Rare Earths* **28**, 957 (2010).
5. T. Horiike, M. Yamashita, *Appl. Environ. Microbiol.* **81**, 3062 (2015).
6. A. Pol *et al.*, *Environ. Microbiol.* **16**, 255 (2014).
7. Y. Hibi *et al.*, *J. Biosci. Bioeng.* **111**, 547 (2011).
8. E. Skovran *et al.*, Gordon Conference on Metals in Biology, Ventura, CA, 25 to 30 January 2015.
9. E. Skovran *et al.*, *J. Bacteriol.* **193**, 6032 (2011).
10. A. Beck *et al.*, *PLOS ONE* **9**, 102458 (2014).
11. J. A. Bogart, A. J. Lewis, E. J. Schelter, *Chemistry* **21**, 1743 (2015).
12. M. L. Wu *et al.*, *Appl. Environ. Microbiol.* **81**, 1442 (2014).

10.1126/science.aaa9091

ORGANIC CHEMISTRY

Streamlining amine synthesis

Bulky amine groups that help make many drugs more bioavailable can be added readily to organic compounds

By **László Kürti**

Amines, a collective name for compounds that contain one or more nitrogen atoms, and their derivatives make up the overwhelming majority of drug molecules and agrochemicals, as well as many compounds that are produced by plants and living organisms (i.e., natural products) (*1*, *2*). Not surprisingly, organic chemists spend a considerable amount of time with the synthesis and late-stage functionalization of amines. On page 886 of this issue, Gui *et al.* (*3*) report a highly innovative iron-catalyzed cross-coupling of olefins with nitroarenes, both of which are readily available and inexpensive, to afford bulky secondary arylamines that are either very difficult to obtain or inaccessible with existing methods.

Aromatic amines (also referred to as arylamines or anilines) appear as substructures in more than one-third of drug candidates (*4*, *5*) that serve as key chemical building blocks for the preparation of biologically active compounds, especially in medicinal chemistry. There are many well-established methods (see the figure, panel A) available in a modern organic chemist's toolbox for the synthesis of amines and, for the nonspecialist, it might appear that amines of any structural type can be quickly and reliably prepared. However, the preparation of sterically hindered (i.e., bulky) *N*-aryl-*N*-alkyl amines (structures I to IV, panel B of the figure) is still a major challenge, as none of the currently used methods allow their rapid and cost-effective synthesis. These bulky amine building blocks are highly sought-after, as the presence of the sterically demanding alkyl groups markedly improves the druglike properties of biologically active compounds, including their lipophilicity (i.e., solubility in fats, oils, and lipids) and metabolic stability toward many enzymes that are present in living organisms (*6*). Thus, the continued development of novel and powerful methods in synthetic organic chemistry is needed to make complex structures quickly and cost-effectively.

The formal hydroamination process is operationally simple, scalable, and avoids the use of protecting groups, which tend to reduce efficiency by adding extra steps to a synthetic sequence. The scope of both coupling partners, especially in terms of their steric and electronic properties, is exceptionally wide and renders this transformation a compelling alternative to currently utilized copper- and palladium-catalyzed cross-coupling (*7*) approaches that proceed with considerably reduced efficiency in the case of sterically demanding arylamine targets.

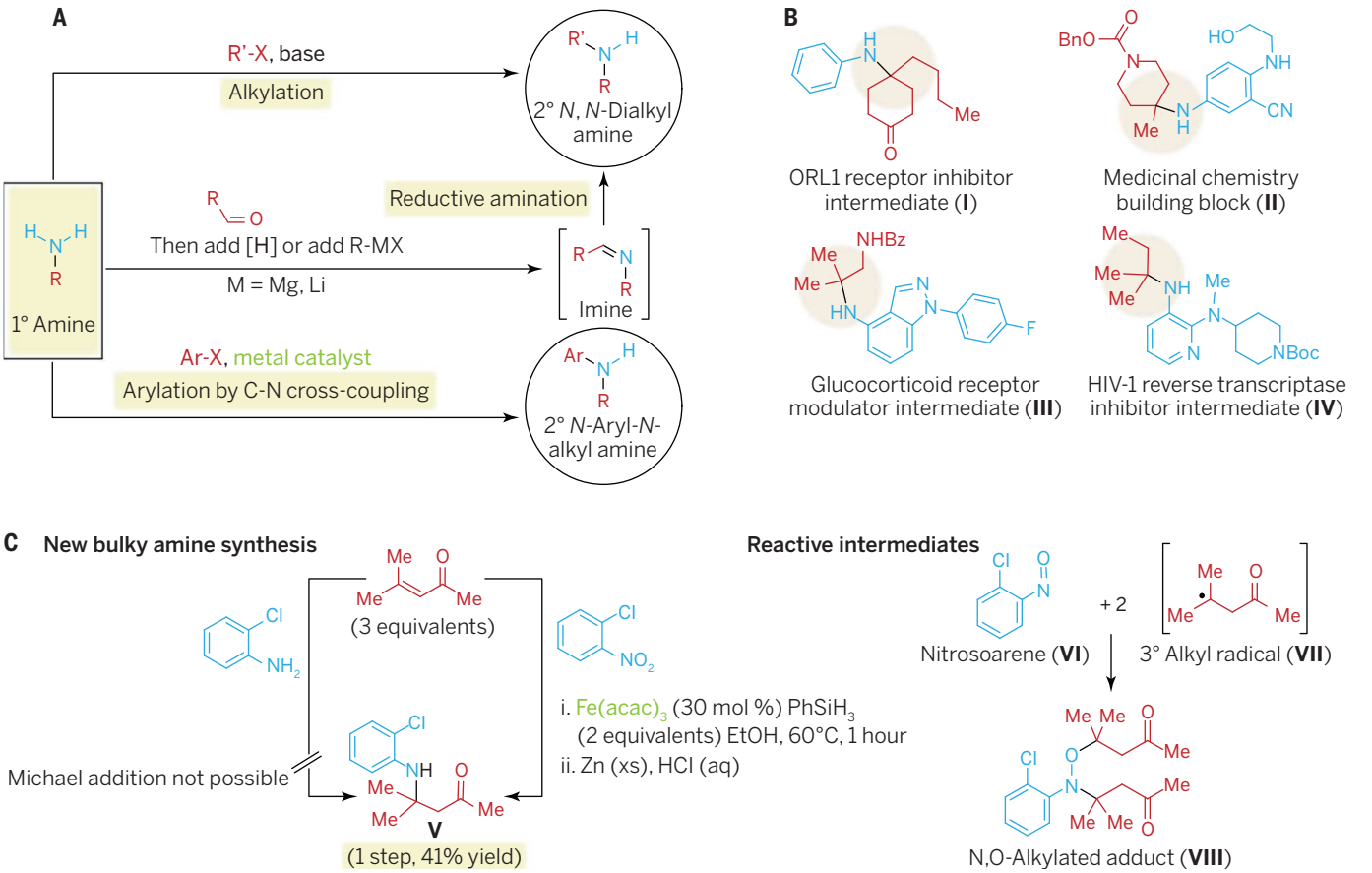
The method developed by Gui *et al.* is orthogonal (i.e., nonoverlapping) to other arylamine syntheses and provides rapid preparative access to structurally diverse secondary amine products via a simple one-pot process that takes place under mild reaction conditions. The chemoselectivity, the preferential reaction of one functional group over others in the same molecule, is excellent, and sensitive functional groups such as ketones, free alcohols/amines, and even boronic acids are well tolerated. Aromatic C(sp^2)-halogen and C(sp^2)-O-triflate bonds remain unchanged, which allows product diversification via classical C-C, C-N, and C-O cross-coupling reactions (*8*).

> ***"...there can be little doubt that this new transformation will find wide applicability in both academic and industrial laboratories."***

Key to the success of this method is the simultaneous generation of a tertiary alkyl radical (VII, panel C of the figure) from the olefin and the efficient reduction of the nitroarene to the corresponding nitrosoarene (VI). An inexpensive iron salt is used as the catalyst and a silane as the stoichiometric reducing agent, a set of conditions that Baran and co-workers had identified for the radical coupling of alkenes (*9*). Two equivalents of the alkyl radical (VII) add across the N=O double bond of the nitrosoarene (VI) to afford an *N*,*O*-alkylated adduct (VIII); the de-

Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. Address as of 1 June 2015: Department of Chemistry, Rice University, Houston, TX 77005, USA. E-mail: kurti.laszlo@rice.edu

Synthetic routes to amines, then and now



Synthetic access to bulky amines. (**A**) Several well-established methods for the synthesis of amines are shown. (**B**) Examples of sterically hindered amine building blocks that are difficult to access with currently available synthetic methods. (**C**) Gui *et al*. use an inexpensive iron salt and reducing agent for one-pot cross-coupling of nitoarenes with olefins to afford bulky secondary arylamines. Two key intermediates are formed in situ: a nitrosoarene and a tertiary alkyl radical that initially afford a N,O-alkylated adduct that is later reduced to the desired product.

sired bulky secondary arylamine product (V) is revealed after a simple reductive workup in which the N-O bond is cleaved. Radicals tend to react in a highly chemoselective fashion, so harsh reaction conditions can be avoided and functional group interconversions can be kept at a minimum (*10*).

Given that an excess olefin coupling partner is needed (i.e., 3 equivalents), structurally complex and valuable olefin building blocks are not practical to use in this transformation. Nonetheless, this previously unexploited C-N bond disconnection invented by Gui *et al.* allows rapid synthetic access to valuable, and heretofore hard-to-prepare, bulky secondary aryl amine building blocks in which molecular complexity is built up in a single step. Since the starting materials and the reagents are inexpensive, the iron catalyst is abundant, and protecting groups are mostly unnecessary, the overall cost and material throughput of a given synthetic sequence that utilizes this new olefin hydroamination process will be vastly improved compared to existing approaches. Thus, there can be little doubt that this new transformation will find wide applicability in both academic and industrial laboratories.

It is expected that modified and improved versions of this transformation will be developed that address some of the current shortcomings such as the need for multiple equivalents of olefin coupling partner and for the final reductive cleavage of the N-O bond. Moreover, the atom economy of the process would improve if the phenylsilane reducing agent could be replaced by cheaper and more abundant sources, such as H_2. Functional group compatibility would improve if the combination of excess zinc metal and strong acid could be substituted with a nonmetal reduction source under neutral conditions for the final N-O bond cleavage. The report of Gui *et al.* raises the intriguing possibility about potentially rendering this olefin hydroamination reaction catalytic and asymmetric, as in several of the products a new and fully substituted carbon stereogenic center is created. ■

REFERENCES

1. A. Ricci, Ed., *Amino Group Chemistry: From Synthesis to the Life Sciences* (Wiley-VCH, Weinheim, Germany, 2008).
2. Y. Ishihara, A. Montero, P. S. Baran, *The Portable Chemist's Consultant* (iBook, Apple Publishing Group, New York, 2013).
3. J. Gui *et al.*, *Science* **348**, 886 (2015).
4. Z. Rappoport, Ed., *The Chemistry of Anilines, Part 1 and 2. Patai Series: The Chemistry of Functional Groups* (Wiley, Chichester, UK, 2007).
5. S. D. Roughley, A. M. Jordan, *J. Med. Chem.* **54**, 3451 (2011).
6. L. Wanka, K. Iqbal, P. R. Schreiner, *Chem. Rev.* **113**, 3516 (2013).
7. A. K. Yudin, Ed., *Catalyzed Carbon-Heteroatom Bond Formation* (Wiley-VCH, Weinheim, Germany, 2011).
8. A. de Meijere, S. Braese, M. Oestreich, Eds., *Metal-Catalyzed Cross-Coupling Reactions and More* (Wiley-VCH, Weinheim, Germany, 2014), vol. 1 to 3.
9. J. C. Lo, Y. Yabe, P. S. Baran, *J. Am. Chem. Soc.* **136**, 1304 (2014).
10. S. L. Castle, *Nature* **516**, 332 (2014).

10.1126/science.aab2812

MARINE BIOLOGY

Uncovering hidden worlds of ocean biodiversity

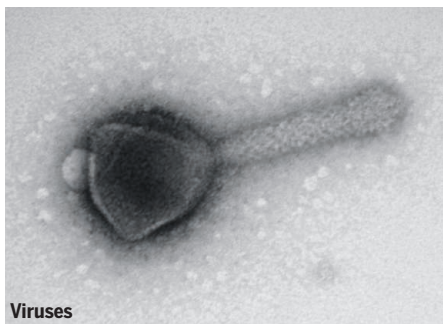
A 3-year expedition yields a treasure trove of data on microorganisms and small animals in the world's oceans

By **E. Virginia Armbrust**[1] *and* **Stephen R. Palumbi**[2]

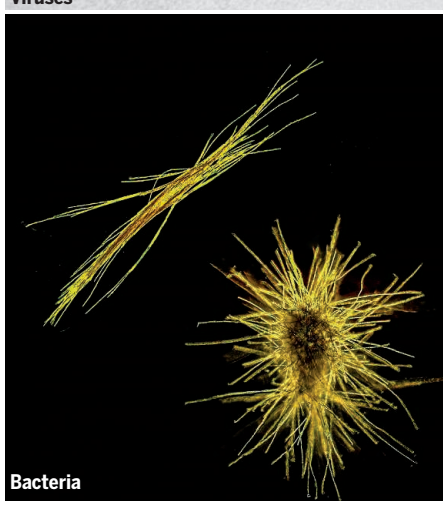
A bewildering swirl of tiny creatures dominates life in the oceans. More numerous than the stars in the universe, these organisms serve as the foundation of all marine food webs, recycling major elements and producing and consuming about half the organic matter generated on Earth each year (*1*). In this issue, five research articles from the *Tara* Oceans expedition (*2*–*6*) provide a vivid, potentially transformative view of the genetic diversity and interconnectivity of these unseen marine communities of viruses, bacteria, archaea, single-celled eukaryotes, and small planktonic animals (see the figure). Together, these studies deliver compelling evidence for extensive networks of previously hidden biological interactions in the sea.

The *Tara* Oceans expedition harkens back to 18th-century sailing voyages that explored uncharted worlds, including Darwin's voyage aboard the *HMS Beagle* and the *Challenger* expedition that heralded the beginning of modern oceanography. The 36-m schooner *Tara* departed Lorient, France, in 2008 and sailed through the Mediterranean Sea and into the Indian, South Atlantic, and Southern oceans (see the map). *Tara* visited coral reefs in the South Pacific Ocean and then sailed through the Panama Canal and back across the North Atlantic Ocean, arriving at her homeport nearly 3 years later. At hundreds of locations along the way, scientists and crew collected thousands of samples from surface waters, from the deep chlorophyll maximum layers where microscopic photosynthetic organisms accumulate, and from deeper waters.

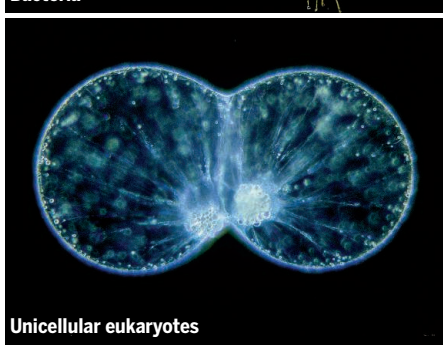
The researchers partitioned the seawater samples into seven size classes, ranging from the smallest viruses to animals less than 2 mm in size. This is where similarities to bygone voyages end: Hundreds of researchers from around the world used modern DNA sequencing, state-of-the-art microscopy, and computational analyses to examine underlying patterns and drivers of biodiversity.



Viruses



Bacteria



Unicellular eukaryotes

Zooplankton

Ocean diversity. During the *Tara* Oceans expedition, scientists studied ocean biodiversity in thousands of samples from the world's oceans. From top to bottom, the images show an aquatic virus; filament colonies formed by the cyanobacterium *Trichodesmium* (colony diameter: about 2 mm); the single-cell dinoflagellate *Noctiluca scintillans* (sea sparkle; typical diameter: 0.5 mm) undergoing cell division; and the about 1.5-mm-long zooplankton *Appendicularian Oikopleura dioca*. Five reports in this issue report results from the expedition (*2*–*6*).

Most small organisms in the sea cannot be grown and studied in the laboratory and are known only by DNA sequence barcodes—16*S* ribosomal gene sequences for bacteria and archaea and 18*S* ribosomal gene sequences for eukaryotes. Even less is known about the genetic makeup of these organisms. Whole-genome sequences are available for the relatively few cultured representatives. Computational approaches (*7*) and single-cell genomics (*8*) have generated whole-genome sequences for a select few of the vast majority of marine organisms that remain uncultured. About 10 years ago, the *Sorcerer II* Global Ocean Survey (GOS) first obtained DNA sequences from entire communities of marine microbes, uncovering previously unknown species and genes across the global ocean (*9*). Building on these results, the *Tara* Oceans expedition has added to the global genetic database a remarkable 7.2 trillion bases of metagenomic DNA sequence from viruses, bacteria, archaea, and those eukaryotes less than ~3 μm in size. In addition, it has yielded millions of 16*S* and 18*S* ribosomal bar codes derived from organisms ranging in size from 0.2 μm to 2 mm.

An important outcome of the expedition is the creation by Sunagawa *et al.* (page 873) (*2*) of an Ocean Microbial Reference Gene Catalog: a collection of more than 40 million nonredundant genes that are blueprints for metabolic function. Given the staggering amount of new sequence information, it is perhaps not surprising that Sunagawa *et al.* found relatively little overlap with the more restricted GOS gene catalog and even less with available reference sequences, emphasizing the likely enormous reservoir of unexplored genetic diversity in these communities. This work also serves as a reminder that the genetic content of domesticated laboratory organisms can differ greatly from that of the abundant taxa found in the wild (*10*). The data set undoubtedly holds numerous additional whole-genome sequences that await computational reconstruction. Sunagawa

[1]*School of Oceanography, University of Washington, Seattle, WA 98195, USA.* [2]*Hopkins Marine Station, Stanford University, Stanford, CA 93950, USA.*
E-mail: armbrust@uw.edu; spalumbi@stanford.edu

et al. estimate that a few tens of thousands of 16*S*-based taxa inhabit the upper ocean, similar to earlier predictions from the International Census of Marine Microbes (*11*). Ocean temperature appears to be the main factor driving these distributions.

Categorizing viruses in the oceans poses a special challenge. Viruses lack a universal molecular identifier, and only a tiny fraction of viruses can be grown in the laboratory. Brum *et al.* (page 874) (*3*) focused on a subset of DNA-based viruses to generate a "viral pan metagenome." Based on comparisons with other studies (*12*), they suggest that the extent of viral genetic diversity in the upper ocean appears well sampled and consists of nearly 1.5 million proteins. As in Sunagawa *et al.*'s study, there was little overlap between their data set and the genetic composition of cultured viral representatives. Although viral community patterns are also influenced by temperature, the biggest driver appears to be related to regional environmental conditions that support seed populations from localized hosts, which are then distributed more broadly via ocean currents.

Despite the impressive computational approaches used in the expedition, applying metagenomics to larger eukaryotic organisms remains a daunting task, in part because few reference genomes exist. Instead, metabarcoding approaches focus on extensive sampling of target genes, such as the gene encoding cytochrome oxidase c (COI), recently used to identify marine benthic animals (*13*). In this vein, de Vargas *et al.* (page 874) (*4*) focused their efforts on generating an extensive database of 18*S* ribosomal bar codes. They first compiled and generated a comprehensive family tree based on available data and then compared their ~2.3 million bar codes with sequences from different branches of the tree. They estimate that the sunlit regions of the ocean harbor ~150,000 DNA-based taxa of small eukaryotes (less than about 2 mm in size), again providing upper limits for taxonomic diversity and highlighting the difference from reference sequences. The authors find that diversity is greatest within three poorly known groups of unicellular eukaryotes: the Alveolata, Rhizaria, and Excavata. Each consists largely of parasites, phagotrophs (cells that engulf other cells), and symbionts. This observation provides strong evidence that organism interactions drive diversification in marine plankton.

Lima-Mendez *et al.* (page 874) (*5*) amplify the results of de Vargas *et al.* by providing statistical evidence that parasitic and viral interactions strongly affect population structure. They predict a potential network of interactions between specific organisms that they refer to as an interactome. Their visual confirmation of one such predicted symbiosis between a flatworm and a microalgal suggests that many more novel interactions are embedded in these networks.

Villar *et al.* (page 875) (*6*) used natural "seas within seas" to study how environmental changes affect marine communities over time and space. The fast-moving Agulhas current hugs the tip of South Africa before slamming into the Antarctic Circumpolar current, a collision that causes most of the Agulhas to turn back on itself to rejoin the Indian Ocean. A small proportion of the current continues westward, however, slipping around Africa into the South Atlantic in the form of giant rotating eddies—the Agulhas rings. These rings, with diameters of hundreds of kilometers, remain physically distinct from surrounding waters and are recognizable in satellite images as they slowly move across the Atlantic toward South America. Villar *et al.* document the

Exploring hidden ocean worlds, the modern way.

This map shows the route taken by the *Tara* Ocean expedition, which sampled microorganisms and small animals in the world's oceans between 2009 and 2012.

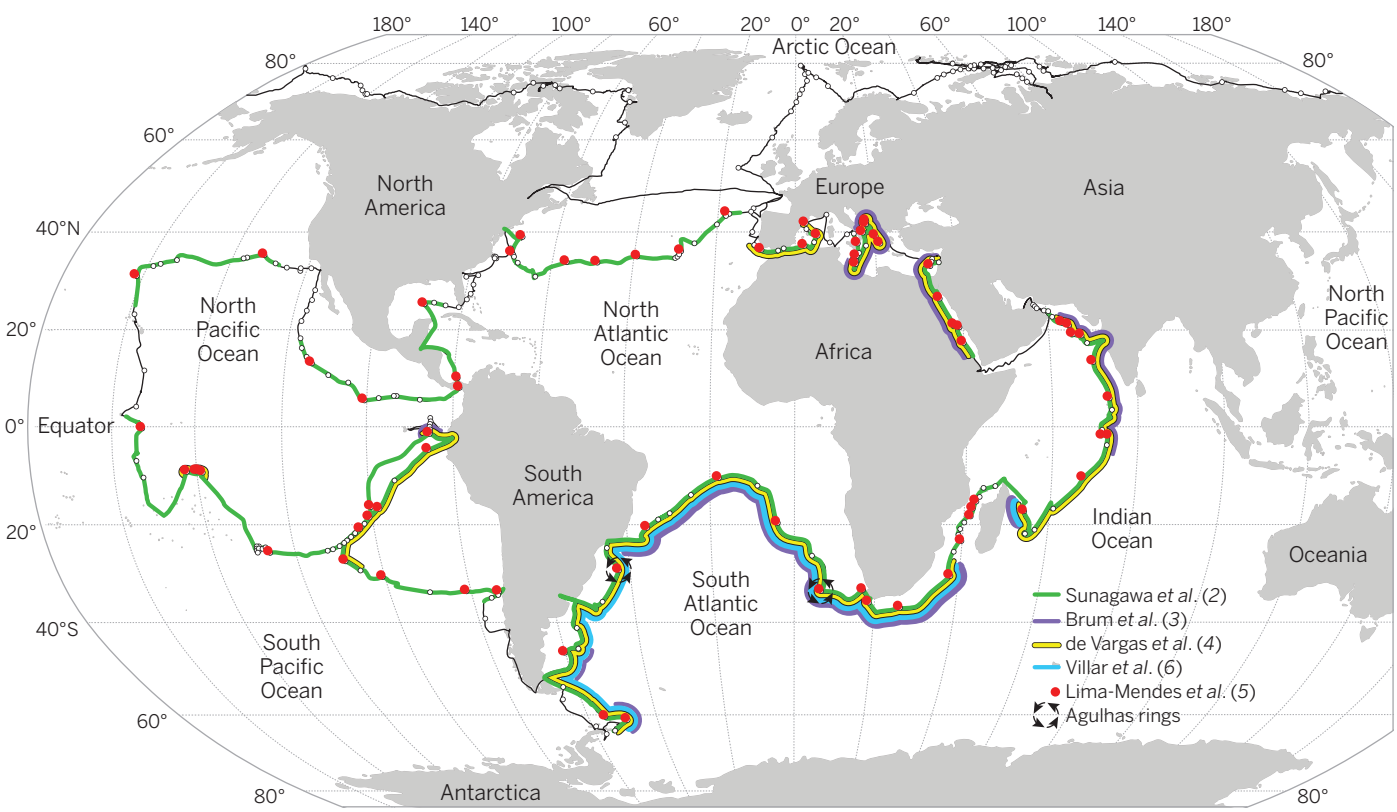


ILLUSTRATION: N. LEBESCOT AND C. DE VARGAS/EPEP-CNRS ROSCOFF

fate of the communities trapped in these rings. They show that the rings are leaky incubators, changing in oceanographic features and nutrient structure on their trek across the Atlantic. Taxonomic composition changes dramatically along the way, emphasizing the importance of environmental factors. As a result, the rings do not inject plankton from the Indian Ocean into the Atlantic Ocean as previously hypothesized (*14*).

The studies illustrate the exquisite complexity of marine ecosystems, in which microscopic organisms interact through competition for limiting resources, predator-prey and parasite-host dependencies, and cross-kingdom synergies. These interactions are embedded in a backdrop of fluctuating temperatures, light, and nutrient concentrations. In this world, viscous rather than gravitational forces dominate, allowing heterogeneous hotspots to develop along microscales (*15*). Some organisms compete for essential metabolites dissolved in the water, whereas others likely rely on various forms of trading alliances. Yet others simply surround small organic particles, consuming them as they drift to the ocean bottom as "marine snow." These diverse interactions across a large number of different species raise the question of whether coevolution acts largely between pairs of closely interacting species or on many species interacting within consortia.

The greatest challenge will be to uncover unifying principles behind these interactions. What key currencies are exchanged, how do organisms recognize one another, and when are abiotic conditions more important than species interactions in determining distribution and abundance? How and when should this complexity be incorporated into ecosystem models? When do these interactions affect climate tipping points for ecosystem function? The *Tara* Oceans expedition has generated a treasure trove of data available to anyone willing to dive in and start addressing these questions. ■

REFERENCES

1. F. Azam, F. Malfatti, *Nat. Rev. Microbiol.* **5**, 782 (2007).
2. S. Sunagawa *et al.*, *Science* **348**, 1261359 (2015).
3. J. R. Brum *et al.*, *Science* **348**, 1261498 (2015).
4. C. de Vargas *et al.*, *Science* **348**, 1261605 (2015).
5. G. Lima-Mendez *et al.*, *Science* **348**, 1262073 (2015).
6. E. Villar *et al.*, *Science* **348**, 1261447 (2015).
7. V. Iverson *et al.*, *Science* **335**, 587 (2012).
8. H. S. Yoon *et al.*, *Science* **332**, 714 (2011).
9. J. C. Venter *et al.*, *Science* **304**, 66 (2004).
10. C. Rinke *et al.*, *Nature* **499**, 431 (2013).
11. L. Zinger *et al.*, *PLOS ONE* **6**, e24570 (2011).
12. B. L. Hurwitz, M. B. Sullivan, *PLOS ONE* **8**, e57355 (2013).
13. M. Leray, N. Knowlton, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 2076 (2015).
14. P. Cermeño, P. G. Falkowski, *Science* **325**, 1539 (2009).
15. R. Stocker, *Science* **338**, 628 (2012).

10.1126/science.aaa7378

CANCER

Preprocancer

Normal skin harbors cancer-causing mutations

By **Douglas E. Brash**

Tumors have a life history (*1*, *2*). A tumor consists of a dominant cell clone containing mutations in key cancer genes called "drivers," together with smaller clones that descended from it but then diverged by accumulating different drivers. Other mutant clones appeared and vanished according to the selection pressures acting upon them or by neutral drift; the resulting diversity confers the means to escape clinical treatment (*3*). The stepwise accumulation of genetic and epigenetic alterations roughly parallels the clinical progression from normal tissue to precancer, cancer, and metastasis. It is widely assumed that driver mutations occur infrequently in long-lived lineages of cells (*2*), and that most arise in cancerous tissue that is too small to be clinically detectable. On page 880 of this issue, Martincorena *et al.* (*4*) overthrow both assumptions and reveal that sun-exposed normal skin is already a polyclonal quilt of driver mutations subjected to selection—a field of preprocancers, as it were—that nevertheless functions as a skin.

The prevailing model for the evolution of cancer is a reiterative process of clonal expansion, genetic diversification, and clonal selection. This multiple-genetic-hit model implies that normal tissue contains a few cells that have driver mutations, but not enough drivers in any one cell to create a cancer or precancer. Indeed, clones of cells with mutations in the tumor suppressor gene *P53* are found in histologically normal human skin and breast tissue; in skin, they constitute a surprising 4% of the epidermis (*5*, *6*).

Martincorena *et al.* used ultradeep sequencing technology to detect other mutations in normal skin. They examined 74 genes in eyelids, a sun-exposed tissue often removed in plastic surgery, and examined the spatial distribution of the mutations by subdividing four lids into 234 minibiopsies. It turns out that *P53* was not an outlier: The burden of mutations in normal skin, in just 74 genes, is about five mutations per megabase. This is only a factor of 10 less than seen in squamous cell carcinoma of the skin and is comparable to the mutation burden seen in cancers of the breast or head and neck. Most mutations in the biopsies have the characteristic signatures of exposure to ultraviolet light (cytosine-to-thymine changes at adjacent pyrimidine residues), and include nonsignature guanine-to-thymine changes that are caused by many mutagens, including ultraviolet light (*7*). Copy number changes are also present. To determine whether these mutations were conferring a selective advantage in normal skin, Martincorena *et al.* used the ratio of mutant sequence reads that do or do not alter the amino acid. Six genes—*NOTCH1*, *2*, and *3*, *P53*, *FAT1*, and *RBM10*—had an excess of nonsilent mutations, indicating that these genes had been selected for and are therefore presumptive drivers. The first five are known drivers for keratinocyte tumors and the last is mutated in other cancers. A seventh gene, *FGFR3*, had recurrent activating mutations. In total, one-quarter of middle-aged skin contains a mutation in one of these drivers.

"...sun-exposed normal skin is already a polyclonal quilt of driver mutations subjected to selection—a field of preprocancers, as it were—that nevertheless functions as a skin."

Whirring in the background of the results is Martincorena *et al.*'s bioinformatics effort to establish gold standard protocols for identifying valid mutations present at low levels. This is critical because a single-copy mutation present in 1% of the cells and sequenced to 500× depth provides on average only 2.5 reads. The supplemental material is an operating manual for identifying rare mutations.

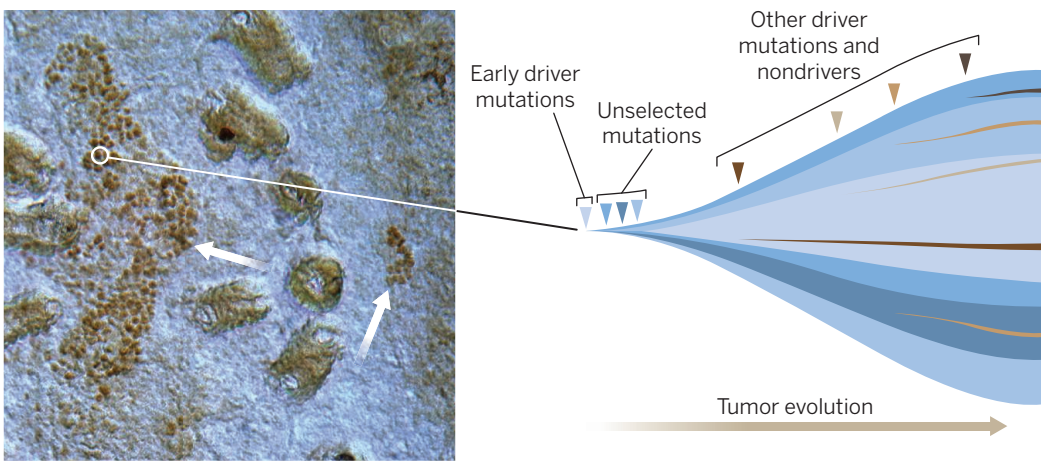
How did our eyelids get so many mutations? In DNA isolated from a homogenized biopsy, a mutant sequencing read could be frequent because its gene has mutation "hotspots" (highly susceptible to mutation) or its cell had a "mutator phenotype" (a mutation that increases the mutation rate at other loci) (*8*); either route would allow scattered cells to acquire the same mutation

Departments of Therapeutic Radiology and Dermatology, Yale School of Medicine, New Haven, CT, USA.
E-mail: douglas.brash@yale.edu

independently. Alternatively, the mutation could drive a single stem cell's clonal expansion (*9*). To investigate this point, Martincorena *et al.* calculated that, although some genes were mutated often in the eyelid, the frequency of any particular base change was so low that fewer than 0.2% of recurrent base changes could have arisen independently in the same biopsy. Hence, recurrent mutations must have been part of the same clone. The authors estimated the clone sizes using the variant allele fraction, after omitting sites with copy number variation near the mutation. For comparison, the descendants of a single fluorescently tagged normal mouse skin cell typically produce a clone of a dozen cells or less in a year (*10*). In the eyelid, most mutant clones occupy 1 to 10% of the ~1 mm^2 biopsies, whereas clones with driver mutations like *P53* and *FGFR3* average up to 0.7 mm^2. Given skin's ~10^6 nucleated cells/cm^2, the latter contain over 5000 cells. This was judged a modest increase over the average size of clones that did not show selection. But the true enhancement may be much greater because, at least for *P53* (*11*), most mutant clones are newly created or unselected clusters of 2 to 25 cells that eventually disappear by neutral drift; these will be missed when the read depth is "only" 500×.

The observed clone size is important because expanding from one mutant cell to 1000 is key to letting a tumor's life history progress to the next stage. DNA replication that allows DNA damage to be converted to a new mutation merely creates new mutations linearly as the number of cell divisions increases. In clonal expansion, mutant daughter cells each create a pair of mutant daughters, exponentially increasing the prevalence of mutant cells (*11*). Only one of this clone's cells needs to acquire the next driver (*12*). Numerically, this is key to making multiple-genetic-hit cancers: ultraviolet light–induced mutation frequencies are less than 10^{-4} per gene per cell division, so mutating four particular genes on both alleles in the same cell will happen in 10^{-32} of the 10^{10} cells of sun-exposed skin. Similar numbers hold for mutations induced in lung cancer by cigarette smoke, in liver by aflatoxin from moldy grain, or by mutator phenotypes in tissues not exposed to carcinogens. But expanding each new mutant to a 1000-cell clone makes the final number 0.1 cancer cell per person, a figure close to the human skin cancer incidence in sunny climes and which can be increased by several orders of magnitude after factoring in a lifetime of cell divisions or allelic loss instead of base changes. The new measurements of Martincorena *et al.* show that the clonal expansion strategy works distressingly well. In 1 cm^2 of normal skin, the authors found six clones each containing up to six driver mutations in the same cell.

A hurdle remains. Clonal expansion in solid tissue needs a special kind of mutation, dubbed a "primer mutation," which allows a mutant stem cell to expand beyond its normal domain into larger territory (*9*). The primer mutation needs to alter stem cell fates or confer enhanced survival relative to the cell's neighbors. A *P53* mutation confers both properties when, long after ultraviolet light's role as a mutagen is finished, cells are exposed to the selection pressure of chronic sunlight (*11*, *12*). It is therefore striking that the panopoly of clones in normal skin is dominated by five mutated stem cell fate genes. The NOTCH1, 2, and 3 receptors control the balance between keratinocyte self-renewal, proliferation, and differentiation; they also regulate survival (*13*). NOTCH expression is up-regulated by P53. P53 switches cells from exponential division resembling self-renewal to linear, asymmetric division (*14*). The cadherin-like FAT1 protein acts during fetal development. These five self-aggrandizing mutations increase the target size for acquiring the next genetic hit and transmit self-aggrandizing ability to that new mutant.



Cancer twists. (**Left**) Ultraviolet light–exposed normal skin is shown, containing a large and small clone of *P53*-mutant keratinocytes (dark nuclei among hair follicles, indicated with white arrows). Normal human skin is now revealed to contain additional cancer-causing mutations. (**Right**) The "big bang" model of cancer (*15*) proposes the rapid clonal expansion of a single cell that had accumulated multiple driver mutations (the first blue arrow) under selection pressure, spawning new subclones containing additional mutations (darker blue arrows), which are dispersed without selection early in the life of a tumor. Subclones accumulate later mutations independently (brown arrows).

Yet, the measurements of Martincorena *et al.* also indicate that priming does not always increase tumor risk. *FGFR3* generated the largest clones, but this gene is mutated in seborrheic keratoses, benign skin lesions that never spawn a cancer cell. Cancer is somehow prevented in that setting. Mutant normal skin may not be many steps removed from cancer. Recently, it was discovered that the most pervasive subclones in a tumor are born when the nascent tumor grows rapidly in a "big bang" event that expands the dominant clone, creates and expands subclones too quickly for selection to choose among them, and fragments these subclones across the growing tumor (*15*) (see the figure).

With normal skin producing variants and selecting on them just as skin tumors do, it's a wonder that we don't have more skin tumors. The conundrum is that a therapy targeting cells containing early driver mutations will erase a large fraction of non-cancerous skin. It may be preferable to act earlier, when the monoclonal sheet of the "freshly scrubbed face" of youth is rarely flecked with preprocancer. ■

REFERENCES

1. L. Foulds, *J. Chronic Dis.* **8**, 2 (1958).
2. S. Nik-Zainal *et al.*, *Cell* **149**, 994 (2012).
3. M. Greaves, C. C. Maley, *Nature* **481**, 306 (2012).
4. I. Martincorena *et al.*, *Science* **348**, 880 (2015).
5. A. S. Jonason *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 14025 (1996).
6. G. Deng, Y. Lu, G. Zlotnikov, A. D. Thor, H. S. Smith, *Science* **274**, 2057 (1996).
7. D. E. Brash, *Photochem. Photobiol.* **91**, 15 (2015).
8. L. A. Loeb, *Cancer Res.* **51**, 3075 (1991).
9. J. Cairns, *Nature* **255**, 197 (1975).
10. A. M. Klein, B. D. Simons, *Development* **138**, 3103 (2011).
11. A. M. Klein, D. E. Brash, P. H. Jones, B. D. Simons, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 270 (2010).
12. A. Ziegler *et al.*, *Nature* **372**, 773 (1994).
13. G. P. Dotto, *Oncogene* **27**, 5115 (2008).
14. J. L. Sherley, P. B. Stadler, D. R. Johnson, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 136 (1995).
15. A. Sottoriva *et al.*, *Nat. Genet.* **47**, 209 (2015).

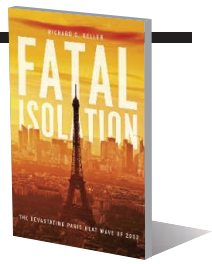
10.1126/science.aac4435

ANTHROPOLOGY

The dynamics of disaster

A social autopsy of the 2003 Paris heat wave

By **Laura Stark**



Fatal Isolation
The Devastating Paris Heat Wave of 2003
Richard C. Keller
University of Chicago Press, 2015. 250 pp.

In August 2003, hundreds of Parisians returned from their summer holidays to an unholy smell. Ascending the stairs in their apartment buildings, they found the source: dead bodies. Between August 1st and 20th, a heat wave baked Europe, and nearly 15,000 people died in France alone. Richard Keller's intrepid new book, *Fatal Isolation*, is a social autopsy of those deaths.

The heat wave was a tragedy in slow motion. In the first week of August 2003, government officials issued tepid warnings about the heat. French journalists mentioned the swelter only to wish middle-class readers a *bon voyage* on their August holiday. Hospital emergency rooms and morgues were overwhelmed by the second week of August, and medical workers described a health infrastructure pressed to its limits. Municipal ice rinks became acceptable venues in which to store dead bodies; newspapers published the names of unclaimed corpses with the hope that someone might retrieve them—and make space for more. When journalists finally began to cover the story, they quickly converged on a cause of death: the decline of social solidarity, exacerbated by government mismanagement. By autumn, spectacles of public remorse about the "forgotten" victims (and about everyone else's fabulous holidays) were *de rigueur*. Yet little ultimately changed, Keller argues, and in *Fatal Isolation*, he explains why.

The story of the 2003 heat wave has been told before but, as Keller shows, the victims have been remembered in odd and unhelpful ways. Media, government, and epidemiological accounts of the heat-wave deaths created an image of the "typical victim" in the public imagination. True, the bulk of the victims were elderly people who had few friends, distant family, and little contact with neighbors. Yet Keller argues that the aggregate and oft-repeated profile of the lonely elderly person gave an incomplete sense of the wide range of victims and, thus, an incomplete understanding of the problem—namely, the biases in how governments participate in the lives and deaths of their citizens.

Richard Keller is a well-regarded historian best known for his work on French colonial medicine in Africa. In *Fatal Isolation*, he welds the perspective of a historian to the tools of an anthropologist in an effort to crack the puzzle of how citizens who lived in a generous welfare state could be consistently and completely abandoned by governments organized to protect them. He interviewed neighbors, shopkeepers, policy-makers, and medical workers. He explored the burst of film and literary nonfiction that the heat wave prompted. He visited addresses of the "forgotten victims" and photographed the crude living conditions of people eking out a bare life in the City of Lights. The result is masterful. Keller synthesizes these disparate sources of information into an impressive new explanation of the heat-wave deaths. More broadly, he demonstrates how social status, not only geographical location, predicts survival during natural disasters.

Individuals sickened by the heat lie in the corridors of Saint Antoine Hospital in August 2003.

PHOTO: AP PHOTO/FRANCK PREVEL

Keller's research shows that heat-wave victims came to be portrayed as people at the margins of French life because of age, infirmity, or personal failings. The narratives crafted and repeated—by locals, by journalists, by policy-makers—cast the victims as individuals who "had withdrawn from society as a consequence of their actions—whether voluntarily or as a function of their erratic behavior, their madness, their addictions," Keller writes. "The rhetorical power of such portrayals is to redistribute culpability and to direct blame toward the victims themselves."

Keller is as likely to follow leads he found buried in the archive as those he found in the cemetery for unclaimed bodies on the outskirts of Paris. In doing so, he suggests new tools for a critical epidemiology of disasters. For example, whereas traditional epidemiology tends to map health in horizontal space—across neighborhoods, for instance—Keller's visits to the victims' homes prompt him to consider the vertical dimension of the problem as well. He finds, for example, that victims tended to live in simple rooms located on the highest stories of popular residential buildings throughout Paris, meaning the victims were the literal neighbors of many well-off urbanites. Moving beyond the basic observation that heat rises, Keller explains how these top-story apartments have historically been low-rent rooms, serving formerly as servants' quarters in Parisian residences. The heat wave produced death by urban design that was decades, not days, in the making.

Still, policy solutions are hard to come by for Keller. He argues that the French government was a major cause of the tragedy, and yet government is also his solution. Although Keller demurs on what precisely might be done—both in the immediate term and in the longer terms of climate change policy—*Fatal Isolation* makes clear that necessary changes will be as ordinary as they are profound.

10.1126/science.aab1097

The reviewer is at the Center for Medicine, Health, and Society, Vanderbilt University, Nashville, TN 37235, USA. E-mail: laura.stark@vanderbilt.edu

VIROLOGY

Exploring the unseen

A new "phage phield guide" sheds light on the habits and habitats of bacteriophages

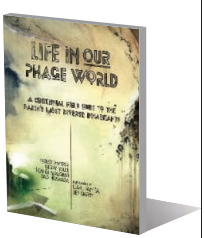
By **Michael Koeris**

Having grown up in the field of bacteriophage biology and launched a company enabled by the power of phages, I am an unabashed proponent of bacteriophages. For those who aren't familiar with them, bacteriophages—or phages—are the ubiquitous viruses that infect bacteria and archaea. They were discovered and described independently by the English bacteriologist Frederick Twort and the French-Canadian microbiologist Félix d'Herelle and have revolutionized our understanding of biology. If I sound like a gushing phage fanboy, that's only because I can't get over the fact that these little viruses effectively act like tiny biologists, bacterial taxonomists, and biochemists in order to survive and thrive in incredibly diverse ecosystems (and they look cool to boot).

Both bacteria and phages are incredibly abundant and omnipresent, but the diversity of bacteria is dwarfed by the huge diversity that exists in the phage world. An oft-cited estimate of their prevalence on earth—10^{31} particles—is enough to boggle the mind.

The reviewer is at Sample6, Cambridge, MA 02139, USA. E-mail: michael.koeris@sample6.com

Life in Our Phage World
A Centennial Field Guide to the Earth's Most Diverse Inhabitants
Forest Rohwer, Merry Youle, Heather Maughan, and Nao Hisakawa
Wholon, 2014. 413 pp.



This, of course, is not a static picture but highly dynamic: Phages replicate at a rate of 10^{24} infections per second (*1*). Large-scale metagenomic sequencing has revealed that phage genomes are most likely not linearly evolved but rather a hypermodular, ever-evolving patchwork. They interact in complicated ways with their environments and with the bacteria they exploit for replication.

Focusing on one representative phage in each chapter, *Life in Our Phage World* traces the infection strategies, replication mechanisms, and recent findings related to each class of phage. Made to look like an Audubon Society field guide from days past, the accompanying illustrations by Leah Pantéa and Ben Darby are a joy to view. Structurally accurate, the drawings imbue the phages (usually visualized via unmoving crystallography) with dynamic motion. Accompanying the illustrations is a brief summary of the phage's genome, encapsidation method, common hosts, habitat, and "lifestyle" (e.g., lytic or temperate). Each section also features a phylogenetic map, a geographic map depicting where the phage has been found and sampled, and both a high-level sketch of the genome and a more detailed, GenBank-style annotated map. Perhaps the only criticism I have for this book is the utility of the last map, which is too detailed for the generalist and not up-to-date or detailed enough for the specialist.

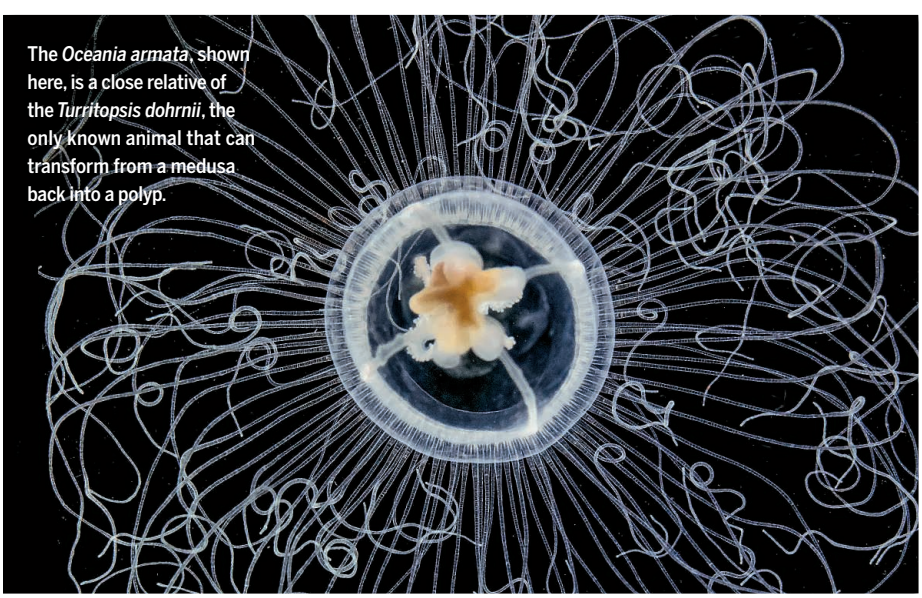
Not only do we learn how a bacteriophage without any motility can effectively navigate the environment (hint: landing gear down or up), the book also delves into the ongoing and unwinnable struggle between phages and bacteria, including all of the mechanisms that are used by each to confer temporary advantages.

With apologies to Douglas Adams, in *Life in Our Phage World*, Rohwer *et al.* have provided us with the 21st-century hitchhiker's guide to the (phage) universe. This book is a welcome refresher on phage complexity and diversity that would serve as an amazing resource for biology instructors at the high school, undergraduate, or graduate level. It is even accessible enough for the casual science aficionado to browse, enjoying a chapter here and there, as time permits.

REFERENCES

1. R. Hendrix, in J. F. Atkins, R. F. Gesteland, Eds., *Recoding: Expansion of Decoding Rules Enriches Gene Expression* (Springer-Verlag, New York, 2010), pp. 249–258.

10.1126/science.aaa9107

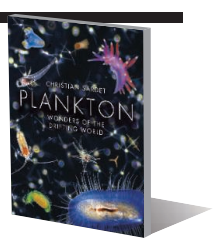


The *Oceania armata*, shown here, is a close relative of the *Turritopsis dohrnii*, the only known animal that can transform from a medusa back into a polyp.

PHOTO: CHRISTIAN SARDET

IN BRIEF

Plankton
Wonders of the Drifting World
Christian Sardet
University of Chicago Press, 2015. 222 pp.



"Take two breaths," begins oceanographer Mark Ohman. "For one of them, you can thank the plankton." In the introduction that follows, Christian Sardet, cofounder and scientific coordinator of the *Tara* Oceans Expedition, expands upon the important role that plankton play in our environment. Striking close-up photos and micrographs take center stage in the remaining chapters, revealing the dazzling diversity of these tiny creatures—from microscopic unicellular organisms to complex crustaceans.

10.1126/science.aac4683

Edited by **Jennifer Sills**

Eugenics lurk in the shadow of CRISPR

IN CALLING THEIR Perspective "A prudent path forward for genomic engineering and germline gene modification" (3 April, p. 36; published online 19 March), D. Baltimore *et al.* show at once the size of the problem and the modesty of their response to it. CRISPR-Cas9, invented by the ninth author, Jennifer Doudna, allows the alteration of specific DNA in the mammalian genome. The authors say that "CRISPR-Cas9 technology, as well as other genome engineering methods, can be used to change the DNA in the nuclei of reproductive cells that transmit information from one generation to the next (an organism's 'germ line')." This is a big deal. It means that we can imagine a day when human chromosomes may be modified in the sperm and egg to assure that one or another aspect of a child's inheritance is designed to order.

This is a huge departure from current understanding, but the authors are remarkably circumspect. They call for the convening of a "globally representative group of developers and users of genome engineering technology and experts in genetics, law, and bioethics, as well as members of the scientific community, the public, and relevant government agencies and interest groups, to further consider these important issues, and where appropriate, recommend policies."

That simply will not do. This opening to germline modification is, simply put, the opening of a return to the agenda of eugenics: the positive selection of "good" versions of the human genome and the weeding out of "bad" versions, not just for the health of an individual, but for the future of the species. I do not think their call is sufficient. Even in its inadequacy, I doubt it will be heeded by the six private corporations that are listed in the paper as supporting their research, nor by the universities listed as holding their patents on continuing CRISPR-Cas9 research.

Rational eugenics is still eugenics. The best in the world will not remove the pain from those born into a world of germ-line modification but who had not been given a costly investment in their gametes. They will emerge with the complexity of a genome different from what this technology will be able to define as "normal." I do not think anything short of a complete and total ban on human germline modification will do, to prevent this powerful force for rational medicine—one patient at a time—from becoming the beginning of the end of the simplest notion of each of us being "endowed by our Creator with certain inalienable rights."

Robert Pollack

Department of Biological Sciences, Columbia University, New York, NY 10027, USA. E-mail: pollack@columbia.edu

Eugenics on the horizon?

PHOTO: LAWRENCE LAWRY / SCIENCE SOURCE

Carnivore coexistence: Wilderness not required

OUR REPORT "RECOVERY of large carnivores in Europe's modern human-dominated landscapes" (19 December 2014, p. 1517) generated a series of Letters, published in the 23 January issue, concerning the importance of wilderness for large carnivore conservation ("Carnivore coexistence: Value the wilderness," J. J. Gilroy *et al.*, p. 382; "Carnivore coexistence: America's recovery," M. E. Gompper *et al.*, p. 382; "Carnivore coexistence: Trophic cascades," T. M. Newsome and W. J. Ripple, p. 383).

Gilroy *et al.* claim that the recovery of large carnivores in Europe is contingent on wilderness and protected areas. However, barely 13% of the European Natura 2000 network contains relatively undisturbed natural habitat (*1*), and the majority of protected areas in Europe are too small and isolated to house even single individuals, let alone sustain viable large carnivore populations (*2*). We by no means argue for a rollback on protected area designation or on the importance of conserving remaining wilderness. We simply argue that European carnivores are not among the species whose conservation depends on either of these conservation strategies.

In contrast to the claim made by Gilroy *et al.*, Swedish bears do not live in wilderness but in some of the most intensively harvested commercial forests in the world (*3*, *4*). Decades of bear hunting in Sweden have not precluded their recovery. Central European lynx populations are not generally linked to protected areas. The Bavaria-Bohemian lynx population is a rare exception (*5*). At their lowest demographic extent, wolves in Mediterranean countries persisted in human-dominated landscapes, and they have made a remarkable comeback to such landscapes in Germany (*6*). The high black bear densities in New Jersey cited in Gompper *et al.*'s Letter are another illustration of large carnivores' ability to coexist with people if they are allowed.

We agree that the apparent dichotomy between Europe's land sharing versus North America's land sparing may be primarily a legacy of the size difference in protected areas available between continents, and may even reflect a difference in rhetoric rather than practice. Where they exist, wilderness areas tend to play an important role as refuges and potential recovery nuclei for large carnivores, but claiming that such areas are a requirement for large carnivore recovery is not supported by the data.

We agree with Newsome and Ripple that preserving the ecological processes driven by large carnivores in human-dominated landscapes is challenging and requires further research on the functionality of the many different levels of completeness in which the ecological processes can be

OUTSIDE THE TOWER

Honing the climate change message

Five years ago, I scheduled my first meeting with a local official to discuss carbon emission reduction in China. I had planned a polished and persuasive argument. First, I presented the main findings from the IPCC Fourth Assessment Report: Climate Change 2007. I then illustrated the various scenarios and the possible turning points we may see under the Kyoto Protocol and beyond. I concluded that we should spring into action to develop a lower-carbon-emission strategy to address global climate change within the regional developmental policy system immediately.


Bing Xue

The official's reaction surprised me. I learned that those in government didn't feel that climate change was a priority. Rather, they were focused on sustaining local economic growth and maintaining socioeconomic stability. This was true despite sustainable development being a national strategy since 1994 (*1*) and the publication of China's first comprehensive policy initiative, China's National Climate Change Programme, in 2007 (*2*).

Based on this meeting, and the others that followed, I honed my message. Instead of emphasizing the local responsibilities in addressing global issues such as climate change, I now tell politicians that local efforts on reducing carbon emissions could lead to substantial cobenefits, such as reduction of local air pollutants (*3*), better economic performance (*4*), new economic growth areas, and job opportunities. I remind them that tax sources would be created by building low-carbon–oriented facilities.

There has always been a language gap between scientists and local officials, particularly in the field of sustainable development. To address climate change effectively, we must bridge that gap. I found that advocating for science was possible if I could articulate our shared goals.

Bing Xue

Institute of Applied Ecology, Chinese Academy of Sciences, 110016, Shenyang, China; Institute for Advanced Sustainability Studies (IASS), 14467, Potsdam, Germany; and Bohai University, 121013, Jinzhou, China. E-mail: xuebing@iae.ac.cn; bing.xue@iass-potsdam.de

REFERENCES

1. J. Liu, *Science* **328**, 50 (2010).
2. National Development and Reform Commission, "China's National Climate Change Programme," (2007); www.ccchina.gov.cn/WebSite/ccchina/Upfile/File188.pdf.
3. B. Xue *et al.*, *Ecol. Indic.* **38**, 272 (2014).
4. B. Xue *et al.*, *Renew. Sustainable Energy Rev.* **41**, 338 (2015).

Outside the Tower is an occasional feature highlighting science advocacy projects led by scientists and citizen scientists. How do you advocate for science? Tell us at submit2science.org.

declined (*7*). There is also no doubt that these processes will be very different in landscapes that are human-dominated. Allowing "nature its way" in an area undisturbed by humans is both important to conserve some elements of biodiversity and scientifically fascinating, but so is the ability of large carnivores to cope with human-dominated landscapes, which, like it or not, is a prerequisite for their survival in large parts of the modern world.

José Vicente López-Bao,[1,2]*

Petra Kaczensky,[3] John D. C. Linnell,[4] Luigi Boitani,[5] Guillaume Chapron[2]

[1]Research Unit of Biodiversity (UO/CSIC/PA), Oviedo University, Mieres, 33600, Spain. [2]Grimsö Wildlife Research Station, Swedish University of Agricultural Sciences, 73091, Riddarhyttan, Sweden. [3]Research Institute of Wildlife Ecology, University of Veterinary Medicine, Vienna, 1160, Vienna, Austria. [4]Norwegian Institute for Nature Research, Sluppen, 7485, Trondheim, Norway. [5]Department of Biology and Biotechnologies, University of Rome "La Sapienza," 00185, Roma, Italy.

*Corresponding author: jv.lopezbao@gmail.com

REFERENCES

1. European Commission Wilderness in Europe (http://ec.europa.eu/environment/nature/natura2000/wilderness/index_en.htm).
2. L. Santini, M. Di Marco, L. Boitani, L. Maiorano, C. Rondinini, *Divers. Dist.* **20**, 698 (2014).
3. C. Fries, O. Johansson, B. Pettersson, P. Simonsson, *Forest. Ecol. Manage.* **94**, 89 (1997).
4. T. F. Moe, J. Kindberg, I. Jansson, J. Swenson, *Can. J. Zool.* **85**, 518 (2007).
5. J. D. Linnell *et al.*, in *Reintroduction of Top-Order Predators*, M. W. Hayward, M. J. Somers, Eds. (Wiley-Blackwell, Chichester, UK, 2009), pp. 72–91.
6. G. Chapron *et al.*, *Science* **346**, 1517 (2014).
7. J. D. Linnell *et al.*, in *Large Carnivores and the Conservation of Biodiversity*, J. Ray, K. H. Redford, R. Steneck, J. Berger, Eds. (Island Press, Washington, DC, 2005), pp. 381–398.

TECHNICAL COMMENT ABSTRACTS

Comment on "Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 B.P."

Jade d'Alpoim Guedes, R. Kyle Bocinsky, Ethan E. Butler

Chen *et al.* (Reports, 16 January 2015, p. 248) argued that early Tibetan agriculturalists pushed the limits of farming up to 4000 meters above sea level. We contend that this argument is incompatible with the growing requirements of barley. It is necessary to clearly define past crop niches to create better models for the complex history of the occupation of the plateau.

Full text at http://dx.doi.org/10.1126/science.aaa4819

Response to Comment on "Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 B.P."

Guanghui Dong, Dongju Zhang, Xinyi Liu, Fengwen Liu, Fahu Chen, Martin Jones

Guedes *et al.* have drawn attention to a mismatch between the predictions of their "thermal niche model" and the records we have published of early barley finds in the northeastern Tibetan Plateau. Here, we consider how that mismatch usefully draws our attention to the additional variables that may account for it—namely, variations in genetic expression and agricultural practice.

Full text at http://dx.doi.org/10.1126/science.aaa7573

TECHNICAL COMMENT

ASIAN ARCHAEOLOGY

Comment on "Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 B.P."

Jade d'Alpoim Guedes,[1]* R. Kyle Bocinsky,[1] Ethan E. Butler[2]

Chen *et al*. (Reports, 16 January 2015, p. 248) argued that early Tibetan agriculturalists pushed the limits of farming up to 4000 meters above sea level. We contend that this argument is incompatible with the growing requirements of barley. It is necessary to clearly define past crop niches to create better models for the complex history of the occupation of the plateau.

In recent years, there has been much interest in understanding the mechanisms by which humans adapted agricultural subsistence patterns to high-altitude environments on the Tibetan Plateau and in the Andes (*1*–*8*). Chen *et al*. (*1*) bring important new data derived from the northeastern Tibetan Plateau (NETP) to bear on this issue.

Ecological factors can place heavy constraints on humans, particularly in areas of high altitude, because even small changes in temperature, precipitation, and land cover can have a major effect on what crops can be grown. Understanding what factors influence crop growth is key to producing more realistic models of past human behavior. Early studies of the mechanisms underlying the spread of wheat and barley considered a single aspect of crop growth patterns: length of the growing season (*2*). They argued that this factor slowed the spread of wheat and barley (*2*, *3*) and facilitated the spread of millets (*2*). A more complete estimate of agricultural potential can be constructed using thermal niche modeling (*4*, *5*, *8*, *9*). In particular, models based on a crop's accumulated heat requirements—growing degree days (GDD)—predict that wheat and barley are more adapted to growth in high-latitude and high-altitude Eurasia than millets (*4*, *5*). Millets were able to flourish only in select niches on the southeastern Tibetan Plateau (SETP) and only during the warmer Holocene climatic optimum [although even here, models indicate that their potential success was low (*5*)]. In contrast, the frost tolerance and lower GDD requirements of wheat and barley enabled these crops to be rapidly adopted as staples on the Tibetan Plateau and its margins (*4*, *5*, *8*). Local ecology and climate, coupled with crop phenology, thus had a marked impact on crop adoption in high-altitude environments, one that differed substantially from the lowlands (*10*).

New data from the NETP provide strong support for the important role played by barley in facilitating agricultural practice on the Tibetan Plateau (*1*). However, as with politics, all agriculture is local and depends on specific local conditions. Thus, a clear understanding of constraints on crop growth in any given locale is crucial to the development of archaeological models for human behavior (*4*, *9*). Examination of the thermal niche occupied by barley on the NETP (Fig. 1) shows that growing barley at an altitude of 4000 m above sea level (masl)—a claim made in Chen *et al*. (*1*)—was unlikely to have been successful in the average year. In the NETP, they could be reliably grown only up to 3000 masl, and the highest site from which they document barley remains is at 3341 masl [table S1 in (*1*)]. Although it is possible that climatic variations could have increased the elevation at which crops could be grown in certain years or that there were particularly favorable micro-environments in otherwise inhospitable zones, the social memory of crop failure and the risks associated with it would have likely driven humans away from farming in such environments.

Furthermore, seeds can arrive on archaeological sites via a wide range of mechanisms; their presence at a site should not be taken as evidence of in situ agriculture. Both hunter-gatherers and pastoralists can be involved in complex patterns of exchange with agriculturalists, trading resources available to them for domestic grain [e.g., (*11*)]. Finally, agriculturalists and agropastoralists can reside in areas well outside of where crops can be grown, while still cultivating fields within an agriculturally productive niche.

Until recently, hunter-gatherers have been relatively invisible in discourses on agricultural and pastoral origins in China, even though evidence shows that they were present on the plateau well into the late Holocene [e.g., (*12*)]. The potential for both hunter-gatherers and pastoralists to move grains outside of the cultivated niche means that caution should be exercised when discussing when and where agriculture was carried out on the Tibetan Plateau and with what permanence it was occupied. There is a need not only to develop

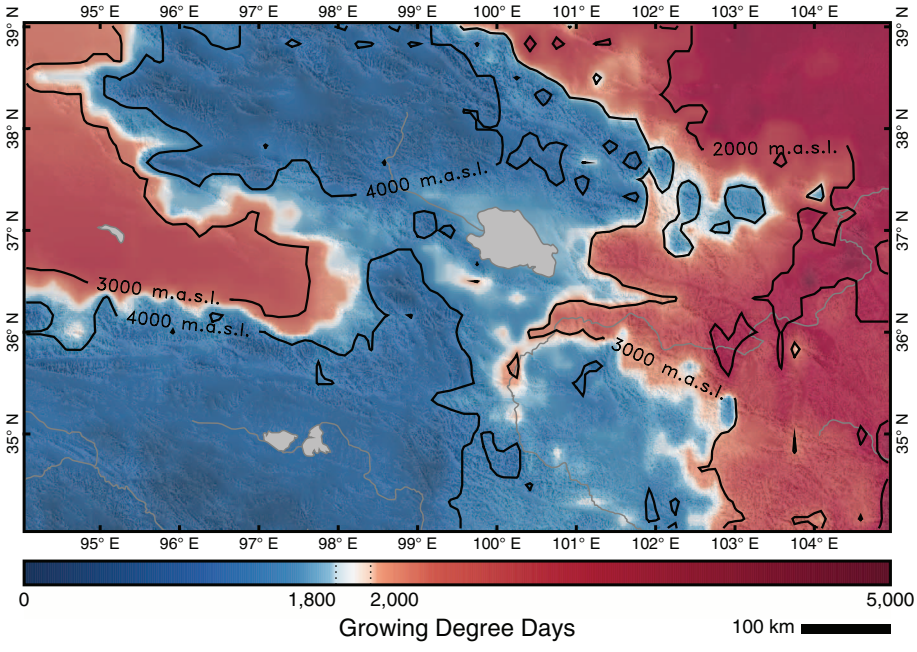


Fig. 1. Thermal growing niche of barley based on accumulated heat requirements (GDD) in the NETP. A GDD base temperature of 0°C is used for barley. Barley can be grown with certainty in regions shaded red. Regions covered by different minimal estimates of growing conditions are shaded white. Barley cannot be grown in the regions shaded blue. The black contours denote elevation above sea level. Lakes and major rivers are shown in gray. [Figure prepared using methods from (*4*) and data from (*14*) and (*15*)]

[1]Department of Anthropology, Washington State University, Pullman, WA 99164, USA. [2]Department of Earth and Planetary Science, Harvard University, Cambridge, MA 02138, USA.
*Corresponding author. E-mail: jade.dalpoimguedes@wsu.edu

the tools to identify and characterize more mobile populations but also to view these populations as important and active actors in the spread of agriculture and pastoralism throughout East Asia (*13*). Disentangling how agricultural crops and pastoral animals spread onto and across the Tibetan Plateau requires not only better evidence and robust models upon which this evidence may be brought to bear but also more complex narratives to interpret the mismatch between archaeological data and models for human behavior.

REFERENCES

1. F. H. Chen *et al*., *Science* **347**, 248–250 (2015).
2. M. Jones *et al*., *World Archaeol.* **43**, 665–675 (2011).
3. C. An, W. Dong, H. Li, Y. Chen, L. Barton, *Quat. Sci. Rev.* **81**, 148–149 (2013).
4. J. d'Alpoim Guedes, E. Butler, *Quat. Int.* **349**, 29–41 (2014).
5. J. d'Alpoim Guedes, thesis, Harvard University, Cambridge, MA (2013).
6. J. R. Harlan, in *Man, Settlement and Urbanism*, P. J. Ucko, R. Tringham, G. W. Dimbleby, Eds. (Duckworth, London, 1972), pp. 239–243.
7. R. W. Jamieson, M. B. Sayre, *Anthropol. Archaeol.* **29**, 208–218 (2010).
8. J. d'Alpoim Guedes *et al*., *Archaeol. Anthropol. Sci.* **6**, 255–269 (2014).
9. R. K. Bocinsky, T. A. Kohler, *Nat. Comm.* **5**, 6618 (2014).
10. N. Boivin, D. Fuller, A. Crowther, *World Archaeol.* **44**, 452–469 (2012).
11. T. Headland, L. Reid, *Curr. Anthropol.* **30**, 43–66 (1989).
12. P. J. Brantingham *et al*., *Geoarchaeology* **28**, 413–431 (2013).
13. G. J. Stein, *Am. Anthropol.* **104**, 903–916 (2002).
14. M. J. Menne *et al*., Global Historical Climatology Network - Daily (GHCN-Daily), Version 3. 2. NOAA National Climatic Data Center (2012).
15. A. Jarvis, H. I. Reuter, A. Nelson, E. Guevara, Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database (2008); http://srtm.csi.cgiar.org.

12 December 2014; accepted 3 March 2015
10.1126/science.aaa4819

TECHNICAL RESPONSE

ASIAN ARCHAEOLOGY

Response to Comment on "Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 B.P."

Guanghui Dong,[1]* Dongju Zhang,[1] Xinyi Liu,[3]* Fengwen Liu,[1] Fahu Chen,[1] Martin Jones[2]

Guedes *et al*. have drawn attention to a mismatch between the predictions of their "thermal niche model" and the records we have published of early barley finds in the northeastern Tibetan Plateau. Here, we consider how that mismatch usefully draws our attention to the additional variables that may account for it—namely, variations in genetic expression and agricultural practice.

We welcome the interest shown by Guedes *et al*. in our paper on human adaption to "the roof of the world" (*1*) and the introduction of their "thermal niche model" into the debate (*2*). As we write, thousands of Tibetan farmers look forward to this coming year's barley harvest, as they have all their lives, at altitudes at which Guedes's *et al*. model predicts that barley cultivation is not sustainable. Modern barley cultivation covers the whole arable region of the Tibetan Plateau, from 1000 m above sea level (masl) up to 4750 masl, but mainly distributed from 3000 to 4000 masl on the plateau (*3*, *4*) (Fig. 1). The earliest historical records of barley cultivation on the Tibetan Plateau are in the Tang dynasty (618 to 907 C.E.) (*5*); in Qaidam Basin higher than 3000 masl regions, no later than the Qing dynasty (1636 to 1911 C.E.) (*6*); in the Yushu region higher than 4200 masl, in the early 20th century (*7*); in western Sichuan province (Luhuo) higher than 3500 masl and the western plateau (A Li region) higher than 4000 masl, no later than the Qing dynasty (*8*, *9*); and in the southern plateau higher than 4400 masl—Sajia and Yamdrok Lake—in the Yuan (1271 to 1368 C.E.) and Qing dynasties, respectively (*9*) (Fig. 1). Archaeological studies show that barley cultivation widely appeared on the plateau higher than 3000 masl regions or even higher than 4000 masl in the southern plateau as early as 3600 years ago (*1*), and at least 3000 years ago in the Nuomuhong region on the edge of Qaidam Basin (Fig. 1) higher than 3000 masl, evidenced by rich and diverse crop remains [see the supplementary materials in (*1*)].

The mismatch between the predictions of Guedes's *et al*. model and the actual growth of barley today and in the past should not, however, be taken as negating the value of the model; scientific models are often at their most productive

[1]Key Laboratory of Western China's Environmental Systems (Ministry of Education), Lanzhou University, Lanzhou 730000, China. [2]McDonald Institute of Archaeological Research, University of Cambridge, Cambridge CB2 3ER, UK. [3]Department of Anthropology, Washington University in St. Louis, St. Louis, MO 63130, USA.
*Corresponding author. E-mail: ghdong@lzu.edu.cn (G.D.); liuxinyi@wustl.edu (X.L.)

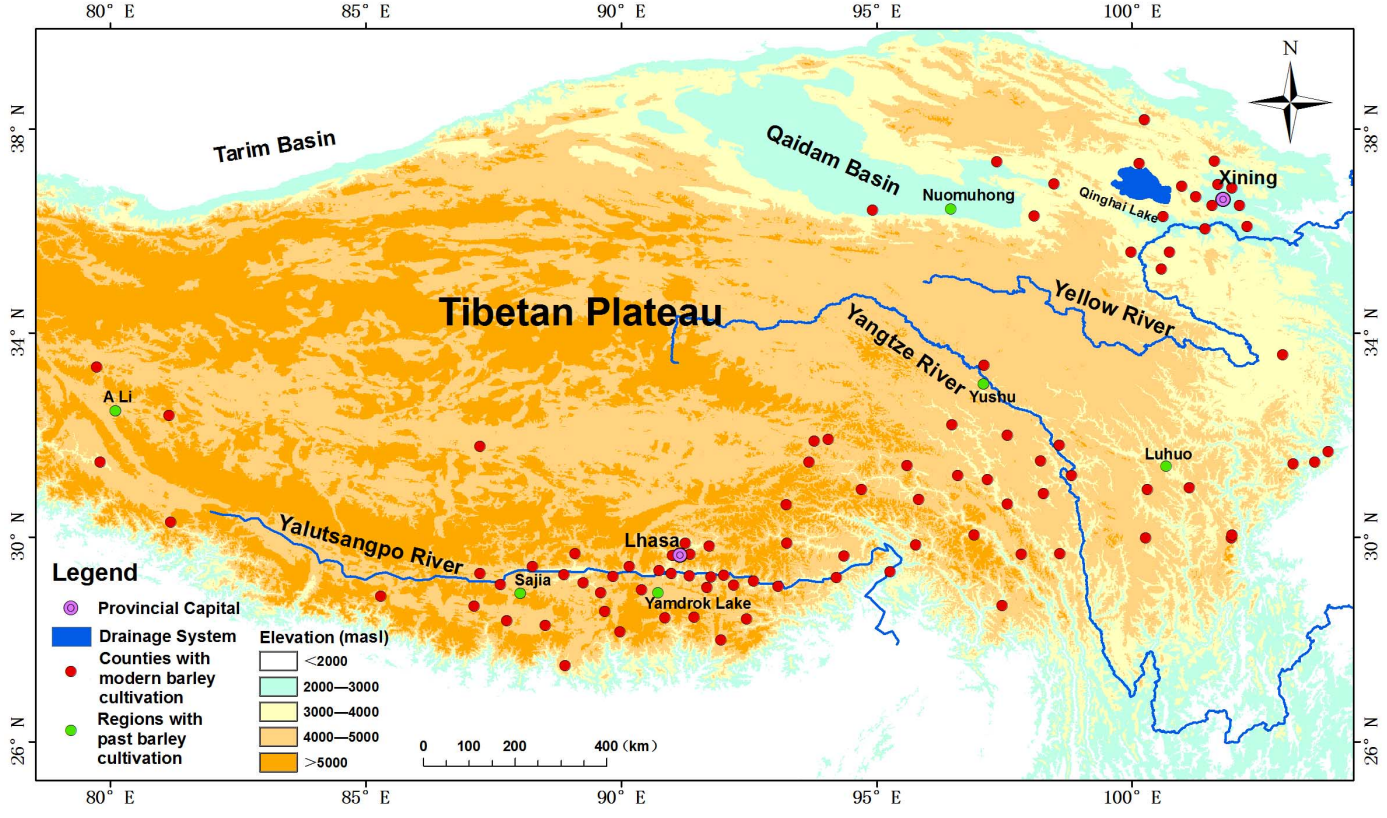


Fig. 1. The modern and past barley cultivation on the Tibetan Plateau. Red circles show locations where barley cultivations are documented since 1949. Green circles show locations of barley cultivation documented in historical records.

when they do not fit the observable data. In these cases, they force us to consider different parameters, not factored into the extant model, that may nonetheless be substantially influencing the data. In this case, we can follow Guedes's *et al*. barley data to its source and reflect on what those further variables and parameters might be.

Alongside a series of equivalent sources for wheat, rice, and millet, the principal data source for barley used in Guedes and Butler 2014 (*10*) is Stewart and Dwyer 1987 (*11*). The latter describes phenotypic observations of 192 *Hordeum vulgare* plants grown in pots in a Canadian greenhouse. Two observations can be made on this study: First, the variety of *H. vulgare* is not specified; second, while preliminary results of a field trial at one location were promising, the authors had not extended their study to field conditions. We can thus reflect on whether within-crop genetic diversity and field conditions (particularly cultivation practice) are the different parameters in question. There is a considerable body of recent research into altitudinal range and Ethiopian barley, examining traits and their expression at a range of altitudes up to 3300 masl. This research provides substantial evidence of diversity in a range of traits in relation to adaption to altitude (*12*–*14*). In terms of cultivation practice, two key well-documented variables can serve to modify and mitigate ambient thermal and seasonal environments—namely, flexibility in sowing times, and cultivation depth and strategy (*3*, *4*, *15*).

Guedes *et al*. also make observations on millet cultivation, to which similar considerations may apply. Although we have here drawn attention to the mismatch between their model output and observable cereal growth today and in the past, we do not discount the potential utility of the model. That utility is in drawing attention to factors that may account for the mismatch. In this instance, the likely factors are diversity in genetic expression and in field cultivation techniques. Both those factors are worthy of further scientific enquiry.

REFERENCES AND NOTES

1. F. H. Chen *et al*., *Science* **347**, 248–250 (2015).
2. J. d'Alpoim Guedes, R. K. Bocinsky, E. Butler, *Science* **348**, 872-b (2015).
3. S. J. Hu, *An Introduction to Agriculture of Tibet* (Sichuan Publishing House of Science and Technology, Chengdu, 1995).
4. The Editorial Committee of Chinese Agriculture Book, *Zhongguo Nongye Quanshu, Qinghaijuan* (*Chinese Agriculture Book, Qinghai Province*) (Chinese Agriculture Press, Beijing, 2001).
5. X. Liu, *Jiu Tang Shu, Liezhuan 146* (*Old Book of Tang, Ranked Biographies 146*) (compiled between 941 and 945 A.D.) (Zhonghua Book Company, Beijing, 1975).
6. *Qingshizong Shilu*, 102 and 125 (*Veritable Records of Shizong Emperor, Qing Dynasty*, **102** and **125**) (compiled in Qing Dynasty) (Zhonghua Book Company, Beijing, 1985).
7. Y. H. Cui, *Qinghai Jingjishi, Jindaijuan* (*The Recent Economic History of Qinghai Province*) (Qinghai People's Publishing House, Xining, 1998).
8. Z. K. Li, *Luhuotun Zhi Lue* (*Record of Luhuo County*) (Library of Minzu University of China, Beijing, 1979).
9. C. H. Chen, *Xizang Difang Jingjishi* (*The Local Economic History of Tibet*) (Ganshu People's Publishing House, Xining, 2008).
10. J. d'Alpoim Guedes, E. E. Butler, *Quat. Int.* **349**, 29–41 (2014).
11. D. W. Stewart, L. M. Dwyer, *Agric. For. Meteorol.* **39**, 37–48 (1987).
12. J. M. M. Engels, *Genet. Resour. Crop Evol.* **41**, 67–73 (1994).
13. T. Tanto Hadado, D. Rau, E. Bitocchi, R. Papa, *BMC Plant Biol.* **10**, 121–141 (2010).
14. T. D. Abebe, A. M. Bauer, J. Léon, *Hereditas* **147**, 154–164 (2010).
15. N. J. Z. Sou, L. Q. Liu, trans., *Xizang Wangtong Ji* (*The Clear Mirror: A Royal Genealogy*) (Minzu Publishing House, Beijing, 2000).

ACKNOWLEDGMENTS

This research was funded by the National Natural Science Foundation of China (grant nos. 41271218 and 41101087), the 111 Program (grant no. B06026) of the Chinese State Administration of Foreign Experts Affairs, and the European Research Council research project Food Globalisation in Prehistory (grant no. 249642).

6 February 2015; accepted 3 March 2015
10.1126/science.aaa7573

Tara Oceans studies plankton at PLANETARY SCALE

By P. Bork,[1] C. Bowler,[2] C. de Vargas,[3] G. Gorsky,[4] E. Karsenti,[5] P. Wincker[6]

The ocean is the largest ecosystem on Earth, and yet we know very little about it. This is particularly true for the plankton that inhabit the ocean. Although these organisms are at least as important for the Earth system as the rainforests and form the base of marine food webs, most plankton are invisible to the naked eye and thus are largely uncharacterized. To study this invisible world, the multinational *Tara* Oceans consortium, with use of the 110-foot research schooner *Tara*, sampled microscopic plankton at 210 sites and depths up to 2000 m in all the major oceanic regions during expeditions from 2009 through 2013 (*1*).

Success depended on collaboration between scientists and the *Tara* Expeditions logistics team. The journey involved not only science but also outreach and education as well as negotiation through the shoals of legal and political regulations, funding uncertainties, threats from pirates, and unpredictable weather (*2*). At various times, journalists, artists, and teachers were also on board. Visitors included Ban Ki-moon (Secretary-General of the United Nations) and numerous youngsters, including schoolchildren from the favelas in Rio de Janeiro.

Research schooner *Tara* supported a multinational, multidisciplinary team in sampling plankton ecosystems around the world.

Sampling, usually 60 hours per site, followed standardized protocols (*3*) to capture the morphological and genetic diversity of the entire plankton community from viruses to small zooplankton, covering a size range from 0.02 μm to a few millimeters, in context with physical and chemical information. Besides the sampling, a lab on board contained a range of online instruments and microscopes to monitor the content of the samples as they were being collected. The main focus was on the organism-rich sunlit upper layer of the ocean (down to 200 m), but the twilight zone below was also sampled. Guided by satellite and in situ data, scientists sampled features such as mesoscale eddies, upwellings, acidic waters, and anaerobic zones, frequently in the open ocean. In addition to being used for genomics and oceanography, many samples were collected for other analyses, such as high-throughput microscopy imaging and flow cytometry. The samples and data collected on board were archived in a highly structured way to enable extensive data processing and integration on land (*4*). The five Research Articles in this issue of *Science* describe the samples, data, and analysis from *Tara* Oceans (based on a data freeze from 579 samples at 75 stations as of November 2013).

De Vargas *et al.* used ribosomal RNA gene sequences to profile eukaryotic diversity in the photic zone. This taxonomic census shows that most biodiversity belongs to poorly known lineages of uncultured heterotrophic single-celled protists. Sunagawa *et al.* used metagenomics to study viruses, prokaryotes, and picoeukaryotes. They established a catalog with >40 million genes and identified temperature as the driver of photic microbial community composition. Brum *et al.*, by sequencing and electron microscopy, found that viruses are diverse on a regional basis but less so on a global basis. The viral communities are passively transported by oceanic currents and structured by local environments. Lima-Mendez et al. modeled interactions between viruses, prokaryotes, and eukaryotes. Regional and global parameters refine resulting networks. Villar *et al.* studied the dispersal of plankton as oceanic currents swirl around the southern tip of Africa, where the Agulhas rings are generated. Vertical mixing in the rings drives nitrogen cycling and selects for specific organisms.

Tara Oceans combined ecology, systems biology, and oceanography to study plankton in their environmental context. The project has generated resources such as an ocean microbial reference gene catalog; a census of plankton diversity covering viruses, prokaryotes, and eukaryotes; and methodologies to explore interactions between them and their integration with environmental conditions. Although many more such analyses will follow, life in the ocean is already a little less murky than it was before.

REFERENCES

1. E. Karsenti *et al.*, *PLOS Biol.* **9**, e1001177 (2011).
2. E. Karsenti, *Mol. Sys. Biol.* **15**, 6271 (2015).
3. S. Pesant *et al.*, *BioRXiv*, http://biorxiv.org/content/early/2015/05/08/019117 (2015).
4. S. Sunagawa *et al.*, *Mol. Sys. Biol.* **15**, 6272 (2015).

10.1126/science.aac5605

[1]*Structural and Computational Biology, European Molecular Biology Laboratory (EMBL), Meyerhofstraße 1, 69117 Heidelberg, Germany.* [2]*Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, F-75005 Paris, France.* [3]*CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.* [4]*CNRS, UMR 7093 Laboratoire d'Océanographie de Villefranche-sur-Mer (LOV), Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France.* [5]*Directors' Research, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany.* [6]*Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Génomique, GENOSCOPE, 2 Rue Gaston Crémieux, 91000 Evry, France.*

OCEAN PLANKTON

Structure and function of the global ocean microbiome

Shinichi Sunagawa,[1*†] Luis Pedro Coelho,[1*] Samuel Chaffron,[2,3,4*] Jens Roat Kultima,[1] Karine Labadie,[5] Guillem Salazar,[6] Bardya Djahanschiri,[1] Georg Zeller,[1] Daniel R. Mende,[1] Adriana Alberti,[5] Francisco M. Cornejo-Castillo,[6] Paul I. Costea,[1] Corinne Cruaud,[5] Francesco d'Ovidio,[7] Stefan Engelen,[5] Isabel Ferrera,[6] Josep M. Gasol,[6] Lionel Guidi,[8,9] Falk Hildebrand,[1] Florian Kokoszka,[10,11] Cyrille Lepoivre,[12] Gipsi Lima-Mendez,[2,3,4] Julie Poulain,[5] Bonnie T. Poulos,[13] Marta Royo-Llonch,[6] Hugo Sarmento,[6,14] Sara Vieira-Silva,[2,3,4] Céline Dimier,[10,15,16] Marc Picheral,[8,9] Sarah Searson,[8,9] Stefanie Kandels-Lewis,[1,17] *Tara* Oceans coordinators‡ Chris Bowler,[10] Colomban de Vargas,[15,16] Gabriel Gorsky,[8,9] Nigel Grimsley,[18,19] Pascal Hingamp,[12] Daniele Iudicone,[20] Olivier Jaillon,[5,21,22] Fabrice Not,[15,16] Hiroyuki Ogata,[23] Stephane Pesant,[24,25] Sabrina Speich,[26,27] Lars Stemmann,[8,9] Matthew B. Sullivan,[13§] Jean Weissenbach,[5,21,22] Patrick Wincker,[5,21,22] Eric Karsenti,[10,17†] Jeroen Raes,[2,3,4†] Silvia G. Acinas,[6†] Peer Bork[1,28†]

Microbes are dominant drivers of biogeochemical processes, yet drawing a global picture of functional diversity, microbial community structure, and their ecological determinants remains a grand challenge. We analyzed 7.2 terabases of metagenomic data from 243 *Tara* Oceans samples from 68 locations in epipelagic and mesopelagic waters across the globe to generate an ocean microbial reference gene catalog with >40 million nonredundant, mostly novel sequences from viruses, prokaryotes, and picoeukaryotes. Using 139 prokaryote-enriched samples, containing >35,000 species, we show vertical stratification with epipelagic community composition mostly driven by temperature rather than other environmental factors or geography. We identify ocean microbial core functionality and reveal that >73% of its abundance is shared with the human gut microbiome despite the physicochemical differences between these two ecosystems.

Microorganisms are ubiquitous in the ocean environment, where they play key roles in biogeochemical processes, such as carbon and nutrient cycling (*1*). With an estimated 10^4 to 10^6 cells per milliliter, their biomass, combined with high turnover rates and environmental complexity, provides the grounds for immense genetic diversity (*2*). These microorganisms, and the communities they form, drive and respond to changes in the environment, including climate change–associated shifts in temperature, carbon chemistry, nutrient and oxygen content, and alterations in ocean stratification and currents (*3*).

With recent advances in community DNA shotgun sequencing (metagenomics) and computational analysis, it is now possible to access the taxonomic and genomic content (microbiome) of ocean microbial communities and, thus, to study their structural patterns, diversity, and functional potential (*4*, *5*). The *Sorcerer II* Global Ocean Sampling (GOS) expedition, for example, collected, sequenced, and analyzed 6.3 gigabases (Gb) of DNA from surface-water samples along a transect from the Northwest Atlantic to the Eastern Tropical Pacific (*6*, *7*) but also indicated that the vast majority of the global ocean microbiome still remained to be uncovered (*7*). Nevertheless, the GOS project facilitated the study of surface picoplanktonic communities from these regions by providing an ocean metagenomic data set to the scientific community. Several studies have demonstrated that such data could, in principle, identify relationships between gene functional compositions and environmental factors (*8*–*10*). However, an extended breadth of sampling (e.g., across depth layers, domains of life, organismal-size classes, and around the globe), combined with in situ measured environmental data, could provide a global context and minimize potential confounders.

To this end, *Tara* Oceans systematically collected ~35,000 samples for morphological, genetic, and environmental analyses using standardized protocols across multiple depths at global scale, aiming to facilitate a holistic study on how environmental factors and biogeochemical cycles affect oceanic life (*11*). Here we report the initial analysis of 243 ocean microbiome samples, collected at 68 locations representing all main oceanic regions (except for the Arctic) from three depth layers, which were subjected to metagenomic Illumina sequencing. By integrating these data with those from publicly available ocean metagenomes and reference genomes, we assembled and annotated a reference gene catalog, which we use in combination with phylogenetic marker genes (*12*, *13*) to derive global patterns of functional and taxonomic microbial community structures. The vast majority of genes uncovered in *Tara* Oceans samples had not previously been identified, with particularly high fractions of novel genes in the Southern Ocean and in the twilight, mesopelagic zone. By correlating genomic and environmental features, we infer that temperature, which we decoupled from dissolved oxygen, is the strongest environmental factor shaping microbiome composition in the sunlit, epipelagic ocean layer. Furthermore, we define a core set of gene families that are ubiquitous in the ocean and differentiate variable, adaptive functions from stable core functions; the latter are compared between ocean depth layers and to those in the human gut microbiome.

Ocean microbial reference gene catalog

To capture the genomic content of prevalent microbiota across major oceanic regions (Fig. 1A), *Tara* Oceans collected seawater samples within the epipelagic layer, both from the surface water and the deep chlorophyll maximum (DCM) layers, as well as the mesopelagic zone (*14*). From 68 selected locations, 243 size-fractionated samples targeting organisms up to 3 μm [virus-enriched fraction (<0.22 μm): n = 45; girus/prokaryote-enriched fractions (0.1 to 0.22 μm, 0.22 to 0.45 μm,

[1]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [2]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [3]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [4]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [5]CEA–Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [6]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [7]Sorbonne Universités, UPMC, Université Paris 06, CNRS-IRD-MNHN, LOCEAN Laboratory, 4 Place Jussieu, 75005 Paris, France. [8]CNRS, UMR 7093, Laboratoire d'Océanographie de Villefranche-sur-Mer, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [9]Sorbonne Universités, UPMC Université Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [10]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, F-75005 Paris, France. [11]Laboratoire de Physique des Océans UBO-IUEM, Place Copernic 29820 Plouzané, France. [12]Aix Marseille Université CNRS IGS UMR 7256, 13288 Marseille, France. [13]Department of Ecology and Evolutionary Biology, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA. [14]Department of Hydrobiology, Federal University of São Carlos (UFSCar), Rodovia Washington Luiz, 13565-905 São Carlos, São Paulo, Brazil. [15]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [16]Sorbonne Universités, UPMC Université Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [17]Directors' Research, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [18]CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [19]Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [20]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [21]CNRS, UMR 8030, CP5706, Evry, France. [22]Université d'Evry, UMR 8030, CP5706, Evry, France. [23]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan. [24]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [25]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [26]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France. [27]Laboratoire de Physique des Océans UBO-IUEM, Place Copernic, 29820 Plouzané, France. [28]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany.

*These authors contributed equally to this work **†Corresponding author. E-mail: sunagawa@embl.de (S.S.); karsenti@embl.de (E.K.); jeroen.raes@vib-kuleuven.be (J.R.); sacinas@icm.csic.es (S.G.A.); bork@embl.de (P.B.)** ‡*Tara* Oceans coordinators and affiliations are listed at the end of this paper. §Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA.

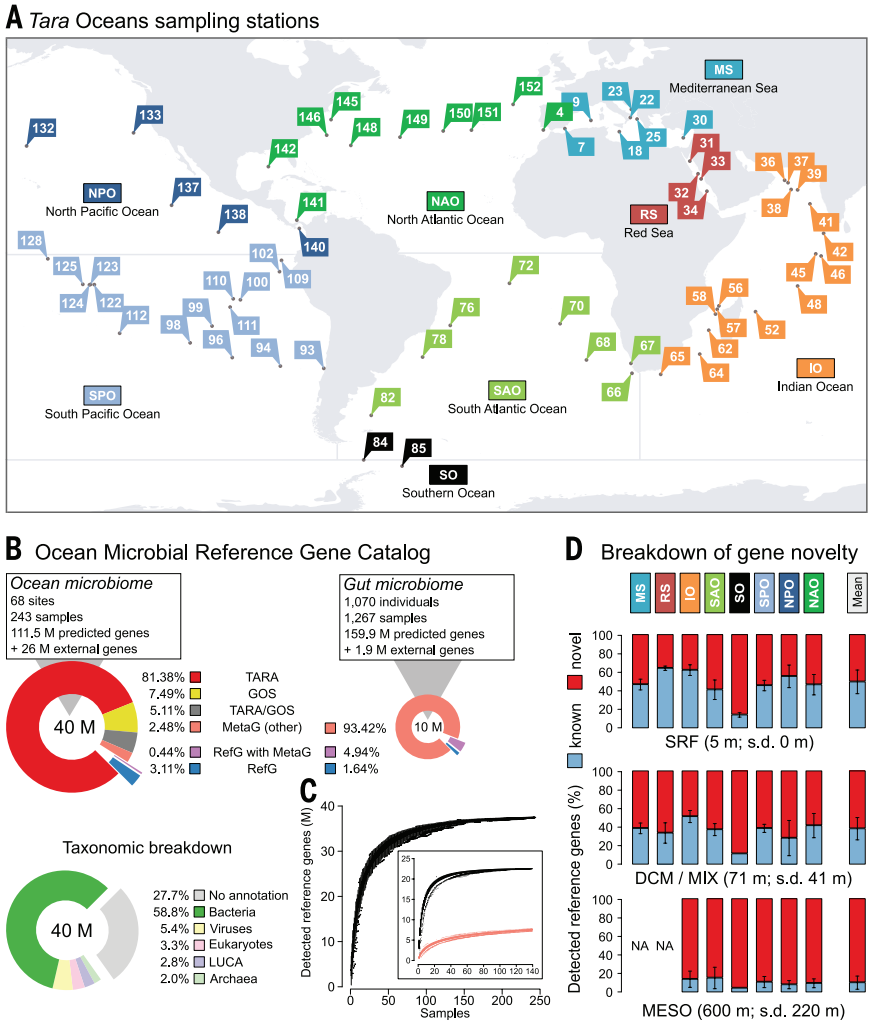


Fig. 1. *Tara* Oceans captures novel genetic diversity in the global ocean microbiome. (**A**) Geographic distribution of 68 (out of >200 in total) representative *Tara* Oceans sampling stations at which seawater samples and environmental data were collected from multiple depth layers. (**B**) Targeting viruses and microbial organisms up to 3 μm in size, deep Illumina shotgun sequencing of 243 samples, followed by metagenomic assembly and gene prediction, resulted in the identification of >111.5 M gene-coding sequences. The currently largest human gut microbial reference gene catalog (*16*) was built with similar amounts of data but from a substantially higher number of samples (n = 1,267). Genes identified in our study were clustered together with >26 M sequences from publicly available data [external genes; see (*14*)] to yield a set of >40 M reference genes (top left), which equals more than four times the number of genes in the human gut microbial reference gene catalog (top right). The combined clustering of genes identified in *Tara* Oceans samples with those obtained from public resources allowed us to annotate genes according to the composition of each cluster. For example, a gene was labeled as: "TARA/GOS" if its original cluster contained sequences from both *Tara* Oceans and GOS samples. More than 81% of the genes were found only in samples collected by *Tara* Oceans. A breakdown of taxonomic annotations (bottom left) shows that the reference gene catalog is mainly composed of bacterial genes (LUCA denotes genes that could not unambiguously be assigned to a domain of life). (**C**) Rarefaction curve of detected genes for 100-fold permuted sampling orders shows only a small increase in newly detected genes toward the end of sampling. The subplot compares sequencing depth-normalized rarefaction curves for 139 prokaryotic ocean samples (black) mapped to the prokaryotic subset of the OM-RGC (24.4 M genes) and the same number of random (100-fold permuted) human gut samples (pink) mapped to a human gut gene catalog (*16*). The lower asymptote for the human gut suggests that the ocean harbors a greater genetic diversity. (**D**) For the subset of 139 prokaryotic samples analyzed, the fraction of detected genes that had previously been available in public databases (blue) are compared to those that were newly identified in samples collected by *Tara* Oceans (red). The breakdown by ocean region and depths shows that the Southern Ocean and the mesopelagic zone had been vastly undersampled prior to *Tara* Oceans. NA, not available. Abbreviations: MS, Mediterranean Sea; RS, Red Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean; GOS, *Sorcerer II* Global Ocean Sampling expedition; MetaG, genes of metagenomic origin; RefG, genes from reference genome sequences; LUCA, last universal common ancestor; SRF, surface water layer; DCM, deep chlorophyll maximum layer; MIX, subsurface epipelagic mixed layer; MESO, mesopelagic zone.

0.45 to 0.8 μm): n = 59; prokaryote-enriched fractions (0.22 to 1.6 μm, 0.22 to 3 μm): n = 139] were paired-end shotgun Illumina sequenced to generate a total of more than 7.2 terabases (Tb), 29.6 ± 12.7 Gb per sample (*14*), enabling comparative analyses with the human gut microbiome for which metagenomic data of the same order of magnitude have been published {U.S. Human Microbiome Project, phase I—stool [1.5 Tb; (*15*)]} and the European Metagenomics of the Human Intestinal Tract project [3.8 Tb; (*16*, *17*)].

To generate a reference gene catalog [see also (*16*, *17*)], we first reconstructed the genomic content of *Tara* Oceans samples by metagenomic assembly and gene prediction (*18*) and combined these data with those from publicly available ocean metagenomes and reference genomes (*14*). Specifically, ~111.5 million (M) protein-coding nucleotide sequences were predicted and clustered at 95% nucleotide sequence identity with 24.4 M sequences from other ocean metagenomes (*14*) and 1.6 M sequences from ocean prokaryotic (n = 433) and viral (n = 114) reference genomes (*14*). This resulted in a global Ocean Microbial Reference Gene Catalog (OM-RGC), which comprises >40 M nonredundant representative genes from viruses, prokaryotes, and picoeukaryotes (Fig. 1B). Compared to a human gut microbial reference gene catalog (*16*), the OM-RGC comprises more than four times the number of genes, most of which (59%) appear prokaryotic (Fig. 1B). Almost 28% of the genes could not be taxonomically annotated. A large fraction is, however, likely of viral origin, because in size fractions targeting organisms smaller than 0.22 μm, 37% (SD = 9%) of the profiled sequence data mapped to nonannotated genes [see also (*19*)], whereas in prokaryote-enriched samples, this fraction decreased to 9% (SD = 2%). As expected, eukaryotic genes (3.3%) include those from protists (unicellular eukaryotes) but also from multicellular, larger organisms whose gametes or fragmented cells may have been sampled (*14*).

In total, 81.4% of the genes were exclusive to *Tara* Oceans samples, with only 5.11 and 0.44% overlapping with GOS sequences and reference genomes, respectively (Fig. 1B), which highlights the extent of the unexplored genomic potential in our oceans. Rarefaction analysis showed that the rate of new gene detection decreased to 0.01% by the end of sampling (Fig. 1C), suggesting that the abundant microbial sequence space appears well represented, at least for the targeted size ranges, sampling locations, and depths. Genes found in only one sample amounted to 3.6% of the OM-RGC, which may originate from localized specialists.

To complement the work of *Tara* Oceans Consortium partners who analyzed viral and protist-enriched size fractions (*19*, *20*) and integrated data across domains of life (*21*, *22*), we focused our analyses on 139 prokaryote-enriched samples, which included 63 surface water samples (5 m; SD = 0 m), 46 epipelagic subsurface water samples mostly from the DCM (71 m; SD = 41 m), and 30 mesopelagic samples (600 m; SD = 220 m). Using this set, we revealed that gene novelty generally

increased from surface to DCM waters and remained relatively stable across ocean regions, with overall about half of the genes being novel. As exceptions to this pattern, we find in Southern Ocean (SO) and mesopelagic samples about 80 and 90% of novelty, respectively. In addition to higher novelty in hitherto uncharted regions, these patterns likely reflect the detection of rare organisms by deep sequencing, although seasonal and locational differences of sampling in relatively well-studied regions may be additional contributing factors.

To put the degree of taxonomic novelty into context, we extracted a total of >14 M metagenomic 16*S* ribosomal RNA gene (16*S*) tags [16*S* $_{mi}$tags; (*12*)] and mapped these to operational taxonomic units (OTUs) based on clustering of reference 16*S* sequences (*23*) at 97% sequence identity. This cutoff has been commonly used to group taxa at the species level, although it may rather represent clades somewhere between species and genus level (*24*). The fraction of total 16*S* $_{mi}$tags not matching any reference OTUs also increased with depth but was on average only 5.5% (*14*). Thus, although the vast majority of prokaryotic clades detected in *Tara* Oceans metagenomes had already been captured by 16*S* sequencing, the OM-RGC now provides a link to their genomic content.

Diversity and depth stratification of the ocean microbiome

Given the global scale of *Tara* Oceans samples, we assessed patterns of diversity and stratifying factors of ocean microbial community composition. 16*S* $_{mi}$tags identified in our metagenomic data set mapped to a total of 35,650 OTUs (2937 OTUs; SD = 585 OTUs), and taxonomic and phylogenetic diversity were highly (R^2 = 0.96) correlated (*14*). The total richness estimate of 37,470 is comparable to the numbers from a previous study, which detected about 44,500 OTUs based on polymerase chain reaction (PCR)–amplified 16*S* rRNA tags from 356 globally distributed pelagic samples (*25*) that were collected in the context of the International Census of Marine Microbes (ICoMM) project (*26*). More than 93% of 16*S* $_{mi}$tags could be annotated at the phylum level. We found that typical members of Proteobacteria, including the ubiquitous clades SAR11 (Alphaproteobacteria) and SAR86 (Gammaproteobacteria), dominate the sampled areas of the ocean both in terms of relative abundance and taxonomic richness (*27*, *28*). Cyanobacteria, Deferribacteres, and Thaumarchaeota were also abundant, although the taxonomic richness within these phyla was smaller (Fig. 2). Photosynthetic cyanobacterial taxa such as *Prochlorococcus* and *Synechococcus* were detected in all mesopelagic samples and contributed about 1% of the abundance (Fig. 2), which is in line with previous reports suggesting a role for cyanobacteria in sinking particle flux (*29*).

To explore the overall variability in community composition, we performed a principal coordinate analysis (PCoA), which revealed that depth explained 73% of the variance (PC1 in Fig. 3A). This is consistent with a vertical stratification of microbial taxa and viruses according to changes in physicochemical parameters, such as light, temperature, and nutrients (*30*, *31*). Given this vertical stratification, we further characterized taxonomic and functional richness, between-sample dissimilarity (β-diversity), total cell abundance, and potential growth rates across three depth layers. Our results revealed an increase in both taxonomic and functional richness with depth, whereas cell abundance, as measured by flow cytometry, and potential maximum growth rates (*32*) decreased with depth (Fig. 3B).

Although increasing species richness from the surface to the mesopelagic has been reported locally, e.g., in the Mediterranean Sea (*33*), our findings emphasize the global relevance of this pattern. The observed increase in taxonomic and functional richness may reflect diversified species adapted to a wider range of niches, such as particle-associated microenvironments in the mesopelagic zone (*34*). In addition, slower growth, due

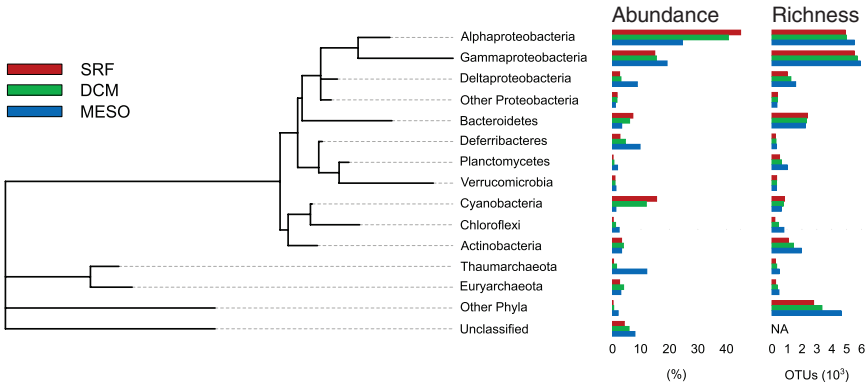


Fig. 2. Taxonomic breakdown of *Tara* Oceans samples. A phylum-level (class-level for Proteobacteria) breakdown of relative abundances is shown for all prokaryotic samples from three depth layers along with the number of detected taxa at the OTU level. SRF, surface water layer; DCM, deep chlorophyll maximum layer; MESO, mesopelagic zone.

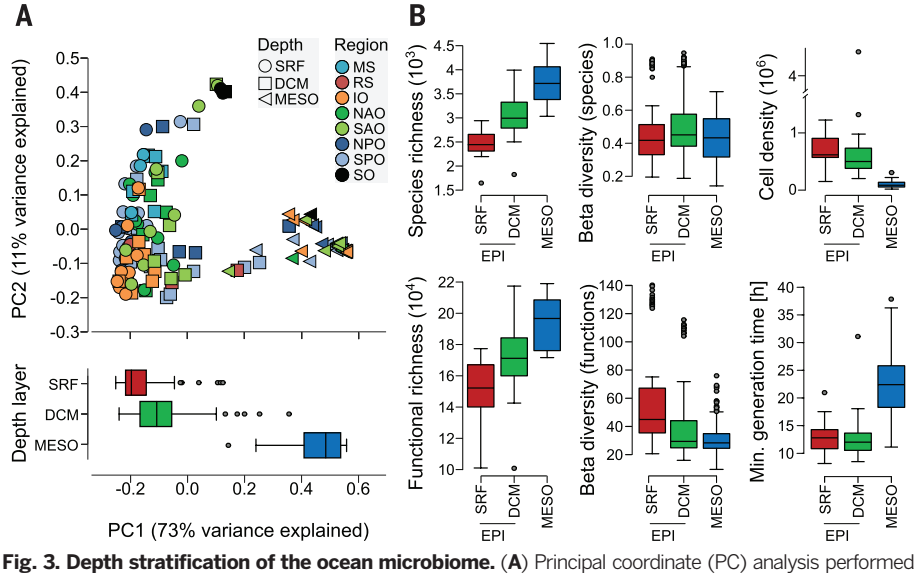


Fig. 3. Depth stratification of the ocean microbiome. (**A**) Principal coordinate (PC) analysis performed on community composition dissimilarities (Bray-Curtis) of 139 prokaryotic samples based on 16*S* $_{mi}$tag relative abundances shows that samples are significantly separated by their depth layer of origin, i.e., surface (SRF), deep chlorophyll maximum (DCM), or mesopelagic (MESO). Boxplots of the first PC illustrate differences between depth layers. Differences between samples from SRF and DCM were significant, but small compared to those with mesopelagic samples. Abbreviations for ocean regions are the same as in Fig. 1. (**B**) For a matched sample set from 20 stations where SRF, DCM, and MESO were sampled, calculations of within-sample species richness (top left) and between-sample diversities (top-center; Bray-Curtis) and cell densities per milliliter (top right) suggest an increase in species richness and a decrease in cell density with depth (pairwise Mann-Whitney U-test: $P < 0.001$), whereas no significant trend was found for between-sample dissimilarity. For gene functional groups (bottom left and center), richness increased with depth, whereas between-sample dissimilarity decreased. Minimum potential generation time of microbial communities (bottom right) is predicted to be higher in the mesopelagic compared to the epipelagic (EPI).

to more limited carbon sources in the mesopelagic zone, and higher motility have been suggested to reduce predation by flagellates and ciliates, as well as viral infection rates (*35*). Our metagenomic analysis now provides molecular support for these models by identifying a significant ($P < 0.001$) enrichment of chemotaxis and motility genes in the mesopelagic zone (see below).

Environmental drivers of community composition

A key question in ocean microbial ecology is the extent to which limited dispersal and historical contingency on the one hand, and global dispersion combined with selection by environmental factors on the other, are responsible for contemporary biogeographic patterns (*4*, *5*). The relationship between absolute latitude and biodiversity is an example of such a pattern, albeit one that is still controversial; while some authors found a negative correlation (*36*), others reported maxima in intermediate latitudinal ranges (*10*, *37*). The latter is supported by our findings (Fig. 4A), as an increase in richness with temperature was found from 4° to about 12°C, followed by a negative correlation for the remainder of the sampled temperature range (up to 30°C). This is also congruent with previous reports on oceanic groups of eukaryotes (*38*). A modeling study predicted season as a driver of biodiversity (*39*). For our data, however, the association of richness with temperature and latitude is robust to the confounding effect of seasonality (partial Mantel test, $P < 0.01$), although more data are needed for a rigorous statistical evaluation of such questions; for example, by periodically sampling the ocean across the globe on the same day (*40*). In addition to latitudinal biodiversity patterns, we found that taxonomic community dissimilarity increased up to about 5000 km within an ocean region (Fig. 4B). Together, our data support biogeographic patterns of microbial communities, in line with previous studies (*10*, *36*, *37*).

To further investigate the underlying mechanisms, we tested whether samples were more similar within than across ocean regions by focusing on surface samples only. If dispersal limitation rather than environmental selection dominated, we would expect a higher similarity within than across ocean regions. By contrast, if environmental selection explained biogeographic patterns, we would expect environmental factors to correlate with community similarity. Previous studies on selected ocean microbial taxa have shown a strong impact of light and temperature (*41*). For entire community assemblages, however, expectations are less clear. In a large-scale meta-analysis, salinity has been suggested as the major determinant across many (including ocean) ecosystems and to exceed the influence of temperature (*42*). In contrast, an analysis of functional trait composition in ocean environments suggested that temperature and light have stronger effects than nutrients or salinity (*10*, *43*).

A PCoA of taxonomic compositions of surface samples does not show a clear separation by regional origin, despite showing on average a higher similarity of communities within than across ocean regions (Fig. 5A). Instead, temperature was found to strongly correlate with PC1 ($R^2 = 0.76$). Thus, to verify the geographic independence of this pattern and to identify environmental drivers in our data set, we correlated distance-corrected dissimilarities of taxonomic and functional community composition with those of environmental factors (Fig. 5B). Overall, temperature and dissolved oxygen were the strongest correlates of both taxonomic and functional composition in the surface layer (Fig. 5B), while no significant correlation was found for salinity. Nutrients were only weakly correlated and, except for silicate, after the removal of a few extreme locations with very low temperatures, the correlations were not statistically significant.

Finally, we tackled the challenge of disentangling the high correlation between temperature and dissolved oxygen ($R^2 = 0.87$) in surface waters. To this end, we first used a machine learning–based approach (*44*) to independently model associations of each of these two factors with taxonomic and functional composition within surface samples (Fig. 6A). We then tested the strength of these associations in DCM layers, where the correlation between the two factors is much weaker ($R^2 = 0.16$), which allowed us to

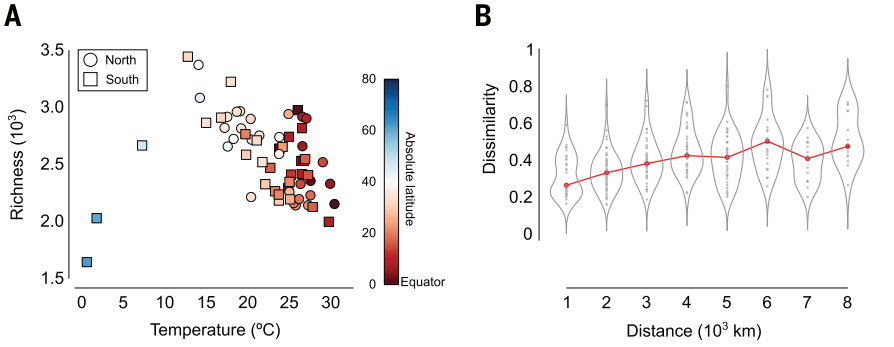


Fig. 4. Latitudinal diversity and distance decay of ocean microbial communities. (**A**) Plotting species richness against the temperature of sampling location shows an initial increase in richness up to about 15°C followed by a decrease toward warmer waters. Richness is highest in mid-latitudinal ranges rather than toward the equator. The color gradient denotes absolute latitudes (with increasing warmth of color from poles to equator). Shape of symbols denotes whether a sample originated from the Northern (circle) or Southern Hemisphere (square). (**B**) Pairwise microbial community dissimilarity (Bray-Curtis) based on relative $_{mi}$tag OTU abundances increases with distance between sampling stations up to about 5000 km. Pairwise distances were calculated only within ocean regions.

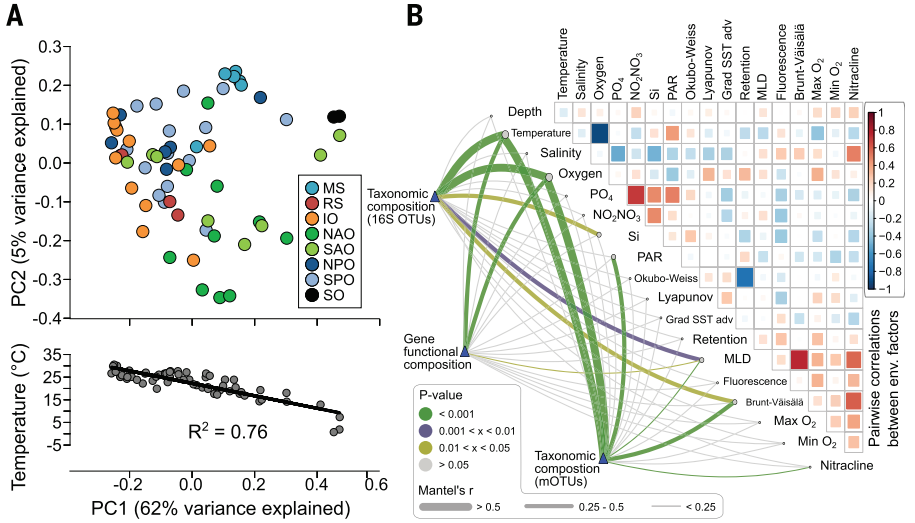


Fig. 5. Environmental drivers of surface microbial community composition. (**A**) Principal coordinate (PC) analysis of surface samples shows that samples are not clearly grouped by their regional origin (top), but rather separated by the local temperatures as shown by the strong correlation (R^2: 0.76) between the first PC and temperature (bottom). (**B**) Pairwise comparisons of environmental factors are shown, with a color gradient denoting Spearman's correlation coefficients. Taxonomic [based on two independent methods: $_{mi}$tags (*12*) and mOTUs (*13*)] and functional (based on biochemical KEGG modules) community composition was related to each environmental factor by partial (geographic distance–corrected) Mantel tests. Edge width corresponds to the Mantel's *r* statistic for the corresponding distance correlations, and edge color denotes the statistical significance based on 9,999 permutations.

effectively decouple dissolved oxygen from temperature. The surface-fitted model of temperature continued to achieve high prediction accuracy when applied at the DCM layers, whereas the oxygen model could not be generalized across depths. To illustrate the strength of these associations, we show that temperature could be predicted with an explained variance of 86%, using only species abundance as information (Fig. 6B). These results were validated with data from the GOS project (R^2 = 0.66) despite differences in sampling and sequencing procedures between the two studies (Fig. 6B).

Taken together, our data suggest that geographic distance plays a subordinate role and reveals temperature to be the major environmental factor shaping taxonomic and functional microbial community composition in the photic open ocean. Thus, a global dispersal potential for microorganisms (*45*) and subsequent environmental selection may, at least for some taxa, represent a mechanism for driving patterns of microbial biogeography. At the same time, localized adaptations by natural selection will lead to differences in spatially distant populations of phylogenetically similar organisms, so that characterizing these variations at strain-level resolution represents an important challenge for the future.

Core functional analysis between ecosystems

The generation of nonredundant gene abundance profiles from a large number (e.g., >100) of samples can be used to define a set of gene families, as a proxy for gene-encoded functions, which are ubiquitously found (core) in microbial communities. Such an analysis was performed for the human gut (*17*), which represents a fundamentally different microbial ecosystem (anoxic, host-associated, dominated by heterotrophs). However, owing to the lack of other large-scale, ecosystem-wide metagenomic data sets, it has been unknown how many of these core functions are shared with any other ecosystem. Thus, we first mapped the OM-RGC to known gene families, represented by clusters of orthologous groups [OGs, (*46*)] and selected prokaryotic genes to ensure comparability between the data sets. In total, we detected 39,246 OGs (19,524 OGs per sample; SD = 2682 OGs). Of those, the number of shared OGs rapidly decreased with sample size, reaching a minimum of 5755 ocean core OGs that were present in all (n = 139) prokaryote-enriched samples (Fig. 7A). Overall, we found that 40% of these ocean core OGs were of unknown function, compared to only 9% of the human gut core OGs (Fig. 7B).

We also sought to determine the overlap of core functions between the two ecosystems and to identify differentially abundant core functional categories (*47*), and contrast their relative importance in each of them (Fig. 7C). The ocean core contained almost twice as many OGs as the gut core, which may reflect the sampling of a greater number and higher complexity of niches in the ocean ecosystem than in the mostly anoxic, thermally stable human gut. However, despite large physicochemical differences between the two ecosystems, we found that most of the prokaryotic gene abundance (73% in the ocean; 63% in the gut) can be attributed to a shared functional core. Significant differential abundances between the two ecosystems were found across many functional categories. Most notably, those for defense mechanisms, signal transduction, and carbohydrate transport and metabolism were considerably more abundant in the gut, whereas those for transport mechanisms in general (coenzyme, lipid, nucleotide, amino acids, secondary metabolites) and energy production (including photosynthesis) were more abundant in the ocean (Fig. 7C).

Functional variability across ocean depths and regions

Functional redundancy across different taxa in microbial communities has been suggested to confer a buffering capacity for an ecosystem in scenarios of biodiversity loss (*48*). When contrasting taxonomic and functional variability in the ocean, we indeed found high taxonomic variability (even at phylum level) accompanied by relatively stable distributions of gene abundances summarized into functional categories (*47*) (Fig. 8A). This is also congruent with previous reports for the human gut, where gene abundances of metabolic pathways were found to be evenly distributed across samples, while taxonomic compositions varied markedly between subjects (*49*). Thus, despite the presumably greater environmental complexity in the ocean, the congruent functional redundancy observed in both ecosystems may be indicative of an ecosystem-independent property of microbial communities.

We next differentiated ocean core from noncore OGs, as the latter are more relevant for environment-specific adaptations. Within the ocean, 67% (SD = 5%) of the total gene abundance was attributed to ocean core OGs. After removing these and the 29% (SD = 5%) of gene abundance from genes that were not assigned to any OG, 4% (SD = 1%) remained as the noncore fraction. The abundance distribution among these noncore OGs, of which the largest fraction encode unknown functions, displayed a much greater variability across samples even when summarized into functional categories (Fig. 8A). Thus, in addition to the stable abundance distribution of core functional processes, as reported here and for human body habitats (*49*), functional variation similar in scale to that of the phylogenetic one can be detected when focusing on noncore, potentially adaptive gene families. As an example for such an environmental adaptation, we found an increase in lipid metabolism in oxygen minimum zones of the Eastern Pacific and Northern Indian Ocean (Fig. 8A).

Finally, to globally investigate the functional basis for the large community structural differences between the epipelagic layer and mesopelagic zone (Fig. 3A), we defined depth-specific core OGs using the approach introduced above. Unexpectedly, we found that the epipelagic core is almost completely contained in the mesopelagic core (Fig. 8B). When testing between-depth functional differences (Fig. 8B), we observed an enrichment of aerobic respiration genes in the ventilated mesopelagic zone, which is coherent with the finding that the mesopelagic zone is a key

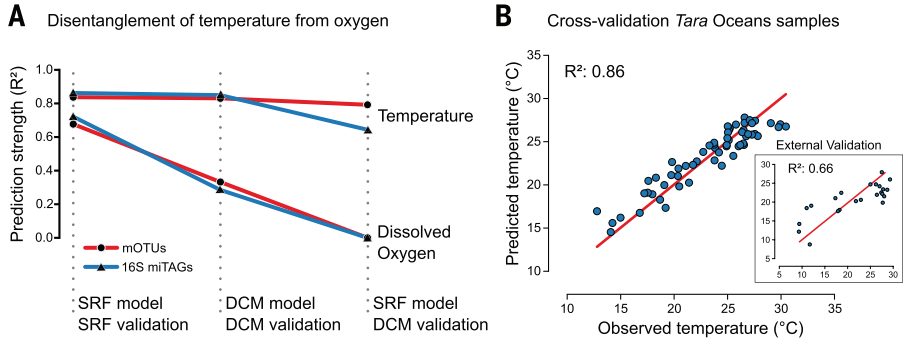


Fig. 6. Temperature as main environmental driver for microbial community composition in the epipelagic layer. (**A**) The strength of association between (meta)genomic and environmental data was tested by statistical models that were first generated with a subset of data for training and then validated on the remaining data. The prediction accuracy was used as a measure for the strength of association. Models that were trained on subsets of taxonomic data from surface water (SRF) samples could predict with high accuracy temperature and dissolved oxygen of samples used for validation (left). Models trained with subsets of taxonomic data from deep chlorophyll maximum (DCM) samples could predict temperature with high accuracy, but could predict dissolved oxygen with only moderate accuracy (middle). To demonstrate across-depth conservation of associations, we show that models trained on data from SRF samples could highly predict temperature, but failed to predict dissolved oxygen in DCM samples. (**B**) To illustrate prediction accuracy, and thus, strength of association between taxonomic composition (using 16S $_{mi}$tag abundances) and temperature, we show that in situ measured temperature could be predicted with 86% explained variance. The red diagonal shows the theoretical curve for perfect predictions. Sanger sequencing reads from the GOS project were used to calculate relative genus abundance tables. Using temperature prediction models trained at genus level using *Tara* Oceans data, we show (inset) that the results could be validated at relatively high accuracy given the large differences in sampling and sequencing methods between these two studies.

remineralization site of exported production (*50*). Flagellar assembly and chemotaxis were also enriched in mesopelagic samples, which is in contrast to previous findings (*51*) but congruent with the model that motility reduces grazing mortality in planktonic bacteria (*52*). In addition, these motility traits are potentially of great utility for bacteria in the dark ocean to colonize sinking particles or marine snow aggregates. Our taxonomic analysis (Fig. 2), combined with the detection of photosynthesis genes in the mesopelagic zone (Fig. 8B), indeed suggests microbial sedimentation from the epipelagic layer into the mesopelagic zone. Moving among aggregates to exploit nutrient patches and potentially new niches (*34*) may drive the diversification of mesopelagic zone–adapted microbial populations (*53*). In the future, matching *Tara* Oceans metatranscriptomic data should help in differentiating active from dead sinking biomass and give further insights into how microbial communities contribute to remineralization and carbon export into the ocean interior.

Conclusions

Tara Oceans has generated, in addition to global biodiversity resources for larger organismal size spectra (*20*), the OM-RGC, which makes ocean microbial genetic diversity accessible for various targeted analyses. Here we analyzed prokaryote-enriched size fractions, whereas related papers studied viral ecology (*19*), cross-kingdom species interactions (*21*), and planktonic community connectivity across an ocean circulation choke-point (*22*). Despite some limitations in the sampled organismal size range, oceanic depth layers, and temporal resolution, our approach generated an ecosystem-wide data set that will be useful for improving predictive models of the ocean. Finding that temperature drives microbial community variation and revealing the high functional redundancy in ocean microbial communities at global scale have wide-ranging implications for potential climate change–related effects. The *Tara* Oceans data set supports progress not only toward a holistic understanding of the ocean ecosystem but also of microbial communities in general, by facilitating comparative analyses between ecosystems.

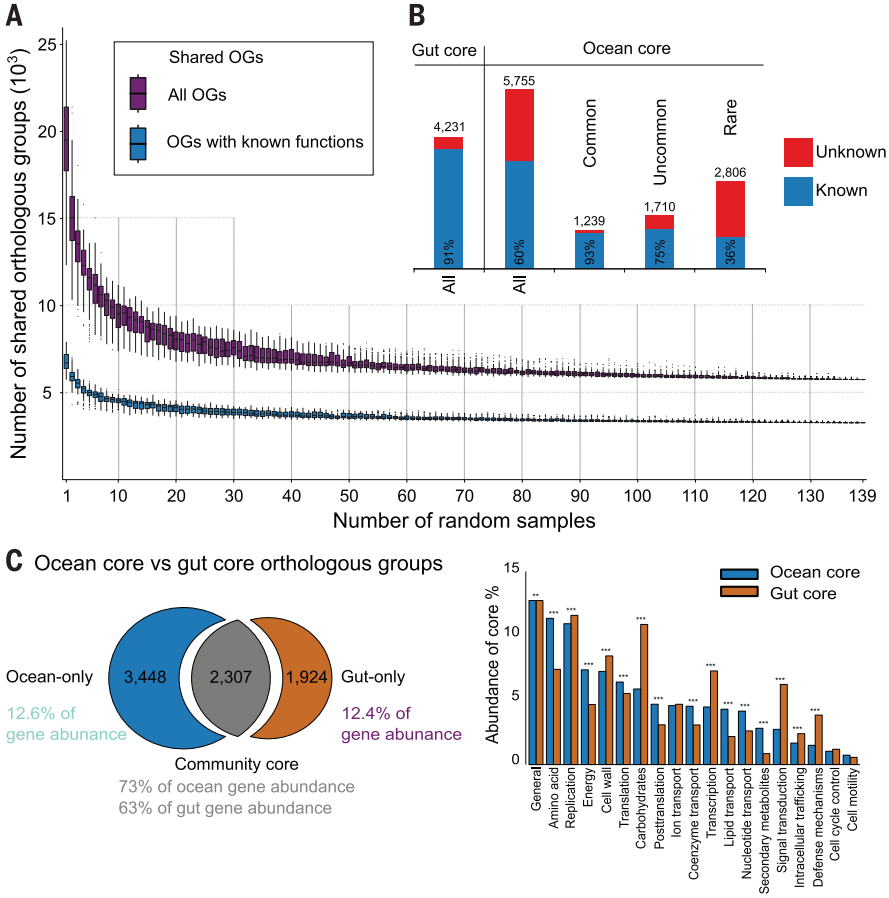


Fig. 7. Ocean versus human gut core orthologous groups. (**A**) The number of orthologous groups (OGs) that were shared among randomly selected sets of samples with sizes ranging from 1 to 139 was computed. With increasing sample size, the number of shared orthologous groups decreased first rapidly, then more gradually to a minimum of 5755 OGs at 139 samples, which was considered the set of ocean core OGs. Purple boxplots show the data for all OGs; blue boxplots show the data for OGs of known function. (**B**) Comparative statistics between ocean and human gut core OGs, showing that for a large fraction of ocean core OGs (40%), the functionality is unknown, which is in stark contrast to the human gut ecosystem (9%). Ocean core OGs are further subdivided into groups of OGs that are commonly (>50%), uncommonly (10% to 50%), or rarely (<10%) found in marine reference genomes. (**C**) A comparison of ocean and human gut core OGs (left) shows a large overlap of functions between these two fundamentally different ecosystems both qualitatively and quantitatively. The bar chart (right) displays a comparison of gene abundance summarized into OG functional categories to illustrate functional enrichments. Asterisks denote Mann-Whitney U-test results (**P < 0.01, ***P < 0.001).

Materials and methods

Sample and environmental data collection

From 2009 to 2013, morphological, genetic, and environmental data were collected at >200 sampling stations across all major oceanic provinces during the *Tara* Oceans expedition. The sampling strategy and methodology are described in (*54*–*57*). Sampling and enumeration of heterotrophic prokaryotes, phototrophic picoplankton, and small eukaryotes by flow cytometry followed previously described procedures, which are summarized in (*58*). Sample provenance is described in table S1 and in (*55*). Sample-associated environmental data and sample-associated biodiversity indexes were inferred at the depth of sampling (*56*, *57*), and additional information is available at (*14*).

Extraction and sequencing of metagenomic DNA

Metagenomic DNA from prokaryote and girus-enriched size fraction filters, and from precipitated viruses, was extracted as described in (*12*), (*59*), and (*19*), respectively. DNA (30 to 50 ng) was sonicated to a 100– to 800–base pair (bp) size range. DNA fragments were subsequently end repaired and 3′-adenylated before Illumina adapters were added by using the NEBNext Sample Reagent Set (New England Biolabs). Ligation products were purified by Ampure XP (Beckmann Coulter), and DNA fragments (>200 bp) were PCR-amplified with Illumina adapter-specific primers and Platinum Pfx DNA polymerase (Invitrogen). Amplified library fragments were size selected (~300 bp) on a 3% agarose gel. After library profile analysis using an Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and quantitative PCR (MxPro, Agilent Technologies, USA), each library was sequenced with 101 base-length read chemistry in a paired-end flow cell on Illumina sequencing machines (Illumina, USA).

Metagenomic sequence assembly and gene predictions

Using MOCAT (version 1.2) (*18*), high-quality (HQ) reads were generated (option *read_trim_filter*; solexaqa with length cut-off 45 and quality cut-off 20) and reads matching Illumina sequencing adapters were removed (option *screen_fastafile* with e-value 0.00001). Screened HQ reads were assembled (option *assembly*; minimum length 500 bp), and gene-coding sequences [minimum length 100 nucleotides (nt)] were predicted on the assembled scaftigs [option *gene_prediction*;

MetaGeneMark (version 2.8) (*60*)], generating a total of 111.5 M gene-coding sequences (*14*). Assembly errors were estimated by testing for colinearity between assembled contigs and genes and unassembled 454 sequencing reads by using a subset of 11 overlapping samples (*58*). From this analysis, we estimate that 1.5% of contigs had breakpoints and thus may suffer from errors (*14*). This error rate is more than a factor of 6.5 less than previous estimates of contig chimericity in simulated metagenomic assemblies (9.8%) with similar N_{50} values (*61*).

Generation of the ocean microbial reference gene catalog

Predicted gene-coding sequences were combined with those identified in publicly available ocean metagenomic data and reference genomes: 22.6 M predicted genes from the GOS expedition (*6*, *7*), 1.78 M from Pacific Ocean Virome study (POV) (*62*), 14.8 thousand from viral genomes from the Marine Microbiology Initiative (MMI) at the Gordon & Betty Moore Foundation (*14*), and 1.59 M from 433 ocean microbial reference genomes (*14*). The reference genomes were selected by the following procedure: An initial set of 3496 reference genomes (all high-quality genomes available as of 23 February 2012) was clustered into 1753 species (*24*), from each of which we selected one representative genome. After mapping all HQ reads against these genomes, a genome was selected if the base coverage was >1× or if the fraction of genome coverage was >40% in at least one sample. In addition, we included prokaryotic genomes for which habitat entries matched the terms "Marine" or "Sea Water" in the Integrated Microbial Genomes database (*63*) or if a genome was listed under the Moore Marine Microbial Sequencing project (*64*) as of 29 July 2013. Finally, we applied previously established quality criteria (*24*), resulting in a final set of 433 ocean microbial reference genomes (*14*). For data from GOS, POV, and MMI, assemblies were downloaded from the CAMERA portal (*64*). A total of 137.5 M gene-coding nucleotide sequences were clustered by using the same criteria as in (*16*); i.e., 95% sequence identity and 90% alignment coverage of the shorter sequence. The longest sequence of each cluster was selected, and after removing sequences <100 nt, we obtained a set of 40,154,822 genes [i.e., nonredundant contiguous gene-coding nucleotide sequences operationally defined as "genes"; see also (*16*, *17*)] that we refer to as the Ocean Microbial Reference Gene Catalog (OM-RGC).

Taxonomic and functional annotation of the OM-RGC

We taxonomically annotated the OM-RGC using a modified dual BLAST-based last common ancestor (2bLCA) approach as described in (*58*). For modifications, we used RAPsearch2 (*65*) rather than BLAST to efficiently process the large data volume and a database of nonredundant protein sequences from UniProt (version: UniRef_2013_07) and eukaryotic transcriptome data not represented in UniRef. The OM-RGC was functionally annotated to orthologous groups in the eggNOG (version 3) and KEGG databases (version 62) with SmashCommunity (version 1.6) (*46*, *66*, *67*). In total, 38% and 57% of the genes could be annotated by homology to a KEGG ortholog group (KO) or an OG, respectively. Functional modules were defined by selecting previously described key marker genes for 15 selected ocean-related processes, such as photosynthesis, aerobic respiration, nitrogen metabolism, and methanogenesis (*14*).

Taxonomic profiling using 16*S* tags and metagenomic operational taxonomic units 16*S* fragments directly identified in Illumina-sequenced metagenomes ($_{mi}$tags) were identified as described in (*12*). 16*S* $_{mi}$tags were mapped to cluster centroids of taxonomically annotated 16*S* reference sequences from the SILVA database (*23*) (release 115: SSU Ref NR 99) that had been clustered at 97% sequence identity with USEARCH v6.0.307 (*68*). 16*S* $_{mi}$tag counts were normalized by the total sum for each sample. In addition, we identified protein-coding marker genes suitable for metagenomic species profiling using fetchMG (*13*) in all 137.5 M gene-coding sequences and clustered them into metagenomic operational taxonomic units (mOTUs) that group organisms into species-level clusters at higher accuracy than 16*S* OTUs as described in (*13*, *24*). Relative abundances of mOTU linkage groups were quantified with MOCAT (version 1.3) (*18*).

Functional profiling using the OM-RGC

Gene abundance profiles were generated by mapping HQ reads from each sample to the OM-RGC (MOCAT options *screen* and *filter* with length and identity cutoffs of 45 and 95%, respectively, and paired-end filtering set to *yes*). The abundance of each reference gene in each sample was calculated as gene length–normalized base and insert counts (MOCAT option *profile*). Functional abundances were calculated as the sum of the relative abundances of reference genes, or key

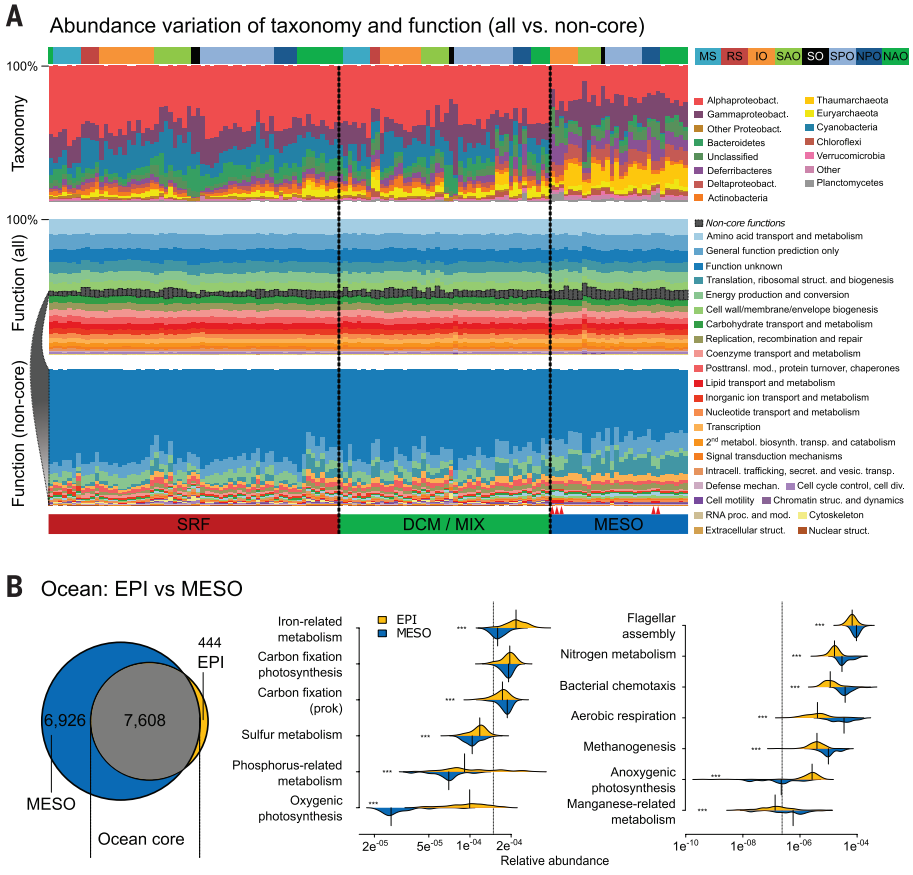


Fig. 8. Functional structuring of the ocean microbiome. (**A**) Phylum-level (class-level for Proteobacteria) taxonomic variability is higher (top, median relative SD = 65%) relative to the functional composition (OG functional categories) of ocean microbial samples (center, median relative SD = 7%). Removal of functions that are ubiquitous in the ocean environment reveals the variable, noncore fraction (bottom, median relative SD = 47%), which amounts on average to 4% of the total gene abundance. Red triangles on *x* axis highlight mesopelagic samples collected in oxygen minimum zones of the Indian Ocean and Eastern Pacific, which show increased levels of lipid metabolism in noncore functions. (**B**) Venn diagram (left) showing that core OGs in the epipelagic layer of the ocean are almost completely contained in mesopelagic core OGs (left). The bean charts (right) display differential abundances of marker genes (based on KO annotations) for selected functional processes in the ocean. Asterisks denote Mann-Whitney U test results (***P < 0.001).

marker genes (*14*), annotated to different functional groups (OGs, KOs, and KEGG modules). For each functional module, the abundance was calculated as the sum of relative abundances of marker KOs normalized by the number of KOs. For comparative analyses with the human gut ecosystem, we used the subset of the OM-RGC that was annotated to Bacteria or Archaea (24.4 M genes). Using a rarefied (to 33 M inserts) gene count table, an OG was considered to be part of the ocean microbial core if at least one insert from each sample was mapped to a gene annotated to that OG. Samples from the human gut ecosystem were processed similarly, and a list of all OGs that were defined in either the ocean or the gut as core is provided in (*14*).

Microbial community structural analyses and prediction of minimum generation times

16*S* $_{mi}$tag counts were rarefied 100 times to the minimum number of total 16*S* $_{mi}$tags per sample (39,410), and OTU richness and Chao1 richness estimators were calculated as the mean of all rarefactions (*14*). A phylogenetic tree of 16*S* $_{mi}$tags was calculated from full-length 16*S* sequences, by using parts of the LotuS 16*S* pipeline (*69*). This phylogenetic tree was midpoint rooted in R and used with the $_{mi}$tag abundance matrix rarefied to 39,000 reads per sample to calculate Faith's phylogenetic diversity (*70*) as the mean value of five repetitions (*14*). Similarly, OG richness was computed as the average of 10 rarefactions (*14*). Community growth potential from genomic traits was estimated as the average minimum generation time of the organisms present in the sample, weighted by their abundance, as previously described (*32*).

Distance correlations between genomic and environmental data

We computed pairwise distances between samples on the basis of (i) relative abundances of taxonomic (16*S* $_{mi}$tags and mOTUs) and gene functional compositions (at KEGG module level)—the compositional data; (ii) in situ measurements of physicochemical data—the environmental data; and (iii) geographic location of sampling stations—the geographic data. Data from the three southernmost stations were removed from the analysis, as these stations are outside the range of the rest of the data in parameters such as temperature, oxygen, and nutrients. For compositional data, we applied a logarithmic transformation to relative abundances using the function $\log_{10}(x + x_0)$, where x is the original relative abundance and x_0 is a small constant, and $x_0 < \min(x)$.

We applied an additional low-abundance filter, which removed features whose relative abundance did not exceed 0.0001 in any sample. Environmental data were transformed to z-scores before calculating distances. We used Euclidean distances for compositional and environmental data and Haversine distances for geographic data. Given these distance matrices, we computed partial Mantel correlations between compositional and environmental data given geographic distance (9,999 permutations) using the *vegan* R software package. Partial Mantel tests were also performed between species richness and both temperature and latitude, while controlling for season.

Statistical modeling and correlation analysis

Compositional data (see above) were normalized to ranks across samples and then used to learn a regression model to predict environmental measures. In particular, we fitted an elastic net model (*44*) using inner cross-validation to set the hyperparameters as implemented by the *scikit-learn* Python package (*71*). For spatial autocorrelation-corrected cross-validation, samples from each ocean basin were iteratively held out for testing on a model learned from the rest of the samples.

As a measure of association between the environmental parameter and the compositional data, we computed the cross-validated R^2 (also known as Q^2) (*72*), defined as $1 - \sum \frac{(y_i - \hat{y}_i)^2}{(y_i - \overline{y}_i)^2}$, where y_i is the value of the parameter for sample i, $\hat{y}_i$ is the prediction for that same sample (obtained by held-out cross-validation), and $\overline{y}$ is the overall mean (the summation runs over all the samples). To disentangle effects of temperature and oxygen, we trained models on surface samples, which were then evaluated in DCM samples. Again, to avoid spatial autocorrelation, cross-validation by ocean basin was used. An external cross-validation was performed by classifying GOS reads using the RDP database (*73*). Only genera detected in both studies were considered. Because of the lower and varying sequencing depth of the GOS data, for each GOS sample, we downsampled *Tara* Oceans data to match the corresponding sequencing depth and learned a model based on this downsampled data set. This model was based on the presence or absence of the taxa (which was modeled by passing a binary input matrix to the elastic net fitting routines).

REFERENCES AND NOTES

1. P. G. Falkowski, R. T. Barber, V. Smetacek, Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* **281**, 200–206 (1998). doi: 10.1126/science.281.5374.200; pmid: 9660741
2. W. B. Whitman, D. C. Coleman, W. J. Wiebe, Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6578–6583 (1998). doi: 10.1073/pnas.95.12.6578; pmid: 9618454
3. S. C. Doney *et al*., Climate Change Impacts on Marine Ecosystems. *Annu. Rev. Mar. Sci.* **4**, 11–37 (2012). doi: 10.1146/annurev-marine-041911-111611
4. J. A. Fuhrman, Microbial community structure and its functional implications. *Nature* **459**, 193–199 (2009). doi: 10.1038/nature08058; pmid: 19444205
5. J. B. H. Martiny *et al*., Microbial biogeography: Putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006). doi: 10.1038/nrmicro1341; pmid: 16415926
6. D. B. Rusch *et al*., The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLOS Biol.* **5**, e77 (2007). doi: 10.1371/journal.pbio.0050077; pmid: 17355176
7. S. Yooseph *et al*., The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PLOS Biol.* **5**, e16 (2007). doi: 10.1371/journal.pbio.0050016; pmid: 17355171
8. A. Barberán, A. Fernández-Guerra, B. J. M. Bohannan, E. O. Casamayor, Exploration of community traits as ecological markers in microbial metagenomes. *Mol. Ecol.* **21**, 1909–1917 (2012). doi: 10.1111/j.1365-294X.2011.05383.x; pmid: 22121910
9. T. A. Gianoulis *et al*., Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1374–1379 (2009). doi: 10.1073/pnas.0808022106; pmid: 19164758
10. J. Raes, I. Letunic, T. Yamada, L. J. Jensen, P. Bork, Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Syst. Biol.* **7**, 473 (2011). doi: 10.1038/msb.2011.6; pmid: 21407210
11. E. Karsenti *et al*., A holistic approach to marine eco-systems biology. *PLOS Biol.* **9**, e1001177 (2011). doi: 10.1371/journal.pbio.1001177; pmid: 22028628
12. R. Logares *et al*., Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* **16**, 2659–2671 (2014). doi: 10.1111/1462-2920.12250; pmid: 24102695
13. S. Sunagawa *et al*., Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013). doi: 10.1038/nmeth.2693; pmid: 24141494
14. Companion website tables W1 to W8, data, and information are available at http://ocean-microbiome.embl.de/companion.html
15. Human Microbiome Project Consortium, *Nature* **486**, 215–221 (2012). doi: 10.1038/nature11209; pmid: 22699610
16. J. Li *et al*., An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014). doi: 10.1038/nbt.2942; pmid: 24997786
17. J. Qin *et al*., A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010). doi: 10.1038/nature08821; pmid: 20203603
18. J. R. Kultima *et al*., MOCAT: A metagenomics assembly and gene prediction toolkit. *PLOS ONE* **7**, e47656 (2012). doi: 10.1371/journal.pone.0047656; pmid: 23082188
19. J. R. Brum *et al*., Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
20. C. de Vargas *et al*., Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
21. G. Lima-Mendez *et al*., Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).
22. E. Villar *et al*., Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science* **348**, 1261447 (2015).
23. C. Quast *et al*., The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013). doi: 10.1093/nar/gks1219; pmid: 23193283
24. D. R. Mende, S. Sunagawa, G. Zeller, P. Bork, Accurate and universal delineation of prokaryotic species. *Nat. Methods* **10**, 881–884 (2013). doi: 10.1038/nmeth.2575; pmid: 23892899
25. L. Zinger *et al*., Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLOS ONE* **6**, e24570 (2011). doi: 10.1371/journal.pone.0024570; pmid: 21931760
26. L. Amaral-Zettler *et al*., in *Life in the World's Oceans*, A. D. McIntyre, Ed. (Wiley-Blackwell, Oxford, 2010), pp. 221–245.
27. C. L. Dupont *et al*., Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* **6**, 1186–1199 (2012). doi: 10.1038/ismej.2011.189; pmid: 22170421
28. R. M. Morris *et al*., SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810 (2002). doi: 10.1038/nature01240; pmid: 12490947
29. K. Lochte, C. M. Turley, Bacteria and cyanobacteria associated with phytodetritus in the deep sea. *Nature* **333**, 67–69 (1988). doi: 10.1038/333067a0
30. S. J. Giovannoni, U. Stingl, Molecular diversity and ecology of microbial plankton. *Nature* **437**, 343–348 (2005). doi: 10.1038/nature04158; pmid: 16163344
31. B. L. Hurwitz, J. R. Brum, M. B. Sullivan, Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome. *ISME J.* **9**, 472–484 (2015). doi: 10.1038/ismej.2014.143; pmid: 25093636
32. S. Vieira-Silva, E. P. C. Rocha, The systemic imprint of growth and its uses in ecological (meta)genomics. *PLOS Genet.* **6**, e1000808 (2010). doi: 10.1371/journal.pgen.1000808; pmid: 20090831
33. T. Pommier *et al*., Spatial patterns of bacterial richness and evenness in the NW Mediterranean Sea explored by pyrosequencing of the 16S rRNA. *Aquat. Microb. Ecol.* **61**, 221–233 (2010). doi: 10.3354/ame01484
34. R. Stocker, Marine microbes see a sea of gradients. *Science* **338**, 628–633 (2012). doi: 10.1126/science.1208929; pmid: 23118182

35. J. Pernthaler, Predation on prokaryotes in the water column and its ecological implications. *Nat. Rev. Microbiol.* **3**, 537–546 (2005). doi: 10.1038/nrmicro1180; pmid: 15953930
36. W. J. Sul, T. A. Oliver, H. W. Ducklow, L. A. Amaral-Zettler, M. L. Sogin, Marine bacteria exhibit a bipolar distribution. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2342–2347 (2013). doi: 10.1073/pnas.1212424110; pmid: 23324742
37. J. A. Fuhrman *et al.*, A latitudinal diversity gradient in planktonic marine bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 7774–7778 (2008). doi: 10.1073/pnas.0803070105; pmid: 18509059
38. D. P. Tittensor *et al.*, Global patterns and predictors of marine biodiversity across taxa. *Nature* **466**, 1098–1101 (2010). doi: 10.1038/nature09329; pmid: 20668450
39. J. Ladau *et al.*, Global marine bacterial diversity peaks at high latitudes in winter. *ISME J.* **7**, 1669–1677 (2013). doi: 10.1038/ismej.2013.37; pmid: 23514781
40. Ocean Sampling Day, www.oceansamplingday.org.
41. Z. I. Johnson *et al.*, Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science* **311**, 1737–1740 (2006). doi: 10.1126/science.1118052; pmid: 16556835
42. C. A. Lozupone, R. Knight, Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 11436–11440 (2007). doi: 10.1073/pnas.0611525104; pmid: 17592124
43. D. P. Herlemann *et al.*, Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J.* **5**, 1571–1579 (2011). doi: 10.1038/ismej.2011.41; pmid: 21472016
44. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005). doi: 10.1111/j.1467-9868.2005.00503.x
45. B. J. Finlay, Global dispersal of free-living microbial eukaryote species. *Science* **296**, 1061–1063 (2002). doi: 10.1126/science.1070710; pmid: 12004115
46. S. Powell *et al.*, eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40**, D284–D289 (2012). doi: 10.1093/nar/gkr1060; pmid: 22096231
47. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000). doi: 10.1093/nar/28.1.33; pmid: 10592175
48. T. Bell, J. A. Newman, B. W. Silverman, S. L. Turner, A. K. Lilley, The contribution of species richness and composition to bacterial services. *Nature* **436**, 1157–1160 (2005). doi: 10.1038/nature03891; pmid: 16121181
49. Human Microbiome Project Consortium, *Nature* **486**, 207–214 (2012). doi: 10.1038/nature11234; pmid: 22699609
50. J. Arístegui, J. M. Gasol, C. M. Duarte, G. J. Herndl, Microbial oceanography of the dark ocean's pelagic realm. *Limnol. Oceanogr.* **54**, 1501–1529 (2009). doi: 10.4319/lo.2009.54.5.1501
51. E. F. DeLong *et al.*, Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496–503 (2006). doi: 10.1126/science.1120250; pmid: 16439655
52. C. Matz, K. Jürgens, High motility reduces grazing mortality of planktonic bacteria. *Appl. Environ. Microbiol.* **71**, 921–929 (2005). doi: 10.1128/AEM.71.2.921-929.2005; pmid: 15691949
53. Y. Yawata *et al.*, Competition-dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5622–5627 (2014). doi: 10.1073/pnas.1318943111; pmid: 24706766
54. S. Pesant *et al.*, Open science resources for the discovery and analysis of *Tara* Oceans Data. http://biorxiv.org/content/early/2015/05/08/019117 (2015).
55. *Tara* Oceans Consortium, Coordinators; *Tara* Oceans Expedition, Participants (2014): Registry of selected samples from the *Tara* Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840721
56. S. Chaffron *et al.*, (2014): Contextual environmental data of selected samples from the *Tara* Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840718
57. S. Chaffron *et al.*, (2014): Contextual biodiversity data of selected samples from the *Tara* Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840698
58. P. Hingamp *et al.*, Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013). doi: 10.1038/ismej.2013.59; pmid: 23575371
59. C. Clerissi *et al.*, Unveiling of the diversity of Prasinoviruses (*Phycodnaviridae*) in marine samples by using high-throughput sequencing analyses of PCR-amplified DNA polymerase and major capsid protein genes. *Appl. Environ. Microbiol.* **80**, 3150–3160 (2014). doi: 10.1128/AEM.00123-14; pmid: 24632251
60. W. Zhu, A. Lomsadze, M. Borodovsky, Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38**, e132–e132 (2010). doi: 10.1093/nar/gkq275; pmid: 20403810
61. D. R. Mende *et al.*, Assessment of metagenomic assembly using simulated next generation sequencing data. *PLOS ONE* **7**, e31386 (2012). doi: 10.1371/journal.pone.0031386; pmid: 22384016
62. B. L. Hurwitz, L. Deng, B. T. Poulos, M. B. Sullivan, Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.* **15**, 1428–1440 (2013). doi: 10.1111/j.1462-2920.2012.02836.x; pmid: 22845467
63. Integrated Microbial Genomes database: https://img.jgi.doe.gov/cgi-bin/w/main.cgi.
64. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis, http://camera.calit2.net/microgenome.
65. Y. Zhao, H. Tang, Y. Ye, RAPSearch2: A fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28**, 125–126 (2012). doi: 10.1093/bioinformatics/btr595; pmid: 22039206
66. M. Arumugam, E. D. Harrington, K. U. Foerstner, J. Raes, P. Bork, SmashCommunity: A metagenomic annotation and analysis tool. *Bioinformatics* **26**, 2977–2978 (2010). doi: 10.1093/bioinformatics/btq536; pmid: 20959381
67. M. Kanehisa *et al.*, KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36** (Database), D480–D484 (2008). doi: 10.1093/nar/gkm882; pmid: 18077471
68. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010). doi: 10.1093/bioinformatics/btq461; pmid: 20709691
69. F. Hildebrand, R. Tadeo, A. Y. Voigt, P. Bork, J. Raes, LotuS: An efficient and user-friendly OTU processing pipeline. *Microbiome* **2**, 30 (2014). doi: 10.1186/2049-2618-2-30
70. D. P. Faith, Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **61**, 1–10 (1992). doi: 10.1016/0006-3207(92)91201-3
71. F. Pedregosa *et al.*, *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
72. H. Wold, in *Systems Under Indirect Observation: Causality, Structure, Prediction,* K. G. Jöreskog, Ed. (North-Holland, 1982), vol. 2, pp. 1-54.
73. Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007). doi: 10.1128/AEM.00062-07; pmid: 17586664

ACKNOWLEDGMENTS

We thank the following individuals and sponsors for their support: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB, Stazione Zoologica Anton Dohrn, Università degli Studi di Milano-Bicocca, Fund for Scientific Research–Flanders, Rega Institute, KU Leuven, The French Ministry of Research, the French Government "Investissements d'Avenir" programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), PSL Research University (ANR-11-IDEX-0001-02), Agence Nationale de la Recherche (projects POSEIDON/ANR-09-BLAN-0348, PHYTBACK/ANR-2010-1709-01, PROMETHEUS/ANR-09-PCS-GENM-217, TARA GIRUS/ANR-09-PCS-GENM-218), European Union FP7 (MicroB3/no.287589, IHMS/HEALTH-F4-2010-261376), European Research Council Advanced Grant Award to C.B. (Diatomite: 294823), Gordon and Betty Moore Foundation grant (no. 3790) to M.B.S., Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to S.G.A., TANIT (CONES 2010-0036) from the Agència de Gestió d´Ajusts Universitaris i Reserca to SGA, Japan Society for the Promotion of Science KAKENHI grant no. 26430184 to H.O., and FWO, BIO5, Biosphere 2 to M.B.S. We also thank the following for their support: Agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, and the *Tara* schooner and its captain and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries that graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge excellent assistance from the European Bioinformatics Institute (EBI), in particular G. Cochrane and P. ten Hoopen, as well as the EMBL Advanced Light Microscopy Facility (ALMF), in particular R. Pepperkok. The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled. Data described herein are available at http://ocean-microbiome.embl.de/companion.html, at the EBI under the project identifiers PRJEB402 and PRJEB7988, and at PANGAEA (*55*–*57*). The data release policy regarding future public release of *Tara* Oceans data is described in (*54*). All authors approved the final manuscript. This article is contribution number 22 of *Tara* Oceans. Additional data are in the supplementary materials.

TARA OCEANS COORDINATORS

Silvia G. Acinas,[1] Peer Bork,[2] Emmanuel Boss,[3] Chris Bowler,[4] Colomban de Vargas,[5,6] Michael Follows,[7] Gabriel Gorsky,[8,9] Nigel Grimsley,[10,11] Pascal Hingamp,[12] Daniele Iudicone,[13] Olivier Jaillon,[14,15,16] Stefanie Kandels-Lewis,[2,17] Lee Karp-Boss,[3] Eric Karsenti,[4,17] Uros Krzic,[18] Fabrice Not,[5,6] Hiroyuki Ogata,[19] Stephane Pesant,[20,21] Jeroen Raes,[22,23,24] Emmanuel G. Reynaud,[25] Christian Sardet,[26,27] Mike Sieracki,[28] Sabrina Speich,[29,30] Lars Stemmann,[8] Matthew B. Sullivan,[31] Shinichi Sunagawa,[2] Didier Velayoudon,[32] Jean Weissenbach,[14,15,16] Patrick Wincker[14,15,16]

[1]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)–CSIC, Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [2]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [3]School of Marine Sciences, University of Maine, Orono, ME, USA. [4]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, F-75005 Paris, France. [5]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [6]Sorbonne Universités, UPMC Université Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [9]Sorbonne Universités, UPMC Université Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [10]CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [11]Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [12]Aix Marseille Université CNRS IGS UMR 7256, 13288 Marseille, France. [13]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [14]CEA–Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [15]CNRS, UMR 8030, CP5706, Evry, France. [16]Université d'Evry, UMR 8030, CP5706, Evry, France. [17]Directors' Research, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [18]Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [19]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan. [20]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [21]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [22]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [23]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [24]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [25]Earth Institute, University College Dublin, Dublin, Ireland. [26]CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [27]Sorbonne Universités, UPMC Université Paris 06, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [28]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA. [29]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France. [30]Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France. [31]Department of Ecology and Evolutionary Biology, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA. [32]DVIP Consulting, Sèvres, France.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/1261359/suppl/DC1
Table S1

16 September 2014; accepted 24 February 2015
10.1126/science.1261359

OCEAN PLANKTON

Environmental characteristics of Agulhas rings affect interocean plankton transport

Emilie Villar,[1]* Gregory K. Farrant,[2,3]† Michael Follows,[11]† Laurence Garczarek,[2,3]† Sabrina Speich,[5,23]¶ Stéphane Audic,[2,3] Lucie Bittner,[2,3,4]‡ Bruno Blanke,[5] Jennifer R. Brum,[6]** Christophe Brunet,[7] Raffaella Casotti,[7] Alison Chase,[8] John R. Dolan,[9,10] Fabrizio d'Ortenzio,[9,10] Jean-Pierre Gattuso,[9,10] Nicolas Grima,[5] Lionel Guidi,[9,10] Christopher N. Hill,[11] Oliver Jahn,[11] Jean-Louis Jamet,[12] Hervé Le Goff,[13] Cyrille Lepoivre,[1] Shruti Malviya,[4] Eric Pelletier,[14,15,16] Jean-Baptiste Romagnan,[9,10] Simon Roux,[6]** Sébastien Santini,[1] Eleonora Scalco,[7] Sarah M. Schwenck,[6] Atsuko Tanaka,[4]§ Pierre Testor,[13] Thomas Vannier,[14,15,16] Flora Vincent,[4] Adriana Zingone,[7] Céline Dimier,[2,3,4] Marc Picheral,[9,10] Sarah Searson,[9,10]|| Stefanie Kandels-Lewis,[17,18] *Tara* Oceans Coordinators¶ Silvia G. Acinas,[19] Peer Bork,[17,20] Emmanuel Boss,[8] Colomban de Vargas,[2,3] Gabriel Gorsky,[9,10] Hiroyuki Ogata,[1]# Stéphane Pesant,[21,22] Matthew B. Sullivan,[6]** Shinichi Sunagawa,[17] Patrick Wincker,[14,15,16] Eric Karsenti,[4,18]* Chris Bowler,[4]* Fabrice Not,[2,3]*†† Pascal Hingamp,[1]*†† Daniele Iudicone[7]*††

Agulhas rings provide the principal route for ocean waters to circulate from the Indo-Pacific to the Atlantic basin. Their influence on global ocean circulation is well known, but their role in plankton transport is largely unexplored. We show that, although the coarse taxonomic structure of plankton communities is continuous across the Agulhas choke point, South Atlantic plankton diversity is altered compared with Indian Ocean source populations. Modeling and in situ sampling of a young Agulhas ring indicate that strong vertical mixing drives complex nitrogen cycling, shaping community metabolism and biogeochemical signatures as the ring and associated plankton transit westward. The peculiar local environment inside Agulhas rings may provide a selective mechanism contributing to the limited dispersal of Indian Ocean plankton populations into the Atlantic.

The Agulhas Current, which flows down the east coast of Africa, leaks from the Indo-Pacific Ocean into the Atlantic Ocean (*1*). This leakage, a choke point to heat and salt distribution across the world's oceans, has been increasing over the last decades (*2*). The influence of the Agulhas leakage on global oceanic circulation makes this area a sensitive lever in climate change scenarios (*3*). Agulhas leakage has been a gateway for planetary-scale water transport since the early Pleistocene (*4*), but diatom fossil records suggest that it is not a barrier to plankton dispersal (*5*). Most of the Agulhas leakage occurs through huge anticyclonic eddies known as Agulhas rings. These 100- to 400-km-diameter rings bud from Indian Ocean subtropical waters at the Agulhas Retroflection (*1*). Each year, up to half a dozen Agulhas rings escape the Indian Ocean, enter Cape Basin, and drift northwesterly across the South Atlantic, reaching the South American continent over the course of several years (*1*, *6*). During the transit of Agulhas rings, strong westerly "roaring forties" winds prevalent in the southern 40s and 50s latitudes cause intense internal cooling and mixing (*7*).

We studied the effect of Agulhas rings and the environmental changes they sustain on plankton dispersal. Plankton such as microalgae, which produce half of the atmospheric oxygen derived from photosynthesis each year, are at the base of open-ocean ecosystem food chains, thus playing an essential role in the functioning of the biosphere. Their dispersal is critical for marine ecosystem resilience in the face of environmental change (*8*). As part of the *Tara* Oceans expedition (*9*), we describe taxonomic and functional plankton assemblages inside Agulhas rings and across the three oceanic systems that converge at the Agulhas choke point: the western Indian Ocean subtropical gyre, the South Atlantic Ocean gyre, and the Southern Ocean below the Antarctic Circumpolar Current (Fig. 1).

Physical and biological oceanography of the sampling sites

The Indian, South Atlantic, and Southern Oceans were each represented by three sites sampled between May 2010 and January 2011 (Fig. 1 and table S1). A wide range of environmental conditions were encountered (*10*). We first sampled the two large contiguous Indian and South Atlantic subtropical gyres and the Agulhas ring structures that maintain the physical connection between them. On the western side of the Indian Ocean, station TARA_052 was characterized by tropical, oligotrophic conditions. Station TARA_064 was located within an anticyclonic eddy representing the Agulhas Current recirculation. Station TARA_065 was located at the inner edge of the Agulhas Current on the South African slope that feeds the Agulhas retroflection and Agulhas ring formation (*3*). In the South Atlantic Ocean, station TARA_070, sampled in late winter, was located in the eastern subtropical Atlantic basin. Station TARA_072 was located within the tropical circulation of the South Atlantic Ocean, and Station TARA_076 was at the northwest extreme of the South Atlantic subtropical gyre. Two stations (TARA_068 and TARA_078) from the west and east South Atlantic Ocean sampled Agulhas rings. Three stations (TARA_082, TARA_084, and TARA_085) in the Southern Ocean were selected to sample the Antarctic Circumpolar Current frontal system. Station TARA_082 sampled sub-Antarctic waters flowing northward along the Argentinian slope, waters that flow along the Antarctic Circumpolar Current (*11*) with characteristics typical

[1]Aix Marseille Université, CNRS, IGS UMR 7256, 13288 Marseille, France. [2]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [3]Sorbonne Universités, Université Pierre et Marie Curie UPMC, Université Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [4]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, F-75005 Paris, France. [5]Laboratoire de Physique des Océans (LPO) UMR 6523 CNRS-Ifremer-IRD-UBO, Plouzané, France. [6]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA. [7]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [8]School of Marine Sciences, University of Maine, Orono, ME, USA. [9]Sorbonne Universités, UPMC Université Paris 06, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [10]INSU-CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [11]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [12]Université de Toulon, Laboratoire PROTEE-EBMA E.A. 3819, BP 20132, 83957 La Garde Cedex, France. [13]CNRS, UMR 7159, Laboratoire d'Océanographie et du Climat LOCEAN, 4 Place Jussieu, 75005 Paris, France. [14]Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Génomique, Genoscope, 2 Rue Gaston Crémieux, 91057 Evry, France. [15]CNRS, UMR 8030, CP5706, Evry, France. [16]Université d'Evry, UMR 8030, CP5706, Evry, France. [17]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [18]Directors' Research, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [19]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM), CSIC, Passeig Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [20]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. [21]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [22]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [23]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD) UMR 8539, Ecole Normale Supérieure, 24 Rue Lhomond, 75231 Paris Cedex 05, France. ***Corresponding author. E-mail: villar@igs.cnrs-mrs.fr (E.V.); not@sb-roscoff.fr (F.N.); hingamp@igs.cnrs-mrs.fr (P.H.); iudicone@szn.it (D.I.); karsenti@embl.de (E.K.), cbowler@biologie.ens.fr (C.B.)** †These authors contributed equally to this work. ‡Present address: CNRS FR3631, Institut de Biologie Paris-Seine, F-75005 Paris, France; Sorbonne Universités, UPMC Université Paris 06, Institut de Biologie Paris-Seine (IBPS), F-75005 Paris, France. §Present address: Muroran Marine Station, Field Science Center for Northern Biosphere, Hokkaido University, Japan. ||Present address: CMORE, University Hawaii, Honolulu, USA. ¶*Tara* Oceans coordinators are listed at the end of this paper. #Present address: Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. **Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA. ††These authors contributed equally to this work.

of summer sub-Antarctic surface waters and are stratified by seasonal heating. Station TARA_084 was located on the southern part of the Antarctic Circumpolar Current, in the Drake Passage between the Polar Front and the South Antarctic Circumpolar Current front (*11*). Station TARA_085 was located on the southern edge of the South Antarctic Circumpolar Current front with waters typical of polar regions.

We compared overall plankton community structures between the three oceans using imaging and genetic surveys of samples from the epipelagic zone of each station (*12*). Prokaryote, phyto-, and zooplankton assemblages were similar across Indian and South Atlantic Ocean samples but different from Southern Ocean samples (Fig. 2A). In the Indian and South Atlantic Oceans, zooplankton communities were dominated by Calanoida, Cyclopoida (Oithonidae), and Poecilostomatoida copepods (*12*); phytoplankton communities were mainly composed of chlorophytes, pelagophytes, and haptophytes (*12*). In contrast, Southern Ocean zooplankton communities were distinguished by an abundance of *Limacina* spp. gastropods and Poecilostomatoida copepods. Southern Ocean phytoplankton were primarily diatoms and haptophytes. The divergence was even more conspicuous with respect to prokaryotes, in that picocyanobacteria, dominant in the Indian and South Atlantic Oceans, were absent in the Southern Ocean. The Southern Ocean had a high proportion of Flavobacteria and Rhodobacterales (*12*). Virus concentrations in the <0.2-μm size fractions were significantly lower in the southernmost Southern Ocean station (*13*). Viral particles were significantly smaller in two of the three Southern Ocean sampling sites, and two Southern Ocean viromes had significantly lower richness compared with the South Atlantic and Indian Oceans (*13*). Although nucleocytoplasmic large DNA viruses were similarly distributed in the South Atlantic and Indian Oceans (*12*), two Southern Ocean sites contained coccolithoviruses also found in the TARA_068 Agulhas ring but not in the other Indian and South Atlantic stations.

Biological connection across the Agulhas choke point

Genetic material as represented by ribosomal RNA gene (rDNA) sequences showed exchange patterns across the oceans (shared barcode richness) (*14*). Despite a smaller interface between the Indian and South Atlantic Oceans than either have with the Southern Ocean, more than three times as much genetic material was in common between the Indian and South Atlantic Oceans than either had with the Southern Ocean (Fig. 2B) (*15*). Indeed, the Indian–South Atlantic interocean shared barcodes richness (32 ± 5%) was not significantly different from typical intraocean values (37 ± 7%, Tukey post hoc, 0.95 confidence). Shared barcode richness involving the Southern Ocean was significantly lower (9 ± 3%) (Fig. 2C). We found that the proportion of whole shotgun metagenomic reads shared between samples, both intraoceanic and Indian–South Atlantic interocean similarities, were in the 18 to 30% range, whereas interocean similarities with Southern Ocean samples were only 5 to 6% (*16*). The statistically indistinguishable Indo-Atlantic intra- and interocean genetic similarities revealed a high Indo-Atlantic biological connection despite the physical basin discontinuity.

Nonetheless, differences on either side of the Agulhas choke point were evident. We found that prokaryote barcode richness was greater in the South Atlantic than in the Indian Ocean (Fig. 3A) (0.2- to 3-μm size fraction). The opposite trend characterized eukaryotes larger than 20 μm in size. We cannot rule out the possibility that the higher prokaryote diversity observed in the South Atlantic Ocean might be due to a protocol artifact resulting from a difference in prefiltration pore size from 1.6 μm (Indian Ocean) to 3 μm (South Atlantic and Southern Oceans). As also evident from the panoceanic *Tara* Oceans data set (*17*), smaller size fractions showed greater eukaryote diversity across the Agulhas system. In all size fractions that we analyzed, samples from the Southern Ocean were less diverse than samples from the South Atlantic Ocean and Indian Ocean (Fig. 3A).

When rDNA barcodes were clustered by sequence similarity and considered at operational taxonomic unit (OTU) level (*14*), more than half (57%) of the OTUs contained higher sub-OTU barcode richness in the Indian Ocean than in the South Atlantic Ocean, whereas less than a third (32%) of OTUs were richer in the South Atlantic Ocean, leaving only 11% as strictly cosmopolitan (Fig. 3B). Taken together, these 1307 OTUs represented 98% of the barcode abundance, indicating that the observed higher barcode richness within OTUs in the Indian Ocean was not conferred by the rare biosphere. Certain taxa displayed unusual sub-OTU richness profiles across the choke point. Consistent with their relatively large size, Opisthokonta (mostly copepods), Rhizaria (such as radiolarians), and Stramenopiles (in particular diatoms) had much higher sub-OTU barcode richness in the Indian Ocean, whereas only small-sized Hacrobia (mostly haptophytes) showed modest increased sub-OTU barcode richness in the South Atlantic Ocean. The plankton filtering that we observed in fractions above 20 μm through the Agulhas choke point might explain the reduction of marine nekton diversity from the Indian Ocean to the South Atlantic Ocean (*18*) by propagating up the food web (*19*).

In situ sampling of two Agulhas rings

To understand whether the environment of Agulhas rings, the main transporters of water across the choke point, might act as a biological filter between the Indian Ocean and the South Atlantic Ocean, we analyzed data collected in both a young and an old Agulhas ring. The young ring sampled at station TARA_068 was located in the Cape Basin, west of South Africa, where rings are often observed after their formation at the Agulhas Retroflection (*7*, *20*). It was a large Agulhas ring that detached from the retroflection about 9 to 10 months before sampling. This ring first moved northward and then westward in the Cape Basin while interacting with other structures (red track in Fig. 1) (*21*). Ocean color data collected by satellite showed that surface chlorophyll concentrations were higher in the Cape Basin than at the retroflection, suggesting that vigorous vertical

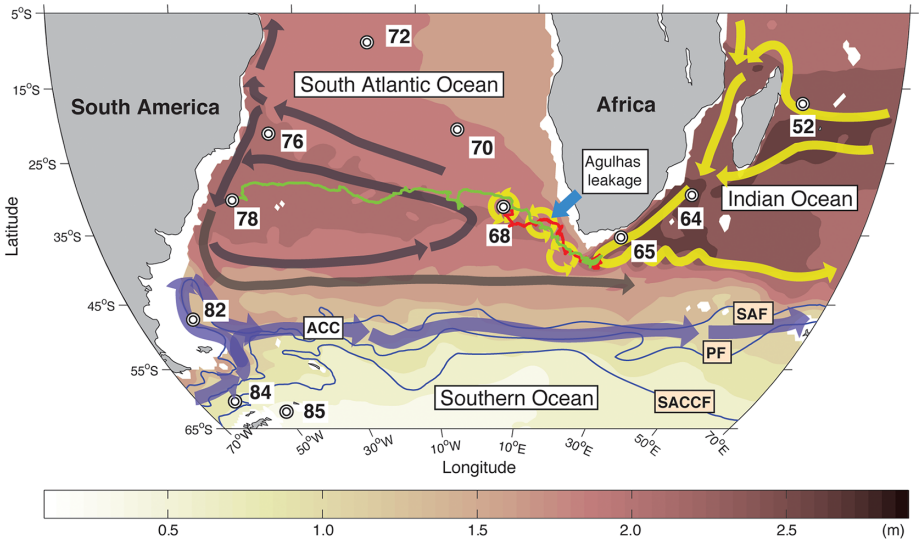


Fig. 1. The oceanic circulation around the Agulhas choke point and location of *Tara* Oceans stations. The map shows the location of sampling stations, together with trajectories of the young and old Agulhas rings (TARA_068 and TARA_078, red and green tracks, respectively). The stations here considered as representative of the main basins are (i) TARA_052, TARA_064, and TARA_065 for Indian Ocean; (ii) TARA_070, TARA_072, and TARA_076 for the South Atlantic Ocean, and (iii) TARA_082, TARA_084, and TARA_085 for the Southern Ocean. The mean ocean circulation is schematized by arrows (currents) and background colors [surface climatological dynamic height (0/2000 dbar from CARS2009; www.cmar.csiro.au/cars)] (*70*). Agulhas rings are depicted as circles. The Antarctic Circumpolar Current front positions are from (*13*).

mixing might have occurred in the Cape Basin (*22*). At the time of sampling, the anticyclonic Agulhas ring was 130 to 150 km in diameter, was about 30 cm higher than average sea surface height, and was flanked by a 130- to 150-km cyclonic eddy to the north and a larger (>200 km) one to the east (Fig. 4A) (*23*). Thermosalinograph data showed that filaments of colder, fresher water surrounded the young ring core (Fig. 4A) (*23*). To position the biological sampling station close to the ring core, a series of conductivity-temperature-depth (CTD) casts was performed (*23*, *24*). The young Agulhas ring had a surface temperature and salinity of 16.8°C and 35.7 practical salinity units (PSU), respectively, and the isopycnal sloping could be traced down to CTD maximal depth (900 to 1000 m). The core of the ring water was 5°C cooler than Indian Ocean subtropical source waters at similar latitudes (TARA_065) (table S1), typical for the subtropical waters south of Africa (17.8°C, 35.56 PSU, respectively) (*25*). The mixed layer of the young ring was deep (>250 m) compared with seasonal cycles of the mixed layer depths in the region (50 to 100 m) (Fig. 4C), typical of Agulhas rings (*26*). At larger scales (Fig. 4B) (*24*), steep spatial gradients were observed, with fresher and colder water in the Cape Basin than in the Agulhas Current because of both lateral mixing with waters from the south and surface fluxes. This confirms that the low temperature of the young Agulhas ring is a general feature of this Indian to South Atlantic Ocean transitional basin. Air-sea exchanges of heat and momentum promoted convection in the ring core, which was not compensated by lateral mixing and advection. The core of the Agulhas ring thus behaved as a subpolar environment traveling across a subtropical region.

At station TARA_078, we sampled a second structure whose origins were in the Agulhas Retroflection, likely a 3-year-old Agulhas ring. This old ring, having crossed the South Atlantic Ocean, was being absorbed by the western boundary current of the South Atlantic subtropical gyre. The structure sampled at station TARA_078 was characterized by a warm salty core (*27*). As for the young Agulhas ring sampled, the old ring also had a 100-m-deeper pycnocline than surrounding waters, typical of large anticyclonic structures.

The plankton assemblage of both Agulhas rings most closely resembled the assemblages found in Indian and South Atlantic samples (Fig. 2A). At higher resolution, barcodes (Fig. 2, B and C) and metagenomic reads (*16*) shared between the Agulhas rings and the Indian or South Atlantic samples showed that the young ring was genetically distinct from both Indian and South Atlantic samples,

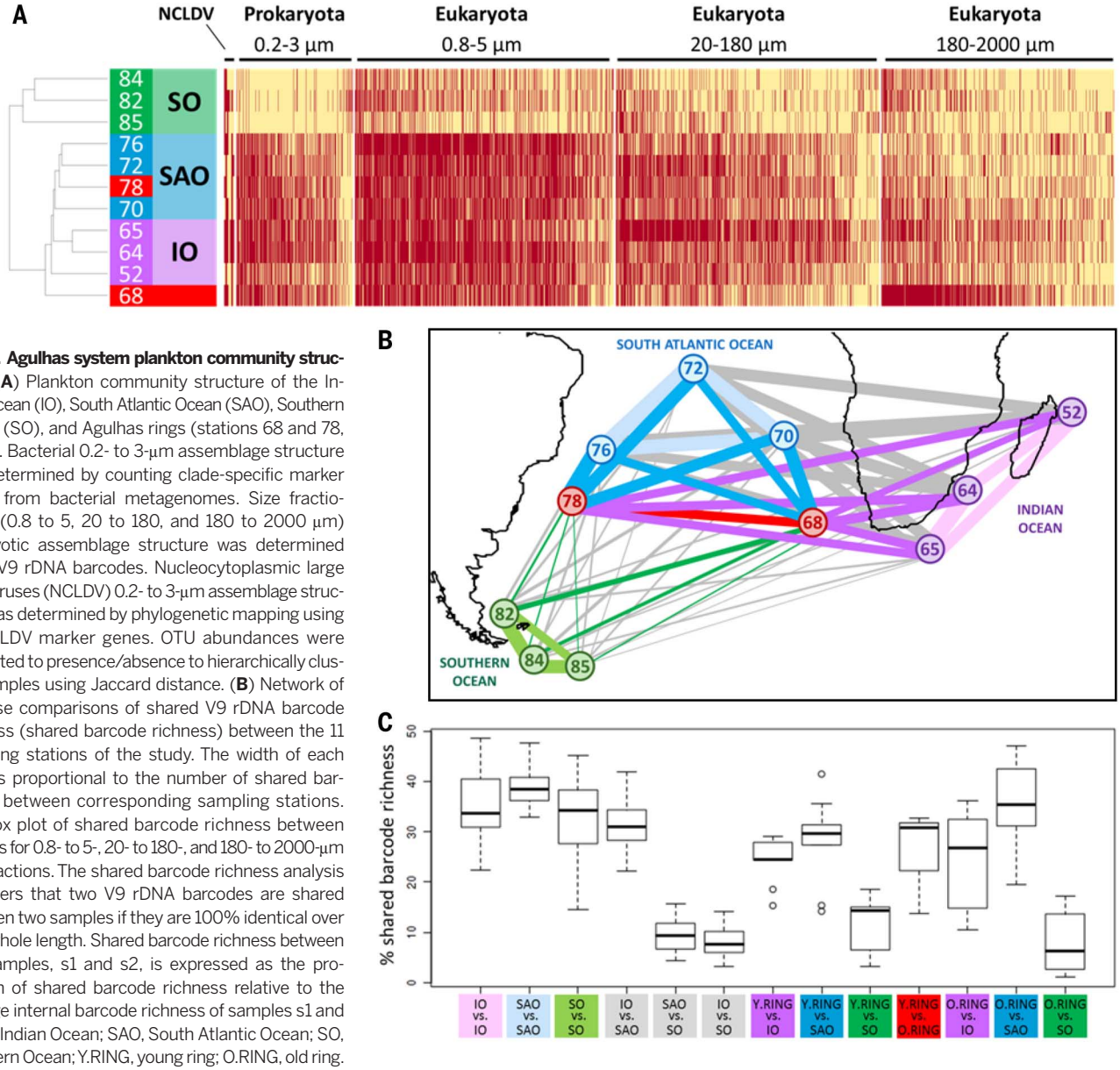


Fig. 2. Agulhas system plankton community structure. (**A**) Plankton community structure of the Indian Ocean (IO), South Atlantic Ocean (SAO), Southern Ocean (SO), and Agulhas rings (stations 68 and 78, in red). Bacterial 0.2- to 3-µm assemblage structure was determined by counting clade-specific marker genes from bacterial metagenomes. Size fractionated (0.8 to 5, 20 to 180, and 180 to 2000 µm) eukaryotic assemblage structure was determined using V9 rDNA barcodes. Nucleocytoplasmic large DNA viruses (NCLDV) 0.2- to 3-µm assemblage structure was determined by phylogenetic mapping using 16 NCLDV marker genes. OTU abundances were converted to presence/absence to hierarchically cluster samples using Jaccard distance. (**B**) Network of pairwise comparisons of shared V9 rDNA barcode richness (shared barcode richness) between the 11 sampling stations of the study. The width of each edge is proportional to the number of shared barcodes between corresponding sampling stations. (**C**) Box plot of shared barcode richness between stations for 0.8- to 5-, 20- to 180-, and 180- to 2000-µm size fractions. The shared barcode richness analysis considers that two V9 rDNA barcodes are shared between two samples if they are 100% identical over their whole length. Shared barcode richness between two samples, s1 and s2, is expressed as the proportion of shared barcode richness relative to the average internal barcode richness of samples s1 and s2. IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; Y.RING, young ring; O.RING, old ring.

whereas the old ring was similar to its surrounding South Atlantic samples (Tukey post hoc, 0.95 confidence). Light microscopy analyses revealed some plankton groups specific to the young Agulhas ring, such as *Pseudo-nitzschia* spp., which represented 20% of the phytoplankton counts but less than 10% in all other stations (*12*). Other potentially circumstantial plankton characteristic of the young Agulhas ring included the tintinnid *Dictyocysta pacifica* (*12*), the diatom *Corethron pennatum* (*12*), and the dinoflagellate *Tripos limulus* (*12*). A tiny (less than 15 μm long) pennate diatom from the genus *Nanoneis,* which we saw only in the young Agulhas ring and Indian Ocean stations around the African coasts (*28*), was an example of the Indo-Atlantic plankton diversity filtering observed at rDNA barcode level and corroborated by microscopy. OTU clustered barcodes revealed a variety of young Agulhas ring sub-OTU richness patterns compared with source and destination oceans (Fig. 5A). Among Copepoda, *Gaetanus variabilis* and *Corycaeus speciosus* were the more cosmopolitan species (Fig. 5B), whereas *Bradya* species found in the young ring were mainly similar to those from the Indian Ocean. *Acartia negligens* and *Neocalanus robustior* displayed high levels of barcode richness specific to each side of the Agulhas choke point. Bacillariophyceae were heavily filtered from Indian to South Atlantic Oceans (Fig. 5C), and most OTUs (17 out of 20) were absent in the young ring, suggesting that diversity filtering could take place earlier in the ring's 9-month history. Consistent with the observed particularities of the plankton in the young ring, continuous underway optical measurements showed that the ring core photosynthetic community differed from surrounding waters (*29*–*31*). Intermediate size cells, and relatively low content of photoprotective pigments, reflected low growth irradiance and suggested a transitional physiological state. Thus, the plankton community in the young Agulhas ring had diverged from plankton communities typical of its original Indian waters but, even 9 months after formation, had not converged with its surrounding South Atlantic waters.

Deep mixing in Agulhas rings promotes plankton bloom

The upper water column of the young ring showed a high nitrite concentration (>0.5 mmol m^{-3}) (Fig. 4D) (*32*). This observation, along with its particularly deep mixed layer (>250 m), suggested that as Agulhas rings proceed westward in the Cape Basin, vigorous deep mixing of their weakly stratified waters may have entrained nitrate and stimulated phytoplankton blooms. Typically, fresh organic material would then either be exported as sinking particles or locally recycled, sustaining heterotrophic production of ammonium that would, in turn, be consumed by photoautotrophs in the euphotic layer but nitrified below. The resulting nitrite, eventually oxidized to nitrate, might remain evident at subsurface as observed in the nitrite anomaly of the young ring detected here. This hypothesis was supported by numerical simulations of the Massachusetts Institute of Technology General Circulation Model (*33*), which resolved Agulhas rings, their phytoplankton populations, and associated nutrient cycling (Fig. 6A). We tracked 12 Agulhas rings in the ocean model and characterized their near-surface biogeochemical cycles (Fig. 6B) (*34*). As the rings moved westward, storms enhanced surface heat loss, stimulating convection and the entrainment of nitrate. In the model simulations, proliferation of phytoplankton generated subsurface nitrite, which persisted because phytoplankton were light-limited at depth and because nitrification was suppressed by light at the surface (*35*). The associated blooms were dominated by large opportunistic phytoplankton and nitrate-metabolizing *Synechococcus* spp. analogs, whereas populations of *Prochlorococcus* spp. analogs dominated the quiescent periods (*34*). Each of the 12 simulated Agulhas rings exhibited this pattern in response to surface forcing by weather systems, and all rings maintained a persistent

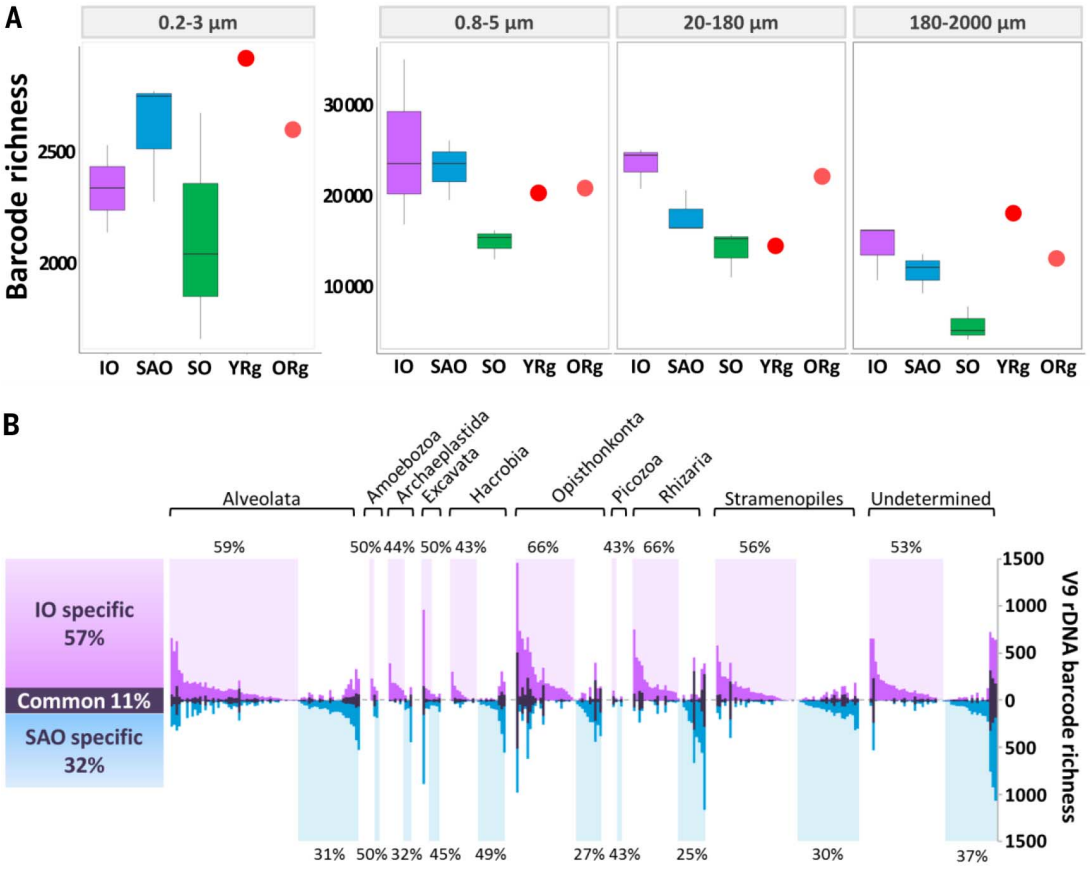


Fig. 3. Diversity of plankton populations specific to Indian and Atlantic Oceans. (**A**) Box plot of 16S (0.2 to 3 μm) and V9 rDNA barcodes richness (0.8- to 5-, 20- to 180-, and 180- to 2000-μm size fractions). Each box represents three sampling stations combined into Indian, South Atlantic, and Southern Ocean. Single Agulhas ring stations are represented as red (young ring) and orange (old ring) crosses. (**B**) Plankton sub-OTU richness filtering across the Agulhas choke point. Each vertical bar represents a single eukaryotic plankton OTU, each of which contains >10 distinct V9 rDNA barcodes (*14*). For each OTU are represented the number of distinct barcodes (sub-OTU richness) found exclusively in the South Atlantic Ocean (blue), exclusively in the Indian Ocean (pink), and in both South Atlantic Ocean and Indian Ocean (gray). OTUs are grouped by taxonomic annotation (indicated above the bar plot). For each taxonomic group, the percentage of OTUs with higher sub-OTU richness in the Indian Ocean (shaded in pink) or in the South Atlantic Ocean (shaded in blue) is indicated, respectively, at the top and bottom of the bar plot. A total of 1307 OTUs are presented, representing 98% of total V9 rDNA barcode abundance.

subsurface nitrite maximum in the region, as observed in TARA_068 and in other biogeochemical surveys (*36*).

The nitrite peak observed at TARA_068 in the young Agulhas ring was associated with a differential representation of nitrogen metabolism genes between the ring and the surrounding South Atlantic and Indian Oceans metagenomes derived from 0.2- to 3-μm size fractions (Fig. 7) (*37*). Agulhas ring overrepresented KEGG (Kyoto Encyclopedia of Genes and Genomes) orthologs (KOs) were involved in both nitrification and denitrification, likely representing the overlap between plankton assemblages involved in the conversion of nitrate to nitrite on the one hand and in denitrification of the accumulating nitrite on the other. Distinct KOs involved in successive denitrification steps were found to be encoded by similar plankton taxa. For instance, KO10945 and KO10946 (involved in ammonium nitrification) and KO00368 (subsequently involved in nitrite to nitrous oxide denitrification) appeared mostly encoded by Nitrosopumilaceae archaea. KO00264 and KO01674 (involved in ammonium assimilation) were mostly assigned to eukaryotic Mamiellales, whereas the opposite KO00367 and KO00366 (involved in dissimilatory nitrite reduction to ammonium), followed by KO01725 (involved in ammonium assimilation), were encoded by picocyanobacteria. In the specific case of the picocyanobacteria, metagenomic reads corresponding to *nirA* genes showed that the observed young Agulhas ring KO00366 (dissimilatory nitrite reduction) enrichment was mainly due to the overrepresentation of genes from *Prochlorococcus* (Fig. 8B). This enrichment was found to be associated with a concomitant shift in population structure from *Prochlorococcus* high-light II ecotypes (HLII, mostly lacking *nirA* genes) to codominance of high-light I (HLI) and low-light I (LLI) ecotypes. Indeed, among the several *Prochlorococcus* and *Synechococcus* ecotypes identified based on their genetic diversity and physiology (*38*, *39*), neutral marker (*petB*) (Fig. 8A) recruitments showed that dominant clades in the Indian Ocean upper mixed layer were *Prochlorococcus* HLII and *Synechococcus* clade II, as expected given the known (sub)tropical preference of these groups (*40*). Both clades nearly completely disappeared (less than 5%) in the mixed cold waters of the young ring and only began to increase again when the surface water warmed up along the South Atlantic Ocean transect. Conversely, young ring water was characterized by a large proportion of *Prochlorococcus* HLI and LLI and *Synechococcus* clade IV, two clades typical of temperate waters. Besides temperature, the *Prochlorococcus* community shift from HLII to HLI + LLI observed in the young ring was likely also driven by the nitrite anomaly. Indeed, whereas most *Synechococcus* strains isolated so far are able to

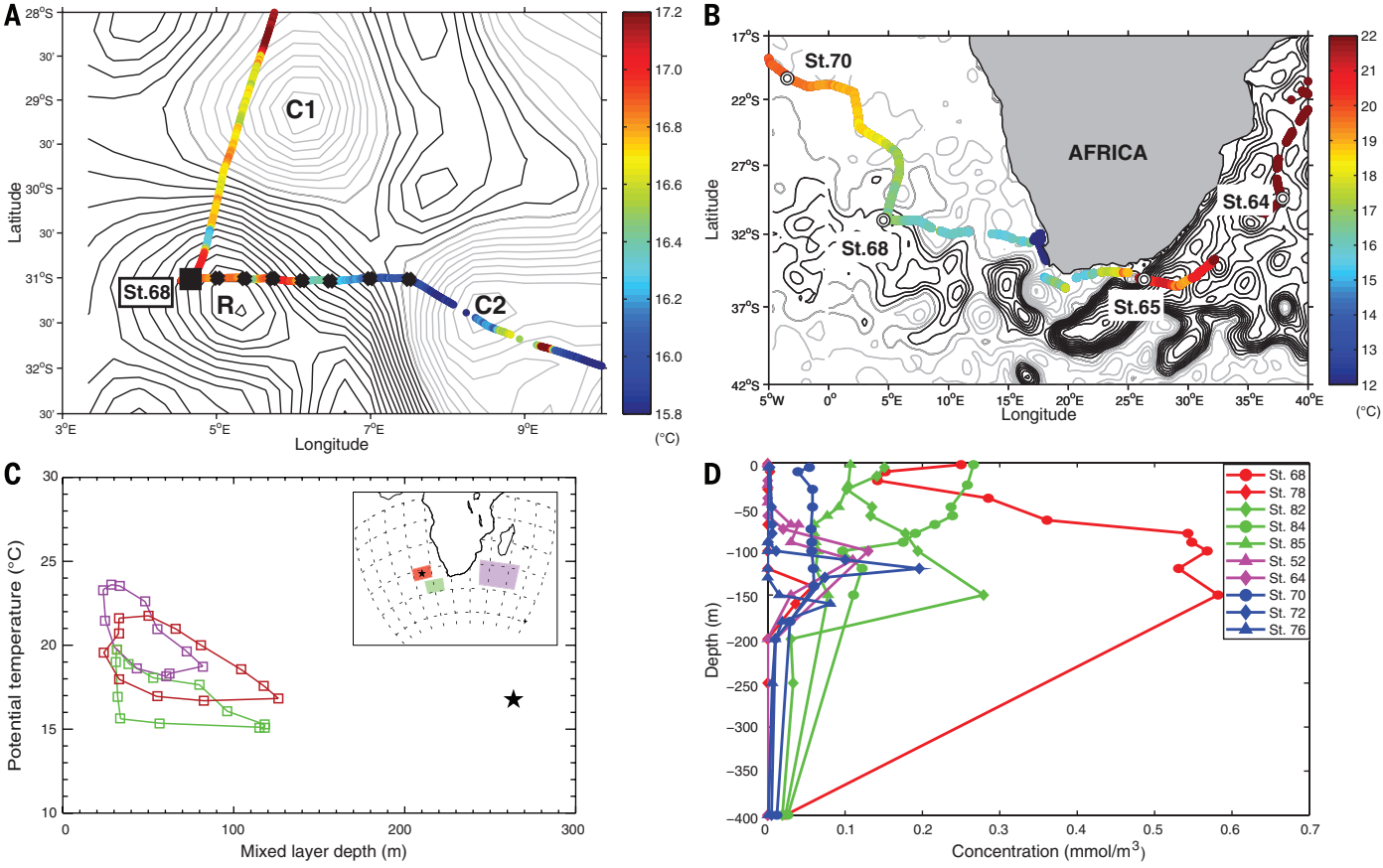


Fig. 4. Properties of the young Agulhas ring (TARA_068). (**A**) Daily sea surface height around young Agulhas ring station TARA_068 [absolute dynamic topography (ADT) from www.aviso.altimetry.fr]. R, C1, and C2, respectively, denote the centers of the Agulhas ring and two cyclonic eddies. The contour interval is 0.02 dyn/m. The ADT values are for 13 September 2010. Light gray isolines, ADT < 0.46 dyn/m. The crosses indicate the CTD stations, and the square symbol indicates the position of the biological station TARA_068. The biological station coincides with the westernmost CTD station. ADT is affected by interpolation errors, which is why CTD casts were performed at sea so as to have a fine-scale description of the feature before defining the position of the biological station (*23*). Superimposed are the continuous underway temperatures (°C) from the on-board thermosalinograph. (**B**) Same as (A) but at the regional scale. Round symbols correspond to biological sampling stations. The contour interval is 0.1 dyn/m. (**C**) Seasonal distribution of the median values of the mixed layer depths and temperatures at 10 m (from ARGO) provided by the IFREMER/LOS Mixed Layer Depth Climatology L2 database (www.ifremer.fr/cerweb/deboyer/mld) updated to 27 July 2011. The mixed layer is defined using a temperature criterion. The star symbol represents the young ring station TARA_068. (Inset) Geographic position of the areas used to select the mixed layer and temperature data. The mixed layer depth measured at TARA_068 is outside the 90th percentile of the distribution of mixed layer depths for the same month for both the subtropical (red and magenta) regions. The temperature matches the median for the same month and region of sampling. (**D**) Nitrite (NO_2) concentrations from CTD casts at different sampling sites (expressed in mmol/m^3).

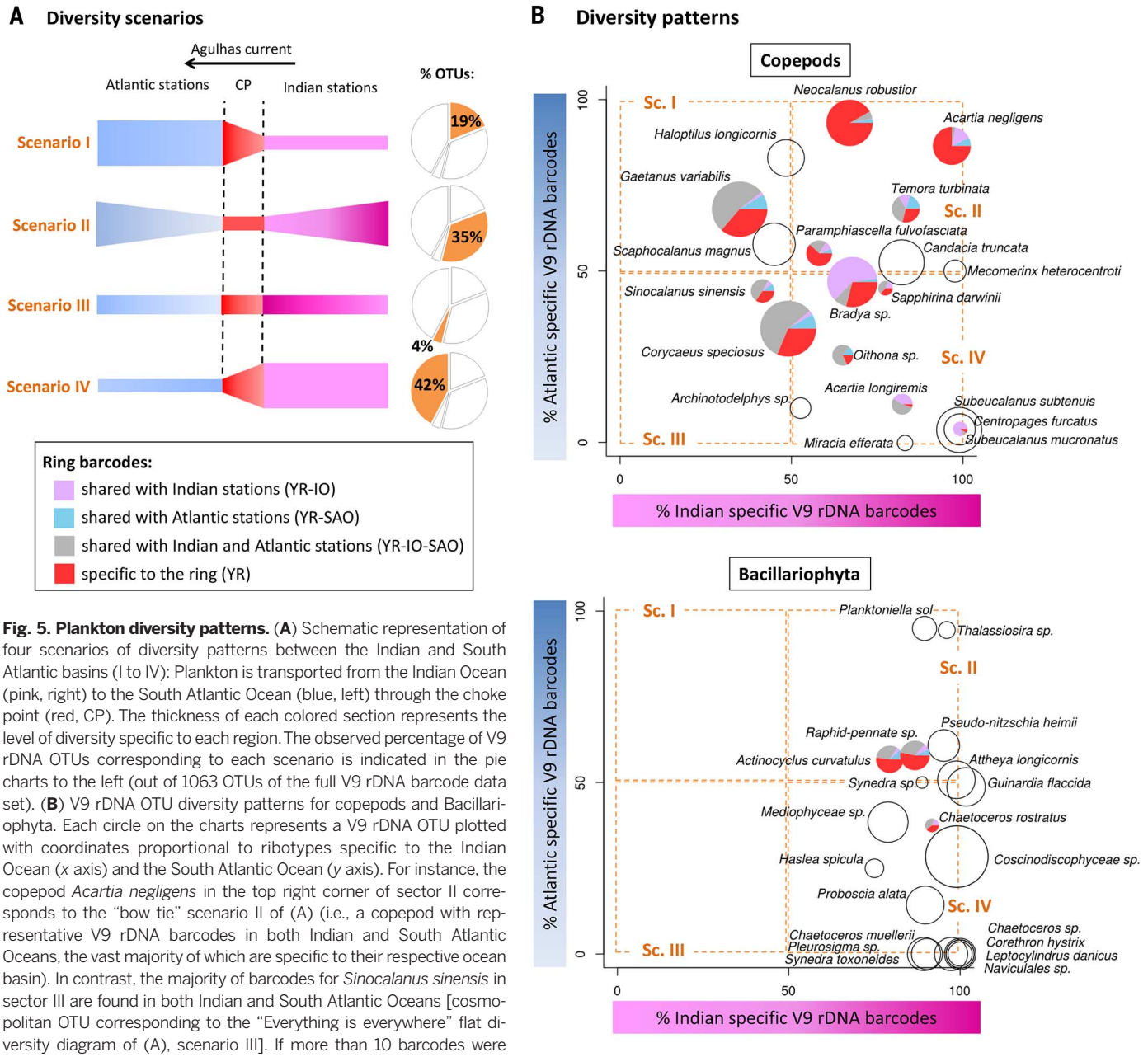


Fig. 5. Plankton diversity patterns. (**A**) Schematic representation of four scenarios of diversity patterns between the Indian and South Atlantic basins (I to IV): Plankton is transported from the Indian Ocean (pink, right) to the South Atlantic Ocean (blue, left) through the choke point (red, CP). The thickness of each colored section represents the level of diversity specific to each region. The observed percentage of V9 rDNA OTUs corresponding to each scenario is indicated in the pie charts to the left (out of 1063 OTUs of the full V9 rDNA barcode data set). (**B**) V9 rDNA OTU diversity patterns for copepods and Bacillariophyta. Each circle on the charts represents a V9 rDNA OTU plotted with coordinates proportional to ribotypes specific to the Indian Ocean (*x* axis) and the South Atlantic Ocean (*y* axis). For instance, the copepod *Acartia negligens* in the top right corner of sector II corresponds to the "bow tie" scenario II of (A) (i.e., a copepod with representative V9 rDNA barcodes in both Indian and South Atlantic Oceans, the vast majority of which are specific to their respective ocean basin). In contrast, the majority of barcodes for *Sinocalanus sinensis* in sector III are found in both Indian and South Atlantic Oceans [cosmopolitan OTU corresponding to the "Everything is everywhere" flat diversity diagram of (A), scenario III]. If more than 10 barcodes were found in the young Agulhas ring (TARA_068), their distribution is indicated in a pie chart (colors are coded in the legend inset); otherwise, the OTU is represented by an empty circle. Circle sizes are proportional to the number of considered barcodes for each OTU. The Bacillariophyta OTU defined as *Raphid pennate* sp. likely corresponds to the *Pseudo-nitzschia* cells observed by light microscopy.

use nitrate, nitrite, and ammonium, only the *Prochlorococcus* LLI and IV and some populations of HL clades, having acquired the *nirA* gene by lateral gene transfer, are able to assimilate nitrite. In the young ring, overrepresentation of cyanobacterial orthologs involved in nitrite reduction could thus have resulted from environmental pressure selecting LLI (87% of the *nirA* recruitments) and HL populations (13%) that possessed this ability. Because the capacity to assimilate nitrite in this latter ecotype reflects the availability of this nutrient in the environment (*41*), these in situ observations of picocyanobacteria indicated that the nitrogen cycle disturbance occurring in the young ring exerts community-wide selective pressure on Agulhas ring plankton.

Discussion

We found that whether or not the Agulhas choke point is considered a barrier to plankton dispersal depends on the taxonomic resolution at which the analysis is performed. At coarse taxonomic resolution, our observations of Indo-Atlantic continuous plankton structure—from viruses to fish larvae—suggested unlimited dispersal, consistent with previous reports (*5*, *42*). However, at finer resolution, our genetic data revealed that the Agulhas choke point strongly affects patterns of plankton genetic diversity. As anticipated in (*5*), the diversity filtering by Agulhas rings likely escaped detection using fossil records because of the limited taxonomic resolution afforded by fossil diatom morphology (*42*). The community-wide evidence presented here confirms observations on individual living species (*43*, *44*), suggesting that dispersal filters mitigate the panmictic ocean hypothesis for plankton above 20 μm.

The lower diversity we observed in the South Atlantic Ocean for micro- and mesoplankton (>20 μm) may be due to local abiotic/biotic pressure or to limitations in dispersal (*33*, *45*). Biogeography emerging from a model with only neutral drift (*46*) predicts

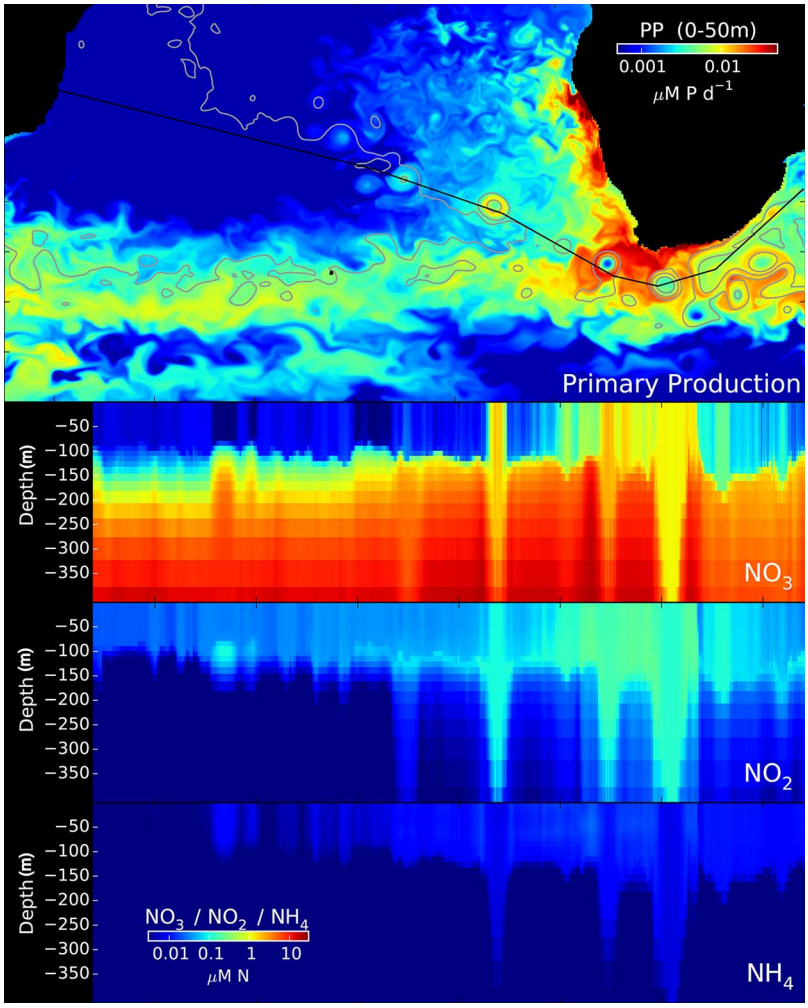


Fig. 6. Modeled nitrogen stocks along Agulhas ring track. (**Top**) Simulated primary production (PP) in the Agulhas system using the MIT-GCM model. The solid black line shows the average northwesterly path of 12 distinct virtual Agulhas rings tracked over the course of the simulation. Color scale for PP is given in the top right inset, with warmer colors indicating higher PP. (**Bottom**) Modeled profiles of NO_3, NO_2, and NH_4 along the Agulhas ring average track (*x* axis) presented in (A). The *y* axis is the depth (in meters) in the water column. The color scale is given in the bottom left inset, with warmer colors indicating higher concentrations of nitrogen compounds.

basin-to-basin genetic differences that are qualitatively consistent with our data. However, the increased proportion of *Prochlorococcus* HL populations carrying the *nirA* gene in the young Agulhas ring indicates that selection is at work in Agulhas rings. Based on our analysis of two Agulhas rings, we propose that environmental disturbances in Agulhas rings reshape their plankton diversity as they travel from the Indian Ocean to the South Atlantic Ocean. Such selective pressure may contribute to the South Atlantic Ocean plankton diversity shift relative to its upstream Indo-Pacific basin. Thus, environmental selection applied at a choke point in ocean circulation may constitute a barrier to dispersal (*47*, *48*). Furthermore, we show that taxonomic groups were not equally affected by the ring transport, both within and between phyla, with a noticeable effect of organism size. The differential effects due to organism size highlight the difficulty in generalizing ecological and evolutionary rules from limited sampling of species or functional types.

Considering the sensitivity of Agulhas leakage to climate change (*1*, *49*), better understanding of the plankton dynamics in Agulhas rings will be required if we are to understand and predict ecosystem resilience at the planetary scale. Considering the breadth of changes already observed in the 9-month-old Agulhas ring, it would be interesting to acquire samples from specific Agulhas rings tracked from early formation to dissipation. Finally, our data suggest that the abundance of Indian Ocean species in South Atlantic Ocean sedimentary records, used as proxies of Agulhas leakage intensity (*4*), may actually also depend on the physical and biological characteristics of the Agulhas rings.

Materials and methods

Sampling

The *Tara* Oceans sampling protocols schematized in Karsenti *et al.* (*9*) are described in Pesant *et al.* (*50*); specific methods for 0.8- to 5-, 20- to 180-, and 180- to 2000-μm size fractions in de Vargas *et al.* (*17*); for 0.2- to 3-μm size fractions in Sunagawa *et al.* (*51*); and for <0.2-μm size fraction in Brum *et al.* (*52*). Due to their fragility, 1.6-μm glass fiber filters initially used for prokaryote sampling were replaced by more resistant 3-μm polycarbonate filters from station TARA_066 onward. In the present text, both 0.2- to 1.6-μm and 0.2- to 3-μm prokaryote size fractions are simply referred to as 0.2 to 3 μm.

Data acquisition

A range of analytical methods covering different levels of taxonomic resolution (pigments, flow cytometry, optical microscopy, marker gene barcodes, and metagenomics) were used to describe the planktonic composition at each sampled station. Viruses from the <0.2 μm size fraction were studied by epifluorescence microscopy, by quantitative transmission electron microscopy, and by sequencing DNA as described in Brum *et al.* (*52*). Flow cytometry was used to discriminate high-DNA-content bacteria (HNA), low-DNA-content bacteria (LNA), *Prochlorococcus* and *Synechococcus* picocyanobacteria, and two different groups (based on their size) of photosynthetic picoeukaryotes, as described previously (*53*). Pigment concentrations measured by high-performance liquid chromatography (HPLC) were used to estimate the dominant classes of phytoplankton using the CHEMTAX procedure (*54*). Tintinnids, diatoms, and dinoflagellates were identified and counted by light microscopy from the 20- to 180-μm lugol or formaldehyde fixed-size fraction. Zooplankton enumeration was performed on formol fixed samples using the ZOOSCAN semi-automated classification of digital images (*55*). Sequencing, clustering, and annotation of 18*S*-V9 rDNA barcodes are described in de Vargas *et al.* (*17*). Metagenome sequencing, assembly, and annotation are described in Sunagawa *et al.* (*51*). NCLDV taxonomic assignations in the 0.2- to 3-μm samples were carried out using 18 lineage-specific markers as described in Hingamp *et al.* (*56*). Virome sequencing and annotation are described in Brum *et al.* (*52*). Samples and their associated contextual data are described at PANGAEA (*57*–*59*).

Data analysis

Origin of sampled Agulhas rings

Using visual and automated approaches, the origins of the TARA_068 and TARA_078 stations were traced back from the daily altimetric data (Fig. 1) (*21*). The automated approach used either the Lagrangian tracing of numerical particles initialized in the center of a given structure and transported by the geostrophic velocity field calculated from sea surface height gradients, or the connection in space and time of adjacent extreme values in sea level anomaly maps.

V9 rDNA barcodes

To normalize for differences in sequencing effort, V9 rDNA barcode libraries were resampled 50 times for the number of reads corresponding to the smallest library in each size fraction: 0.8 to 5 μm, 776,358 reads; 20 to 180 μm, 1,170,592 reads; and 180 to 2000 μm, 767,940 reads. V9 rDNA barcode counts were then converted to the average number

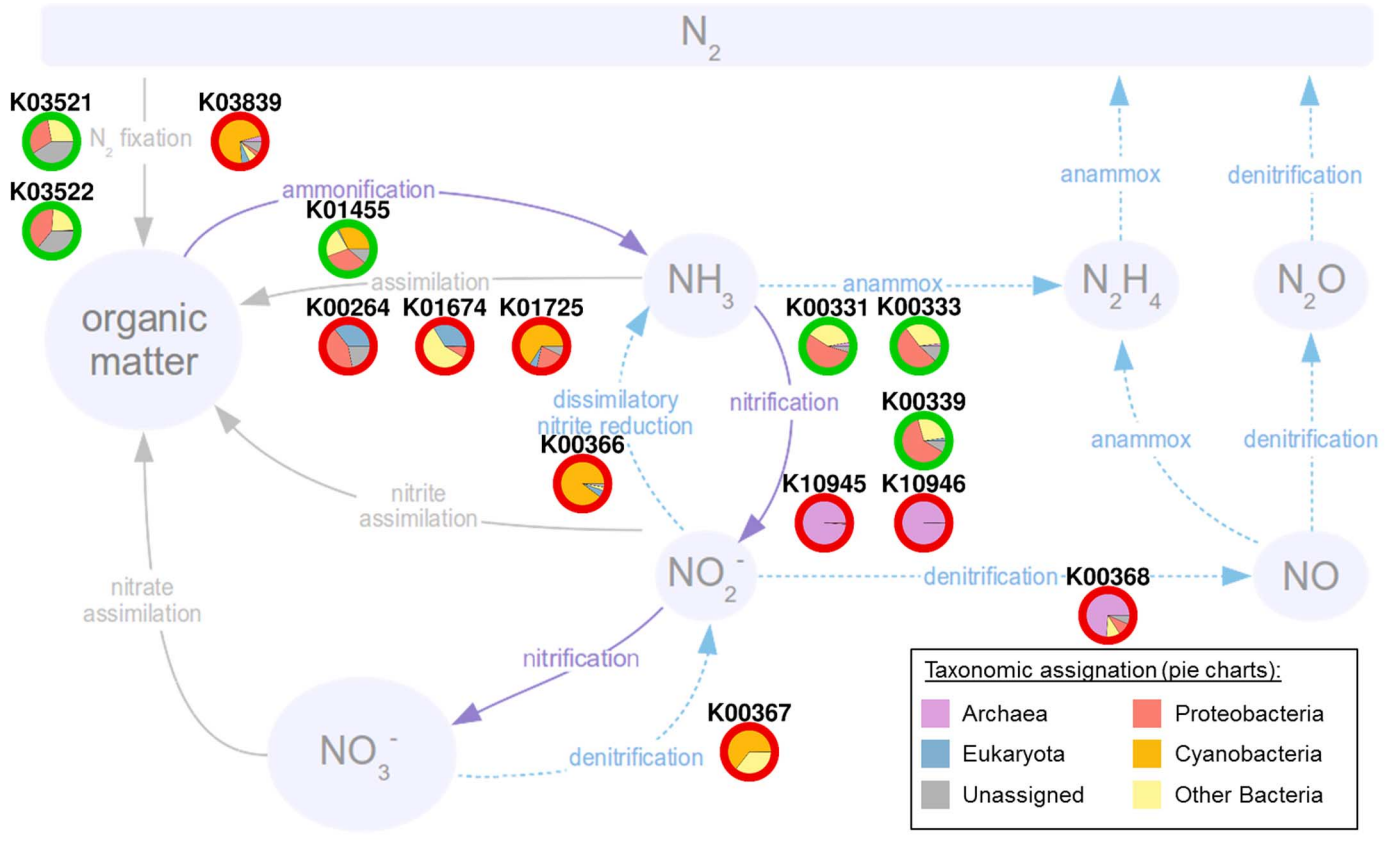


Fig. 7. Nitrite anomaly in the young Agulhas ring is accompanied by shifts in nitrogen pathway–related genes. Metagenomic over- and underrepresented nitrogen pathway genes in young Agulhas ring. Over- (red circles) and under- (green circles) represented metagenome functional annotations (KEGG Orthologs, KO#) involved in the nitrogen pathway in the young ring compared to Indian and South Atlantic Oceans reference stations, at surface and deep chlorophyll maximum depth. Pie charts inside circles represent the taxonomic distribution for each ortholog.

of times seen in the 50 resampling events, and barcodes with less than 10 reads were removed as potential sequencing artifacts. We used downsampled barcode richness (number of distinct V9 rDNA barcodes) as a diversity descriptor because using V9 rDNA barcode abundances to compare plankton assemblages would likely be biased due to (i) technical limitations described in de Vargas *et al.* (*17*) and (ii) seasonality effects induced by the timing of samplings (table S1). Barcode richness was well correlated with Shannon and Simpson indexes (0.94 and 0.78, respectively). The shared barcode richness between each pair of samples (*14*) was estimated by counting, for the three larger size fractions (0.8 to 5, 20 to 180, and 180 to 2000 μm), the proportion of V9 rDNA barcodes 100% identical over their whole length. V9 rDNA barcodes were clustered into OTUs by swarm clustering as described by de Vargas *et al.* (*17*). The sub-OTU richness comparison between two samples s1 and s2 (*14*) produces three values: the number of V9 rDNA barcodes in common, the number of V9 rDNA barcodes unique to s1, and the number of V9 rDNA barcodes unique to s2. These numbers can be represented directly as bar graphs (Fig. 3B) or as dot plots of specific V9 rDNA barcode richness (Fig. 5).

Metagenomic analysis

Similarity was estimated using whole shotgun metagenomes for all four available size fractions (0.2 to 3, 0.8 to 5, 20 to 180, and 180 to 2000 μm). Because pairwise comparisons of all raw metagenome reads are intractable given the present data volume, we used a heuristic in which two metagenomic 100–base pair (bp) reads were considered similar if at least two nonoverlapping 33-bp subsequences were strictly identical (Compareads method) (*60*). For prokaryotic fractions (0.2 to 3 μm), taxonomic abundance was estimated using the number of $16S_{mi}$tags (*51*). The functional annotation, taxonomic assignation, and gene abundance estimation of the panoceanic Ocean Microbial Reference Gene Catalog (OM-RGC) (243 samples, including all those analyzed here) generated from *Tara* Oceans 0.2- to 3-μm metagenomic reads are described in Sunagawa *et al.* (*51*). Gene abundances were computed for the set of genes annotated to the nitrogen metabolism KO (*61*) group by counting the number of reads from each sample that mapped to each KO-associated gene. Abundances were normalized as reads per kilobase per million mapped reads (RPKM). Gene abundances were then aggregated (summed) for each KO group. To compare abundances between the young ring (TARA_068) and other stations, a *t* test was used. KOs with a *P* value <0.05 and a total abundance (over all stations) >10 were considered as significant (*37*). *Prochlorococcus* and *Synechococcus* community composition was analyzed in the 0.2- to 3-μm size fraction at the clade level by recruiting reads targeting the high-resolution marker gene *petB*, coding for cytochrome b_6 (*62*). The *petB* reads were first extracted from metagenomes using Basic Local Alignment Search Tool (BLASTx+) against the *petB* sequences of *Synechococcus* sp. WH8102 and *Prochlorococcus marinus* MED4. These reads were subsequently aligned against a reference data set of 270 *petB* sequences using BLASTn (with parameters set at -G 8 -E 6 -r 5 -q -4 -W 8 -e 1 -F "m L" -U T). *petB* reads exhibiting >80% identity over >90% of sequence length were then taxonomically assigned to the clade of the best BLAST hit. Read counts per clade were normalized based on the sequencing effort for each metagenomic sample. A similar approach was used with *nirA* (KO 00366) and *narB* genes (KO 00367), which were highlighted in the nitrogen-related KO analysis (Fig. 7). Phylogenetic assignment was realized at the highest possible taxonomic level using a reference data set constituted of sequences retrieved from Cyanorak v2 (www.sb-roscoff.fr/cyanorak/) and Global Ocean Sampling (*41*, *63*) databases.

Nitrogen cycle modeling

Numerical simulations of global ocean circulation were based on the Massachusetts Institute of Technology General Circulation Model (MIT-GCM) (*64*), incorporating biogeochemical and ecological components (*65*, *66*). It resolved mesoscale

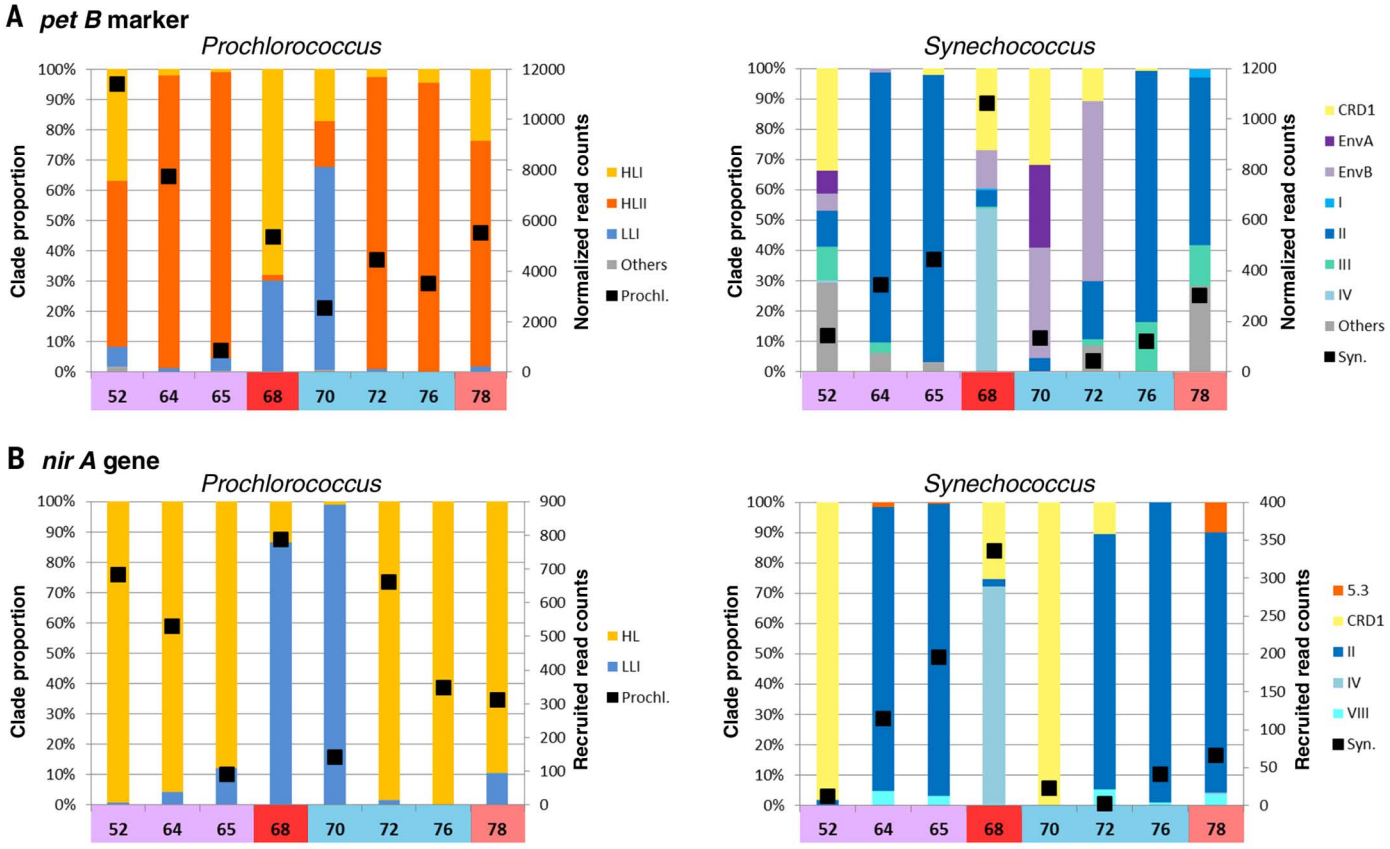


Fig. 8. Picocyanobacterial clade shift in the young Agulhas ring. (**A**) Relative abundance of *Prochlorococcus* and *Synechococcus* clades, estimated by *petB* read recruitments from 0.2- to 3-µm metagenomes. Solid squares correspond to read counts normalized based on the sequencing effort (right axis). (**B**) Relative abundance of *nirA* gene from *Prochlorococcus* and *Synechococcus* clades estimated by number of reads recruited from 0.2- to 3-µm metagenomes. The bar colors correspond to cyanobacterial clades indicated in the inset legends for each panel. Solid squares correspond to the number of reads recruited (right axis). Data are shown for stations TARA_052 to TARA_078 only, because too few cyanobacteria were found in Southern Ocean stations TARA_082, TARA_084, and TARA_085.

features in the tropics and was eddy-permitting in subpolar regions. The physical configurations were integrated from 1992 to 1999 and constrained to be consistent with observed hydrography and altimetry (*67*). Three inorganic fixed nitrogen pools were resolved—nitrate, nitrite, and ammonium—as well as particulate and dissolved detrital organic nitrogen. Phytoplankton types were able to use some or all of the fixed nitrogen pools. Aerobic respiration and remineralization by heterotrophic microbes was parameterized as a simple sequence of transformations from detrital organic nitrogen, to ammonium, then nitrification to nitrite and nitrate. In accordance with empirical evidence (*35*), nitrification was assumed to be inhibited by light. Nitrification is described in the model by simple first-order kinetics, with rates tuned to qualitatively capture the patterns of nitrogen species in the Atlantic (*66*).

Continuous spectral analysis

A continuous flow-through system equipped with a high-spectral-resolution spectrophotometer (AC-S, WET Labs, Inc.) was used for data collection during the *Tara* Oceans expedition, as described previously (*68*). Phytoplankton pigment concentrations, estimates of phytoplankton size γ, total chlorophyll a concentration, and particulate organic carbon (POC) are derived from the absorption and attenuation spectra (*69*) for the 1-km^2-binned *Tara* Oceans data set available at PANGAEA (http://doi.pangaea.de/10.1594/PANGAEA.836318).

REFERENCES AND NOTES

1. A. Biastoch, C. W. Böning, J. R. E. Lutjeharms, Agulhas leakage dynamics affects decadal variability in Atlantic overturning circulation. *Nature* **456**, 489–492http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=19037313&dopt=Abstract (2008). doi: 10.1038/nature07426
2. A. Biastoch, C. W. Böning, F. U. Schwarzkopf, J. R. E. Lutjeharms, Increase in Agulhas leakage due to poleward shift of Southern Hemisphere westerlies. *Nature* **462**, 495–498 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=19940923&dopt=Abstract (2009). doi: 10.1038/nature08519
3. L. M. Beal *et al*., On the role of the Agulhas system in ocean circulation and climate. *Nature* **472**, 429–436 (2011). doi: 10.1038/nature09983; pmid: 21525925
4. F. J. C. Peeters *et al*., Vigorous exchange between the Indian and Atlantic oceans at the end of the past five glacial periods. *Nature* **430**, 661–665 (2004). doi: 10.1038/nature02785; pmid: 15295596
5. P. Cermeño, P. G. Falkowski, Controls on diatom biogeography in the ocean. *Science* **325**, 1539–1541 (2009). pmid: 19762642
6. A. L. Gordon, Oceanography: The brawniest retroflection. *Nature* **421**, 904–905 (2003). doi: 10.1038/421904a; pmid: 12606984
7. H. M. van Aken *et al*., Observations of a young Agulhas ring, Astrid, during MARE in March 20. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **50**, 167–195 (2003). doi: 10.1016/S0967-0645(02)00383-1
8. J. R. Bernhardt, H. M. Leslie, Resilience to climate change in coastal marine ecosystems. *Annu. Rev. Mar. Sci.* **5**, 371–392 (2013). doi: 10.1146/annurev-marine-121211-172411; pmid: 22809195
9. E. Karsenti *et al*., A holistic approach to marine eco-systems biology. *PLOS Biol.* **9**, e1001177 (2011). doi: 10.1371/journal.pbio.1001177; pmid: 22028628
10. Companion Web site, tables W2 and W3; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TablesW
11. A. H. Orsi, T. Whitworth III, W. D. Nowlin Jr., On the meridional extent and fronts of the Antarctic Circumpolar Current. *Deep-Sea Res.* **42**, 641–673 (1995). doi: 10.1016/0967-0637(95)00021-W
12. Companion Web site, tables W4 to W12; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TablesW
13. Companion Web site, figure W1; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW1
14. Companion Web site, figure W2; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW2
15. Companion Web site, tables W13 and W14; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TablesW
16. Companion Web site, figure W3; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW3
17. C. de Vargas *et al*., Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
18. B. W. Bowen, L. A. Rocha, R. J. Toonen, S. A. KarlToBo Laboratory, The origins of tropical marine biodiversity. *Trends Ecol. Evol.* **28**, 359–366 (2013). doi: 10.1016/j.tree.2013.01.018; pmid: 23453048
19. R. L. Cunha *et al*., Ancient divergence in the trans-oceanic deep-sea shark *Centroscymnus crepidater*. *PLOS ONE* **7**, e49196 (2012). doi: 10.1371/journal.pone.0049196; pmid: 23145122
20. C. Schmid *et al*., Early evolution of an Agulhas Ring. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **50**, 141–166 (2003). doi: 10.1016/S0967-0645(02)00382-X

21. Companion Web site, figure W4; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW4
22. Companion Web site, figure W5; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW5
23. Companion Web site, figure W6; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW6
24. Companion Web site, figure W7; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW7
25. L. Gordon, J. R. Lutjeharms, M. L. Gründlingh, Stratification and circulation at the Agulhas Retroflection. *Deep-Sea Res. A, Oceanogr. Res. Pap.* **34**, 565–599 (1987). doi: 10.1016/0198-0149(87)90006-9
26. V. Faure, M. Arhan, S. Speich, S. Gladyshev, Heat budget of the surface mixed layer south of Africa. *Ocean Dyn.* **61**, 1441–1458 (2011). doi: 10.1007/s10236-011-0444-1
27. Companion Web site, figure W8; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW8
28. Companion Web site, figure W9; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW9
29. Companion Web site, text W1; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TextW1
30. Companion Web site, figure W10; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW10
31. Companion Web site, figure W11; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW11
32. Companion Web site, figure W12; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW12
33. S. Clayton, S. Dutkiewicz, O. Jahn, M. J. Follows, Dispersal, eddies, and the diversity of marine phytoplankton. *Limnol. Oceanogr. Fluids Environ.* **3**, 182–197 (2013). doi: 10.1215/21573689-2373515
34. Companion Web site, figure W13; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#FigW13
35. R. J. Olson, Differential photoinhibition of marine nitrifying bacteria: A possible mechanism for the formation of the primary nitrite maximum. *J. Mar. Res.* **39**, 227–238 (1981).
36. S. Levitus *et al*., The World Ocean Database. *Data Sci. J.* **12**, WDS229–WDS234 (2013). doi: 10.2481/dsj.WDS-041
37. Companion Web site, table W15; available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas/#TablesW
38. D. J. Scanlan *et al*., Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev.* **73**, 249–299 (2009). doi: 10.1128/MMBR.00035-08; pmid: 19487728
39. Z. I. Johnson *et al*., Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science* **311**, 1737–1740 (2006). doi: 10.1126/science.1118052; pmid: 16556835
40. K. Zwirglmaier *et al*., Global phylogeography of marine Synechococcus and Prochlorococcus reveals a distinct partitioning of lineages among oceanic biomes. *Environ. Microbiol.* **10**, 147–161 (2008). pmid: 17900271
41. A. C. Martiny, S. Kathuria, P. M. Berube, Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10787–10792 (2009). doi: 10.1073/pnas.0902532106; pmid: 19549842
42. C. Hubert *et al*., A constant flux of diverse thermophilic bacteria into the cold Arctic seabed. *Science* **325**, 1541–1544 (2009). doi: 10.1126/science.1174012; pmid: 19762643
43. C. K. C. Churchill, Á. Valdés, D. Ó. Foighil, Afro-Eurasia and the Americas present barriers to gene flow for the cosmopolitan neustonic nudibranch *Glaucus atlanticus*. *Mar. Biol.* **161**, 899–910 (2014). doi: 10.1007/s00227-014-2389-7
44. N. Selje, M. Simon, T. Brinkhoff, A newly discovered *Roseobacter* cluster in temperate and polar oceans. *Nature* **427**, 445–448 (2004). doi: 10.1038/nature02272; pmid: 14749832
45. G. Casteleyn *et al*., Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12952–12957 (2010). doi: 10.1073/pnas.1001380107; pmid: 20615950
46. F. L. Hellweger, E. van Sebille, N. D. Fredrick, Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. *Science* **345**, 1346–1349 (2014). doi: 10.1126/science.1254421; pmid: 25214628
47. D. H. Janzen, Why mountain passes are higher in the tropics. *Am. Nat.* **101**, 233–249 (1967). doi: 10.1086/282487
48. G. Wang, M. E. Dillon, Recent geographic convergence in diurnal and annual temperature cycling flattens global thermal profiles. *Nature Climate Change* **4**, 988–992 (2014). doi: 10.1038/nclimate2378
49. C. Backeberg, P. Penven, M. Rouault, Impact of intensified Indian Ocean winds on mesoscale variability in the Agulhas system. *Nature Clim. Change* **2**, 608–612 (2012). doi: 10.1038/nclimate1587
50. S. Pesant *et al*., Open science resources for the discovery and analysis of *Tara* Oceans data. http://biorxiv.org/content/early/2015/05/08/019117 (2015).
51. S. Sunagawa *et al*., Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
52. J. R. Brum *et al*., Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
53. J. M. Gasol, P. A. Del Giorgio, Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. *Sci. Mar.* **64**, 197–224 (2000).
54. M. Mackey, D. Mackey, H. Higgins, S. Wright, CHEMTAX - a program for estimating class abundances from chemical markers: Application to HPLC measurements of phytoplankton. *Mar. Ecol. Prog. Ser.* **144**, 265–283 (1996). doi: 10.3354/meps144265
55. G. Gorsky *et al*., Digital zooplankton image analysis using the ZooScan integrated system. *J. Plankton Res.* **32**, 285–303 (2010). doi: 10.1093/plankt/fbp124
56. P. Hingamp *et al*., Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013). doi: 10.1038/ismej.2013.59; pmid: 23575371
57. Tara Oceans Consortium Coordinators; Tara Oceans Expedition, Participants (2014): Registry of selected samples from the Tara Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840721
58. S. Chaffron, L. Guidi, F. D'Ovidio, S. Speich, S. Audic, S. De Monte, D. Iudicone, M. Picheral, S. Pesant; Tara Oceans Consortium Coordinators, Tara Oceans Expedition, Participants (2014): Contextual environmental data of selected samples from the Tara Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840718
59. S. Chaffron, F. D'Ovidio, S. Sunagawa, S. G. Acinas, L. P. Coelho, S. De Monte, G. Salazar, S. Pesant; Tara Oceans Consortium Coordinators, Tara Oceans Expedition, Participants (2014): Contextual biodiversity data of selected samples from the Tara Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840698
60. N. Maillet, C. Lemaitre, R. Chikhi, D. Lavenier, P. Peterlongo, Compareads: Comparing huge metagenomic experiments. *BMC Bioinformatics* **13** (suppl. 19), S10 (2012). doi: 10.1186/1471-2105-13-S19-S10; pmid: 23282463
61. M. Kanehisa *et al*., KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36** (Database), D480–D484 (2008). doi: 10.1093/nar/gkm882; pmid: 18077471
62. S. Mazard, M. Ostrowski, F. Partensky, D. J. Scanlan, Multi-locus sequence analysis, taxonomic resolution and biogeography of marine Synechococcus. *Environ. Microbiol.* **14**, 372–386 (2012). doi: 10.1111/j.1462-2920.2011.02514.x; pmid: 21651684
63. D. B. Rusch *et al*., The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLOS Biol.* **5**, e77 (2007). doi: 10.1371/journal.pbio.0050077; pmid: 17355176
64. J. Marshall, A. Adcroft, C. Hill, L. Perelman, C. Heisey, A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers. *J. Geophys. Res.* **102** (C3), 5753–5766 (1997). doi: 10.1029/96JC02775
65. M. J. Follows, S. Dutkiewicz, S. Grant, S. W. Chisholm, Emergent biogeography of microbial communities in a model ocean. *Science* **315**, 1843–1846 (2007). doi: 10.1126/science.1138544; pmid: 17395828
66. S. Dutkiewicz, M. J. Follows, J. G. Bragg, Modeling the coupling of ocean ecology and biogeochemistry. *Global Biogeochem. Cycles* **23**, GB4017 (2009). doi: 10.1029/2008GB003405
67. D. Menemenlis *et al*., ECCO2: High resolution global ocean and sea ice data synthesis. *Mercator Ocean Quarterly Newsletter* **31**, 13–21 (2008).
68. A. Chase *et al*., Decomposition of in situ particulate absorption spectra. *Methods in Oceanography.* **7**, 110–124 (2013). doi: 10.1016/j.mio.2014.02.002
69. E. Boss *et al*., The characteristics of particulate absorption, scattering and attenuation coefficients in the surface ocean; Contribution of the Tara Oceans expedition. *Methods in Oceanography.* **7**, 52–62 (2013). doi: 10.1016/j.mio.2013.11.002
70. K. R. Ridgway, J. R. Dunn, J. L. Wilkin, Ocean interpolation by four-dimensional least squares - Application to the waters around Australia. *J. Atmos. Ocean. Technol.* **19**, 1357–1375 (2002). doi: 10.1175/1520-0426(2002)019<1357:OIBFDW>2.0.CO;2

ACKNOWLEDGMENTS

We appreciate the commitment of the following people and sponsors: CNRS (in particular, Groupement de Recherche GDR3280); European Molecular Biology Laboratory (EMBL); Genoscope/CEA; VIB; Stazione Zoologica Anton Dohrn; UNIMIB; Fund for Scientific Research–Flanders; Rega Institute, KU Leuven; the French Ministry of Research; the French government Investissements d'Avenir programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), and ANR (projects POSEIDON/ANR-09-BLAN- 0348, PHYTBACK/ANR-2010-1709-01, PROMETHEUS/ANR-09-PCS-GENM-217, SAMOSA/ANR-13-ADAP-0010, and TARAGIRUS/ANR-09-PCS-GENM-218); European Union FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376, and MaCuMBA/No.311975); ERC Advanced Grant Award to C.B. (Diatomite: 294823); Gordon and Betty Moore Foundation grant (no. 3790) to M.B.S.; Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to S.G.A.; TANIT (CONES 2010-0036) from the Agència de Gestió d´Ajusts Universitaris i Reserca to S.G.A.; JSPS KAKENHI grant no. 26430184 to H.O.; NASA Ocean Biology and Biogeochemistry program (NNX11AQ14G and NNX09AU43G) to E.B.; The Italian Research for the Sea (Flagship Project RITMARE) to D.I.; and FWO, BIO5, and Biosphere 2 to M.B.S. We also appreciate the support and commitment of Agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, and the *Tara* schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries that graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge excellent assistance from the European Bioinformatics Institute (EBI), in particular G. Cochrane and P. ten Hoopen, as well as the EMBL Advanced Light Microscopy Facility (ALMF), in particular R. Pepperkok. We thank Y. Timsit for stimulating scientific discussions and critical help during writing of the manuscript. The altimeter products were produced by Ssalto/Duacs and CLS, with support from CNES. The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled. Data described herein are available at http://www.igs.cnrs-mrs.fr/Tara_Agulhas, at EBI under the project identifiers PRJEB402 and PRJEB7988, and at PANGAEA (*57*–*59*). The data release policy regarding future public release of *Tara* Oceans data is described in Pesant *et al*. (*50*). All authors approved the final manuscript. This article is contribution number 21 of *Tara* Oceans. The supplementary materials contain additional data. See also http://doi.pangaea.de/10.1594/PANGAEA.840721; http://doi.pangaea.de/10.1594/PANGAEA.840718; and http://doi.pangaea.de/10.1594/PANGAEA.840698

Tara Oceans Coordinators

Silvia G. Acinas,[1] Peer Bork,[2] Emmanuel Boss,[3] Chris Bowler,[4] Colomban de Vargas,[5,6] Michael Follows,[7] Gabriel Gorsky,[8,9] Nigel Grimsley,[10,11] Pascal Hingamp,[12] Daniele Iudicone,[13] Olivier Jaillon,[14,15,16] Stefanie Kandels-Lewis,[2,17] Lee Karp-Boss,[3] Eric Karsenti,[4,17] Uros Krzic,[18] Fabrice Not,[5,6] Hiroyuki Ogata,[19] Stephane Pesant,[20,21] Jeroen Raes,[22,23,24] Emmanuel G. Reynaud,[25] Christian Sardet,[26,27] Mike Sieracki,[28] Sabrina Speich,[29,30] Lars Stemmann,[8] Matthew B. Sullivan,[31] Shinichi Sunagawa,[2] Didier Velayoudon,[32] Jean Weissenbach,[14,15,16] Patrick Wincker[14,15,16]

[1]Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)-CSIC, Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [2]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [3]School of Marine Sciences, University of Maine, Orono, Maine, USA. [4]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, F-75005 Paris, France. [5]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [6]Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [9]Sorbonne Universités, UPMC Univ Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [10]CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [11]Sorbonne Universités

Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [12]Aix Marseille Université CNRS IGS UMR 7256, 13288 Marseille, France. [13]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [14]CEA, Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [15]CNRS, UMR 8030, CP5706, Evry, France. [16]Université d'Evry, UMR 8030, CP5706, Evry, France. [17]Directors' Research, European Molecular Biology Laboratory, Heidelberg, Germany. [18]Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [19]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. [20]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [21]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [22]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [23]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [24]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [25]Earth Institute, University College Dublin, Dublin, Ireland. [26]CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [27]Sorbonne Universités, UPMC Univ Paris 06, UMR 7009 Biodev, F-06230 Observatoire Océanologique, Villefranche-sur-Mer, France. [28]Bigelow Laboratory for Ocean Sciences, East Boothbay, USA. [29]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond 75231 Paris Cedex 05, France. [30]Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France. [31]Department of Ecology and Evolutionary Biology, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA. [32]DVIP Consulting, Sèvres, France.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/1261447/suppl/DC1
Table S1

18 September 2014; accepted 23 February 2015
10.1126/science.1261447

OCEAN PLANKTON

Patterns and ecological drivers of ocean viral communities

Jennifer R. Brum,[1*†] J. Cesar Ignacio-Espinoza,[2*†] Simon Roux,[1*†] Guilhem Doulcier,[1,3] Silvia G. Acinas,[4] Adriana Alberti,[5] Samuel Chaffron,[6,7,8] Corinne Cruaud,[5] Colomban de Vargas,[9,10] Josep M. Gasol,[4] Gabriel Gorsky,[11,12] Ann C. Gregory,[13†] Lionel Guidi,[11,12] Pascal Hingamp,[14] Daniele Iudicone,[15] Fabrice Not,[9,10] Hiroyuki Ogata,[16] Stéphane Pesant,[17,18] Bonnie T. Poulos,[1] Sarah M. Schwenck,[1] Sabrina Speich,[19‡] Celine Dimier,[9,10,20] Stefanie Kandels-Lewis,[21,22] Marc Picheral,[11,12] Sarah Searson,[11,12] *Tara* Oceans Coordinators,[§] Peer Bork,[21,23] Chris Bowler,[20] Shinichi Sunagawa,[21] Patrick Wincker,[5,24,25] Eric Karsenti,[20,22‖] Matthew B. Sullivan[1,2,13†‖]

Viruses influence ecosystems by modulating microbial population size, diversity, metabolic outputs, and gene flow. Here, we use quantitative double-stranded DNA (dsDNA) viral-fraction metagenomes (viromes) and whole viral community morphological data sets from 43 *Tara* Oceans expedition samples to assess viral community patterns and structure in the upper ocean. Protein cluster cataloging defined pelagic upper-ocean viral community pan and core gene sets and suggested that this sequence space is well-sampled. Analyses of viral protein clusters, populations, and morphology revealed biogeographic patterns whereby viral communities were passively transported on oceanic currents and locally structured by environmental conditions that affect host community structure. Together, these investigations establish a global ocean dsDNA viromic data set with analyses supporting the seed-bank hypothesis to explain how oceanic viral communities maintain high local diversity.

Ocean microbes produce half of the oxygen we breathe (*1*) and drive much of the substrate and redox transformations that fuel Earth's ecosystems (*2*). However, they do so in a constantly evolving network of chemical, physical, and biotic constraints—interactions that are only beginning to be explored. Marine viruses are presumably key players in these interactions (*3*, *4*), as they affect microbial populations through lysis, reprogramming of host metabolism, and horizontal gene transfer. Here, we strive to develop an overview of ocean viral community patterns and ecological drivers.

The *Tara* Oceans expedition provided a platform for sampling ocean biota from viruses to fish larvae within a comprehensive environmental context (*5*). Prior virus-focused work from this expedition has helped optimize the double-stranded DNA (dsDNA) viromic sample-to-sequence workflow (*6*), evaluate ecological drivers of viral community structure as inferred from morphology (*7*), and map ecological patterns in the large dsDNA nucleo-cytoplasmic viruses using marker genes (*8*). Here, we explore global patterns and structure of ocean viral communities using 43 samples from 26 stations in the *Tara* Oceans expedition (see supplementary file S1) to establish dsDNA viromes from viral-fraction (<0.22 μm) concentrates and quantitative whole viral community morphological data sets from unfiltered seawater. Viruses lack shared genes that can be used for investigation of community patterns. Therefore, we used three levels of information to study such patterns: (i) protein clusters (PCs) (*9*) as a means to organize virome sequence space commonly dominated by unknown sequences (63 to 93%) (*10*), (ii) populations, using established metrics for viral contig recruitment (*11*), and (iii) morphology, using quantitative transmission electron microscopy (qTEM) (*7*).

The *Tara* Oceans Viromes (TOV) data set

The 43 *Tara* Oceans Viromes (TOV) data set comprises 2.16 billion ~101-base pair (bp) paired-end Illumina reads (file S1), which largely represent epipelagic ocean viral communities from the surface (ENVO:00002042) and deep chlorophyll maximum (DCM; ENVO:01000326) throughout seven oceans and seas; only 1 of 43 viromes is from mesopelagic waters, Environment Ontology feature ENVO:00000213 (file S1). The TOV data set offers deeper sampling of surface ocean viral communities but underrepresents the deep ocean relative to the Pacific Ocean Viromes data set (POV) (*10*), which includes 16 viromes from aphotic zone waters. In all viromes, sampling and processing affects which viruses are represented (*6*, *12*–*14*). We filtered TOV seawater samples through 0.22-μm-pore–sized filters and then concentrated viruses in the filtrate using iron chloride flocculation (*15*). These steps would have removed most cells but also would have excluded any viruses larger than 0.22 μm. We then purified the resulting TOV viral concentrates using deoxyribonuclease (DNase) treatment, which is as effective as density gradients for purifying ocean viral concentrates (*14*). This DNAse-only step is unlikely to affect viral representation in the viromes but reduces nonviral DNA contamination. Finally, we extracted DNA from the samples and prepared sequence libraries using linker amplification (*13*). These steps preserve quantitative representation of dsDNA viruses in the resulting viromes (*12*, *13*), but the ligation step excludes RNA viruses and is biased against single-stranded DNA (ssDNA) viruses (*12*).

We additionally applied quantitative transmission electron microscopy (qTEM) (*7*) to paired whole seawater samples to evaluate patterns in whole viral communities. This method simultaneously considers ssDNA, dsDNA, and RNA viruses, although without knowledge of their relative abundances because particle morphology does not identify nucleic acid type. In the oceans, total virus abundance estimates based on TEM analyses, which include all viral particles, are similar to estimates based on fluorescent staining, which inefficiently stains ssDNA and RNA viruses (*16*–*24*). This suggests that most

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA. [2]Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721, USA. [3]Environmental and Evolutionary Genomics Section, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS, UMR8197, INSERM U1024, 75230 Paris, France. [4]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Pg. Marítim de la Barceloneta 37-49, Barcelona, E08003, Spain. [5]Genoscope, Commissariat à l'Energie Atomique (CEA)—Institut de Génomique, 2 rue Gaston Crémieux, 91057 Evry, France. [6]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [7]Center for the Biology of Disease, VIB KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [8]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [9]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [10]Sorbonne Universités, Université Pierre et Marie Curie, Université Paris 06, and UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [11]CNRS, UMR 7093, Laboratoire d'océanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-mer, France. [12]Sorbonne Universités, Uiversité Pierre et Marie Curie, Université Paris 06, UMR 7093, Laboratoire d'océanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-mer, France. [13]Soil, Water, and Environmental Science, University of Arizona, Tucson, AZ 85721, USA. [14]Aix Marseille Université, CNRS IGS UMR 7256, 13288 Marseille, France. [15]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [16]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0001, Japan. [17]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, 28359 Bremen, Germany. [18]MARUM, Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany. [19]Laboratoire de Physique des Océans, Institut Universitaire Européen de la Mer, Université de Bretagne Occidentale (UBO-IUEM), Place Copernic, 29820 Plouzané, France. [20]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), and INSERM U1024, and CNRS UMR 8197, Paris, 75005, France. [21]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [22]Directors' Research, European Molecular Biology Laboratory Meyerhofstrasse 1, 69117 Heidelberg, Germany. [23]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. [24]CNRS, UMR 8030, CP5706, 91057 Evry, France. [25]Université d'Evry, UMR 8030, CP5706, 91057 Evry, France. *These authors contributed equally to this work. †Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA. ‡Present address: Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond 75231 Paris, Cedex 05, France. §*Tara* Oceans coordinators and affiliations are listed after the Acknowledgments. ‖Corresponding authors. E-mail: mbsulli@gmail.com (M.B.S.); karsenti@embl.de (E.K.)

ocean viruses are dsDNA viruses. However, one study quantifying nucleic acids at a single marine location suggests that RNA viruses may constitute as much as half of the viral community there (*16*). It remains unknown what the relative contribution of these viral types is to the whole viral community, but our analyses suggest small dsDNA viruses likely dominate as follows. The viromes capture the <0.22-μm dsDNA viruses of bacteria and archaea that are thought to dominate marine viral communities, whereas qTEM analysis includes all viruses regardless of size, nucleic acid type, or host (*7*). In these whole seawater samples used for qTEM, we found that viral capsid diameters ranged from 26 to 129 nm, with the per-sample average capsid diameter constrained at 46 to 66 nm (Fig. 1). We detected no viral particles larger than 0.22 μm among 100 randomly counted particles from each of 41 qTEM samples. These findings are similar to those from a subset of these *Tara* Oceans stations (14 of the 26 stations) (*7*) and indicate that size fractionation using 0.22-μm filtration to prepare viromes did not substantially bias the TOV data set.

TOV protein clusters for comparison of local and global genetic richness and diversity

Across the 43 viromes, a total of 1,075,763 PCs were observed, with samples beyond the 20th virome adding few PCs (Fig. 2A). When we combined TOV with 16 photic-zone viromes from the POV data set (*10*), the number of PCs increased to 1,323,921 but again approached a plateau (Fig. 2B). These results suggest that, although it is impossible to sample completely, the sequence space corresponding to dsDNA viruses from the epipelagic ocean is now relatively well sampled. This contrasts results from marine microbial metagenomic surveys using older sequencing technologies (*9*) but is consistent with those from this expedition (*25*), as well as findings from viral sequence data sets that suggest a limited range of functional diversity derived from bacterial and archaeal viral isolates (*26*) and the POV data set (*27*).

PCs were next used to establish the core genes shared across the TOV data set (Fig. 2A). Broadly, there were 220, 710, and 424 core PCs shared across all surface and DCM viromes, surface viromes only, and DCM viromes only, respectively. The number of core PCs in the upper-ocean TOV samples (220 PCs) was thus less than the number of photic-zone core PCs in POV (565 PCs) (*28*), likely because the POV data set includes only the Pacific Ocean, whereas TOV includes samples from seven oceans and seas. However, the number of core PCs in the upper-ocean TOV samples exceeded the total number of core PCs observed in POV (180 PCs) (*28*), likely because of deep-ocean representation in POV (half of the samples in POV are from the aphotic zone). Consistent with the latter finding, the addition of the sole deep-ocean TOV sample, TARA_70_MESO, decreased the number of core PCs shared by all TOV samples from 220 to 65, which suggests that deep-ocean viral genetic repertoires are different from those in the upper oceans. Indeed, niche-differentiation has been observed in viromes sampled across these oceanic zones in the POV data set (*28*), and similar findings were observed in the microbial metagenomic counterparts from the *Tara* Oceans Expedition (*25*). Thus, viral communities from the deep ocean remain poorly explored and appear to hold different gene sets from those in the epipelagic oceans.

Beyond core and pan metagenomic analyses, PCs also provide a metric for viral community diversity comparisons (Fig. 3A and file S1) from which three trends emerge in the TOV data set. First, high-latitude viromes (82_DCM and 85_DCM) were least diverse [the entropy calculated with the natural log of diversity, Shannon's H', of 8.93 and 9.22 natural digits (nats)], consistent with patterns in marine macroorganisms (*29*) and epipelagic ocean bacteria (*25*, *30*). Second, the remaining viromes had similar diversity (Shannon's H' between 9.47 and 10.55 nats) and evenness (Pielou's J from 0.85 to 0.91), which indicated low dominance of any particular PCs (*31*). Third, local diversity was relatively similar to global diversity (local: global ratios of H' from 0.73 to 0.87), which suggested high dispersal of viral genes (*32*) across the sampled ocean viral communities.

TOV viral populations for assessing global viral community structure

We next estimated abundances of the 5476 dominant viral populations in TOV, which represented up to 9.97% of aligned reads in a sample and were defined by applying empirically derived recruitment cut-offs from naturally occurring T4-like cyanophages (*11*) to high-confidence contigs from bacterial and archaeal viruses (see Methods). Assigning viral populations on the basis of virome data remains challenging (*11*, *33*), but here, the assembly of large contigs (up to 100 kb) aided our ability to accomplish not only analyses at the gene-level using PCs but also the genome-level using viral populations. Viral populations were rarely endemic to one station (15%) and, instead, were commonly observed across >4 stations (47%) and up to 24 of the 26 stations (Figs. 4 and 5A). Exceptional samples include those from the Benguela upwelling region (TARA_67_SUR) and high-latitude samples from the Falklands and Antarctic Circumpolar currents (TARA_82_DCM and TARA_85_DCM, respectively). These samples were also divergent when we assessed microbial communities (TARA_82_DCM and TARA_85_DCM displayed lower microbial genetic richness) (*25*) and eukaryotic communities (TARA_67_SUR had specific and unique eukaryotic communities in all size fractions) (*34*). Although many viral populations were broadly distributed, they were much more abundant at the original location (origin inferred from longest contig assembled; see Methods) compared with alternate stations (Fig. 5B). Thus, most populations were relatively widespread but with variable sample-to-sample abundances. As was observed with PCs, diversity and evenness estimates based on viral populations were similar across all samples except for high-latitude samples (TARA_82_DCM and

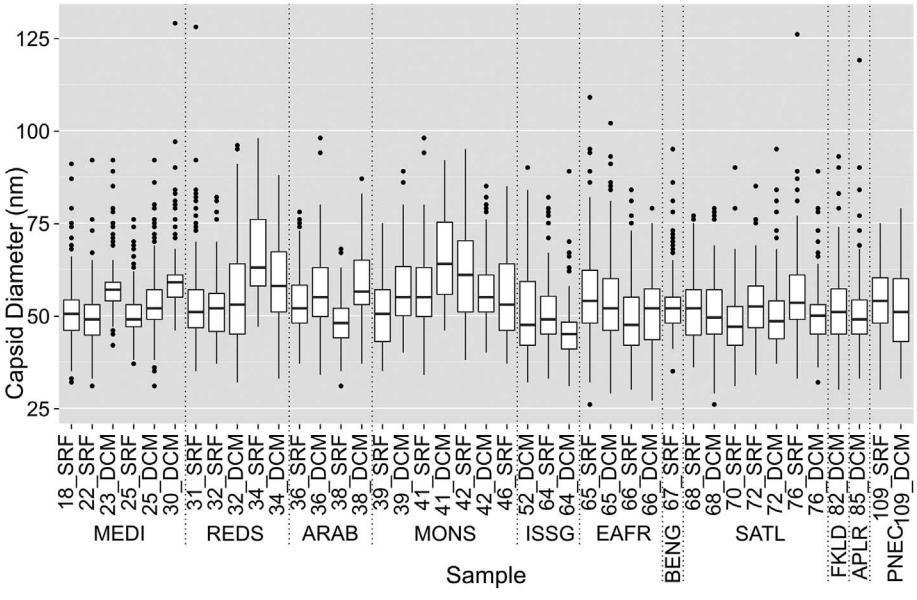


Fig. 1. Distribution of viral capsid diameters in each sample (n = 100 viruses per sample). Data are not available for samples TARA_18_DCM and TARA_70_MESO. Boxplots are constructed with the upper and lower lines corresponding to the 25th and 75th percentiles; outliers are displayed as points. Longhurst provinces are indicated below samples (MEDI, Mediterranean Sea; REDS, Red Sea; ARAB, NW Arabian Upwelling; MONS, Indian Monsoon Gyres; ISSG, Indian S. Subtropical Gyre; EAFR, E. Africa Coastal; BENG, Benguela Current Coastal; SATL, S. Atlantic Gyre; FKLD, SW Atlantic Shelves; APLR, Austral Polar; PNEC, N. Pacific Equatorial Countercurrent).

TARA_85_DCM) and one sample in the Red Sea (TARA_32_DCM) that displayed lower diversity (Fig. 3B and file S1). Finally, local diversity was relatively similar to global diversity (local:global ratios of H' from 0.23 to 0.86, average 0.74) (file S1) and reflected the high dispersal of viruses as highlighted by PC analysis.

Only 39 of the 5476 populations we identified could be affiliated to cultured viruses, which reflects the dearth of reference viral genomes in databases. These cultured viruses include those infecting the abundant and widespread hosts SAR11, SAR116, *Roseobacter*, *Prochlorococcus*, and *Synechococcus* (Fig. 6). The most abundant and widespread viral populations observed in TOV lack cultured representatives (Fig. 6), which suggests that most upper-ocean viruses remain to be characterized even though viruses from known dominant microbial hosts (*35*–*39*) have been cultured. Methods independent of cultivation—including viral tagging (*11*) and mining of microbial genomic data sets (*40*, *41*)—show promise to expand the number of available viral reference genomes (*33*).

Drivers of global viral community composition and distribution

We next leveraged this global data set to evaluate ecological drivers (including environmental variables, sample location, and microbial abundances) (file S1) of viral community structure using all three data types—morphology, populations, and PCs. These metrics revealed increasing resolution, respectively, and showed that viral community structure was influenced by region and/or environmental conditions (Table 1). We conducted the analysis of ecological drivers using all samples in this study, as well as a sample subset that omitted samples with exceptional environmental conditions and divergent viral communities observed using PC and population analyses (see above; TARA_67_SUR, TARA_82_ DCM, TARA_85_DCM, and TARA_70_MESO). Within the sample subset, oceanic viral communities varied significantly with Longhurst province, biome, latitude, temperature, oxygen concentration, and microbial concentrations (including total bacteria, *Synechococcus*, and *Prochlorococcus*). Viral communities were not structured by depth (surface versus DCM) except when considering PCs, which likely reflects the minimal variation between samples in the epipelagic zone compared with that of globally sourced samples, as well as the higher resolution provided by PCs. Nutrients influenced viral community structure when we considered the whole data set but were much less explanatory when the few high-nutrient samples were removed, except for the influence of phosphate concentration on viral populations. Thus, nutrient concentrations may influence viral community structure, but testing this hypothesis would require analysis of samples across a more continuous nutrient gradient.

Global-scale analyses of oceanic macro- (*29*) and microorganisms (*30*) have been conducted, including a concurrent *Tara* Oceans study showing that temperature and oxygen influence microbial community structure (*25*). Environmental conditions have also been shown to affect global viral community morphological traits (*7*). Our TOV study is consistent with these earlier findings in that viral communities are influenced by temperature and oxygen concentration, but not chlorophyll concentration (Table 1). Biogeographic structuring of TOV viral communities on the basis of the significant influence of latitude and Longhurst provinces is also consistent with the conclusion that geographic region influences community structure in Pacific Ocean viruses (*42*). Although only PC analysis showed depth-based divergence, this likely reflects poor (n = 1) deep sample representation in the TOV data set as discussed above. Prior POV viral investigation and concurrent *Tara* Oceans microbial analysis, both of which have better deep-water representation, show stronger depth patterns whereby photic and aphotic zone communities diverge (*25*, *28*, *42*). Thus, our results suggest that the biogeography of upper-ocean viral communities is structured by environmental conditions.

Because viruses require host organisms to replicate, viral community structure follows from environmental conditions shaping the host community, as observed in paired *Tara* Oceans microbial samples (*25*), which would then indirectly affect viral community composition. However, global distribution of viruses can also be directly influenced by environmental conditions, such as salinity, that affect their ability to infect their hosts (*43*). Additionally, the variable decay rates observed for cultivated viruses and whole viral communities (*44*) could also influence their distribution as viruses with lower inherent decay rates will persist for longer in the environment, and environments with more favorable conditions (such as fewer extracellular enzymes) will also contribute to increased viral persistence.

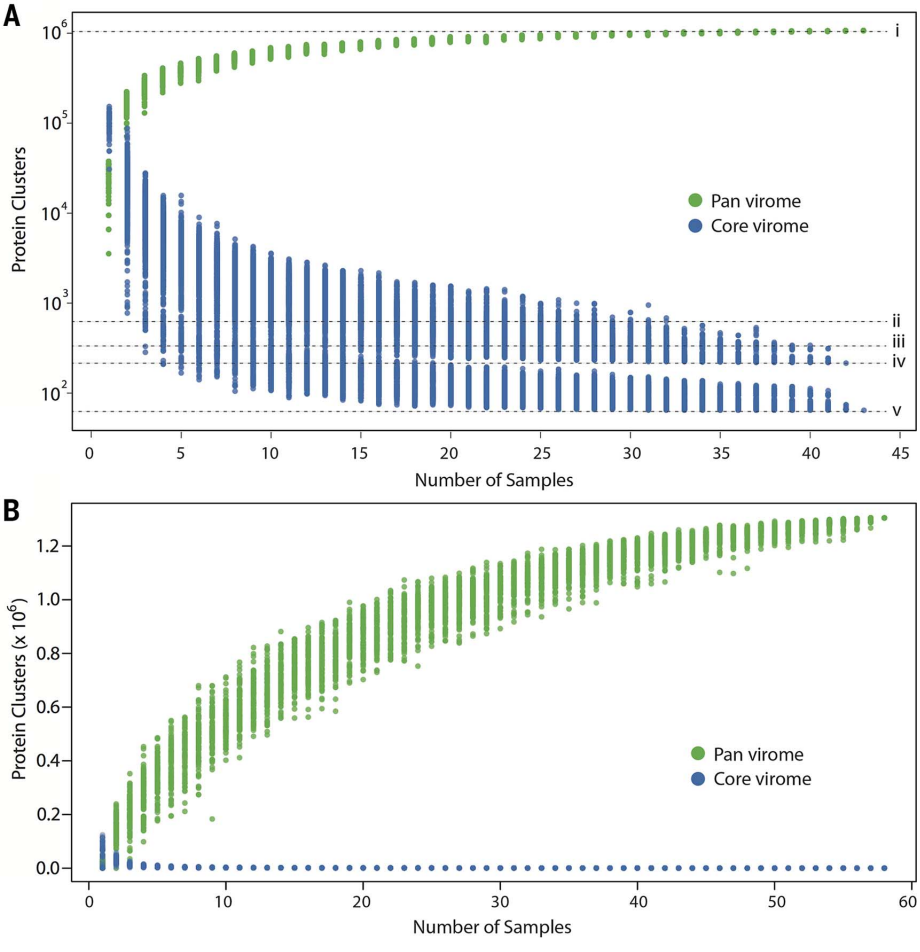


Fig. 2. PC richness in core and pan viromes from the TOV and POV data sets. (**A**) Accumulation curves of core and pan PCs in the TOV data set. Vertical axis shows the number of shared (core virome) and total (pan virome) PCs when n viromes are compared (n = 1 to 43; from 3 to 41 only 1000 combinations are shown). Lines: (i) total number of PCs (1,075,763 PCs), (ii) core surface virome (710 PCs), (iii) core DCM virome (424 PCs), (iv) core surface and DCM virome (220 PCs), (v) all samples (including the deep-ocean sample TARA_70_MESO; 65 PCs). (**B**) Core and pan PCs in all TOV and photic-zone POV samples combined. Vertical axis shows the number of shared (core virome) and total (pan-virome) PCs when n viromes are compared (n = 1 to 58; from 3 to 58 only 1000 combinations are shown). Overall, 1,323,921 PCs were identified in all viromes combined.

Until methods to link viruses to their host cells in natural communities mature to the point of investigating this issue at larger scales [emerging possible methods reviewed by (*33*, *45*)], analyses such as ours remain the only means to assess ecological drivers of viral community structure.

To further investigate how ocean viral communities are distributed throughout the oceans, we compared population abundances between neighboring samples to assess the net direction and magnitude of population exchange (Fig. 7 and see Methods). These genomic signals revealed that population exchange between dsDNA viral communities was largely directed along major oceanic current systems (*46*). For example, the Agulhas current and subsequent ring formation (*47*) connects viral communities between the Indian and Atlantic Oceans, as also observed in planktonic communities from the *Tara* Oceans expedition (*48*), whereas increased connection between the high-latitude stations (TARA_82 and TARA_85) reflects their common origin at the divergence of the Falklands and Antarctic Circumpolar currents. Further, current strength (*46*) was generally related to the magnitude of intersample population exchange, as higher and lower exchange was observed, respectively, in stronger currents, such as the Agulhas current, and within the open ocean gyres or between land-restricted basins such as the Mediterranean and Red Seas. These findings suggest that the intensity of water mass movement, in addition to environmental conditions, may explain the degree to which viral populations cluster globally (Fig. 4). Beyond such current-driven biogeographic evidence, vertical viral transport from surface to DCM samples was also observed (Fig. 4). This is consistent with POV observations wherein deep-sea viromes include a modest influx of genetic material derived from surface-ocean viruses that are presumably transported on sinking particles (*28*). Exceptions include areas such as the Arabian Sea upwelling region, where increased mixing and upwelling likely exceed sinking within the upper ocean.

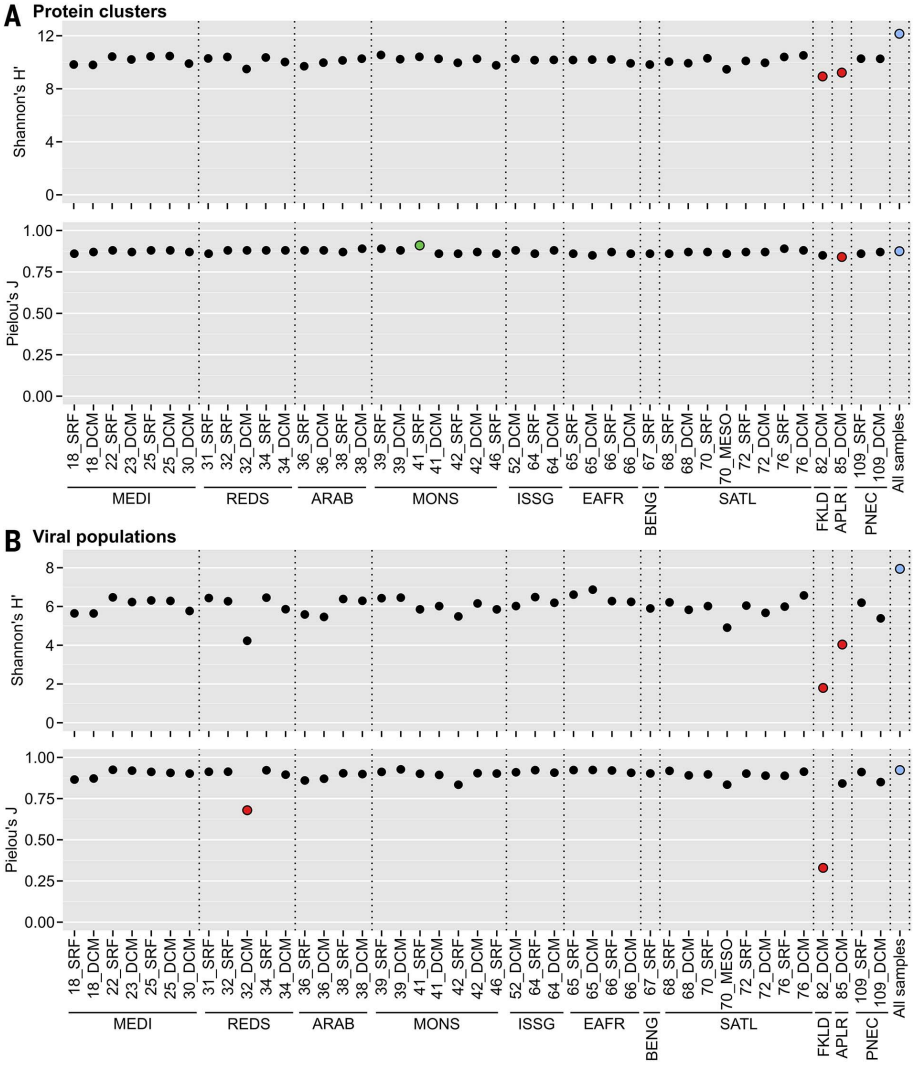


Fig. 3. Alpha diversity measurements in TOV data set. (**A**) Shannon's diversity *H*′ and Pielou's evenness *J* calculated from protein cluster counts for each sample and a pool of all samples, normalized to 5 million reads. (**B**) Shannon's diversity *H*′ and Pielou's evenness *J* calculated from relative abundances of viral populations for each sample and a pool of all samples, with subsamples of 100,000 reads. Outliers corresponding to values outside of the average value ±2 SD are colored green and red, respectively. Values calculated from the pool of all samples are colored blue. Longhurst provinces are indicated below samples using the same abbreviations as in Fig. 1.

Our TOV results enabled evaluation of a hypothesis describing the structure of viral communities in the environment. Gene marker–based studies targeting subsets of ocean viruses previously found high local and low global diversity (*49*), a pattern also recently observed genome-wide in natural cyanophage populations (*11*). To explain this, a seed-bank viral community structure has been invoked, whereby high local genetic diversity can exist by drawing variation from a common and relatively limited global gene pool (*49*). Our results support this hypothesis regarding viral community structure. Ecological driver analyses suggests that the numerically dominant members in local communities are influenced by environmental conditions, which directly impact their microbial hosts and then indirectly restructure viral communities. These dominant communities then form the "bank" in neighboring samples, presumably when passively transported by ocean currents as shown here through the population-level analyses of net viral movement between samples. This systematically sampled global data set suggests that large- and small-scale processes play roles in structuring viral communities and offers empirical grounding for the seed-bank hypothesis with regard to viral community distribution and structure.

Conclusions

Our large-scale data set provides a picture of global upper-ocean viral communities in which we assessed patterns using multiple parameters, including morphology, populations, and PCs. Our data provide advanced and complementary views on viral community structure including diversity estimates not based on marker genes and broad application of population-based viral ecology. We affirm the seed-bank model for viruses, hypothesized nearly a decade ago (*49*), which explains how high local viral diversity can be consistent with limited global diversity (*11*, *27*). The mechanism underlying this seed-bank population structure appears to be a local production of viruses under small-scale environmental constraints and passive dispersal with oceanic currents. Improving sequencing, assembly, and experimental methods are transforming the investigation of viruses in nature (*33*, *45*) and pave the way toward assessment of viral community structure and analysis of virus-host co-occurrence networks (*50*) without requiring marker genes (*51*, *52*). Such experimental and analytical progress, coupled to sampling opportunities from the *Tara* Oceans expedition, are advancing viral ecology toward the quantitative science needed to model the nanoscale

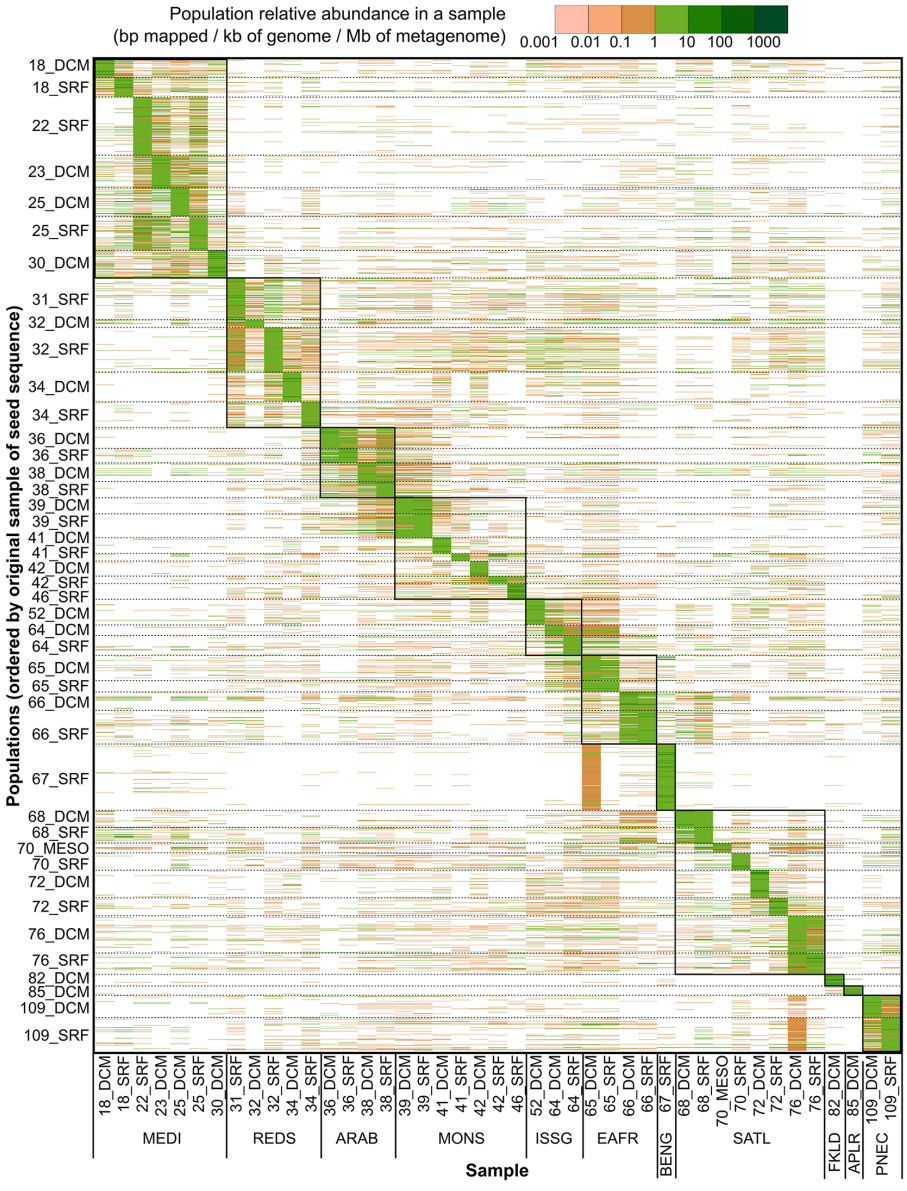


Fig. 4. Relative abundance of viral populations in TOV by sample. This heat map displays the relative abundance of each population (sorted according to its original sample, *y* axis) in each sample (*x* axis). Relative abundance of one population in a sample is based on recruitment of reads to the population reference contig and is only considered if more than 75% of the reference contig is covered. Longhurst provinces are indicated below samples (using the same abbreviations as in Fig. 1) and are outlined in black on the heat map.

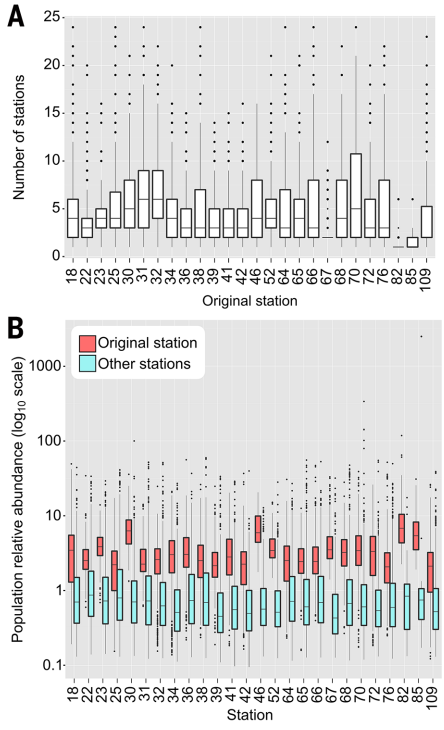


Fig. 5. Relative abundance of viral populations in TOV by station. (**A**) Evaluation of viral population distribution showing the number of stations (*y* axis) in which each population (sorted by their original station, *x* axis) is distributed. Populations are grouped by station, merging surface, and DCM samples from the same station. (**B**) Relative abundance of populations (bp mapped per Kb of contig per Mb of metagenome) at the original stations where the contigs were assembled compared with their abundance at other stations. Box plots are constructed as in Fig. 1.

(viruses) and microscale (microbes) entities driving Earth's ecosystems.

Materials and methods

Sample collection

Forty-three samples were collected between 2 November 2009 and 13 May 2011, at 26 locations throughout the world's oceans (file S1) through the *Tara* Oceans Expedition (*5*). These included samples from a range of depths (surface, deep chlorophyll maximum, and one mesopelagic sample) located in seven oceans and seas, four different biomes, and 11 Longhurst oceanographic provinces (file S1). Longhurst provinces and biomes are defined based on Longhurst (*53*) and environmental features are defined based on Environment Ontology (http://environmentontology.org/). Sampling strategy and methodology for the *Tara* Oceans Expedition is fully described by Pesant *et al.* (*54*).

Environmental parameters

Temperature, salinity, and oxygen data were collected from each station by measuring conductivity, temperature, depth, and pressure using a CTD (Sea-Bird Electronics, Bellevue, WA, USA; SBE 911plus with Searam recorder) and with a dissolved oxygen sensor (Sea-Bird Electronics; SBE 43). Nutrient concentrations were determined using segmented flow analysis (*55*) and included nitrite, phosphate, nitrite plus nitrate, and silica. Nutrient concentrations below the detection limit (0.02 μmol kg^{-1}) are reported as 0.02 μmol kg^{-1}. Chlorophyll concentrations were measured using high-performance liquid chromatography (*56*, *57*). These environmental parameters are available in PANGAEA (www.pangaea.de) by using the accession numbers in file S1.

Microbial abundances

Flow cytometry was used to determine the concentration of *Synechococcus*, *Prochlorococcus*, total bacteria, low-DNA bacteria, high-DNA bacteria, and the percentage of bacteria with high DNA in each sample (*58*).

Morphological analysis of viral communities

qTEM was used to evaluate the capsid diameter distributions of viral communities as previously described (*7*). Briefly, preserved unfiltered samples (electron microscopy–grade glutaraldehyde; Sigma-Aldrich, St. Louis, MO, USA; 2% final concentration) were flash-frozen and stored at −80°C until analysis. Viruses were deposited onto TEM grids using an air-driven ultracentrifuge (Airfuge CLS, Beckman Coulter, Brea, CA, USA), followed

by positive staining of the deposited material with 2% uranyl acetate (Ted Pella, Redding, CA, USA). Samples were then examined by using a transmission electron microscope (Philips CM12 FEI, Hilsboro, OR, USA) with 100 kV accelerating voltage. Micrographs of 100 viruses were collected per sample using a Macrofire Monochrome charge-coupled device camera (Optronics, Goleta, CA, USA) and analyzed using ImageJ software (U.S. National Institutes of Health, Bethesda, MD, USA) (*59*) to measure the capsid diameter. A subset (21) of the 41 samples presented here had previously been analyzed in a different study (*7*).

Virome construction

For each sample, 20 L of seawater were 0.22-μm–filtered, and viruses were concentrated from the filtrate using iron chloride flocculation (*15*) followed by storage at 4°C. After resuspension in ascorbic-EDTA buffer (0.1 M EDTA, 0.2 M Mg, 0.2 M ascorbic acid, pH 6.0), viral particles were concentrated using Amicon Ultra 100-kD centrifugal devices (Millipore), treated with DNase I (100 U/mL) followed by the addition of 0.1 M EDTA and 0.1 M EGTA to halt enzyme activity, and extracted as previously described (*14*). Briefly, viral particle suspensions were treated with Wizard Polymerase Chain Reaction Preps DNA Purification Resin (Promega, Fitchburg, WI, USA) at a ratio of 0.5-ml sample to 1-ml resin, and eluted with TE buffer (10 mM Tris, pH 7.5, 1 mM EDTA) using Wizard Minicolumns. Extracted DNA was Covaris-sheared and size-selected to 160 to 180 bp, followed by amplification and ligation per the standard Illumina protocol. Sequencing was done on a HiSEq 2000 system at the Genoscope facilities (Paris, France).

Quality control of reads and assembly

Individual reads of 43 metagenomes were controlled for quality by using a combination of trimming and filtering as previously described (*60*). Briefly, bases were trimmed at the 5′ end if the number of base calls for any base (A, T, G, C) diverged by more than 2 SD from the average across all cycles. Conversely, bases were trimmed at the 3′ end of reads if the quality score was <20. Finally, reads that were shorter than 95 bp or reads with a median quality score <20 were removed from further analyses. Assembly of reads was done using SOAPdenovo (*61*), where insert and *k*-mer size are calculated at runtime and are specific to each virome as implemented in the MOCAT pipeline (*62*). On average, 34.2% of the virome reads were included in the assembled contigs (min: 21.08%, max: 48.52%). Virome reads were deposited in the European Nucleotide Archive (www.ebi.ac.uk/ena/) under accession numbers reported in file S1.

Protein clustering

Open reading frames (ORFs) were predicted from all quality-controlled contigs using Prodigal (*63*) with default settings. Predicted ORFs were clustered on the basis of sequence similarity as described previously (*9*, *10*). Briefly, ORFs were initially mapped to existing clusters [POV, Global Ocean Sampling expedition, and phage genomes), by using cd-hit-2d ("-g 1 -n 4 -d 0 -T 24 -M 45000"; 60% identity and 80% coverage). Then, the remaining, unmapped ORFs were self-clustered, using cd-hit with the same options as above. Only PCs with more than two ORFs were considered bona fide and were used for subsequent analyses. To develop read counts per PC for statistical analyses, reads were mapped back to predicted ORFs in the contigs data set using Mosaik with the following settings: "-a all -m all -hs 15 –minp 0.95 –mmp 0.05 -mhp 100 -act 20" (version 1.1.0021; http://bioinformatics.bc.edu/mathlab/Mosaik). Read counts to PCs were normalized by sequencing depth of each virome. Shannon's diversity index (H') was calculated from PC read counts by using only PCs with more than two predicted ORFs. Observed richness is reported as the total number of reads in each PC. Pielou's evenness (J) was calculated as the ratio of H'/H_{max}, where $H_{max} = \ln N$, and N = total number of observed PCs in a sample.

Analysis of viral populations

Considering the size of the entire data set (3,821,756 assembled contigs), we decided to focus the analysis of viral populations using contigs originating from bacterial or archaeal viruses. For this, we mined only the 22,912 contigs with more than 10 predicted genes (corresponding to an average of 6.41% of the assembled reads per sample, min: 1.29%, max: 14.52%), as the origin of contigs with only a few predicted genes can be spurious. First, we removed 6706 contigs suspected of having originated from cellular genomes (*64*), whether due to free genomic DNA contamination or viral-encapsidation of cellular DNA (for example, in gene transfer agents or generalized transducing phages). These suspect cellular contigs were those containing no typical viral genes (such as virion-related genes including major capsid proteins and large subunits of the terminase) and displaying as many genes with a significant similarity to a PFAM domain through Hmmsearch (*65*) as a typical cellular genome, whereas phage genomes are typically enriched in uncharacterized genes (*40*). We also removed all contigs posited to originate from eukaryotic viruses. These were contigs that contained at least three predicted proteins with best BLAST hits to a eukaryotic virus, and more than half of the affiliated proteins were not associated with bacteriophages or archaeal viruses. Not surprisingly, given that eukaryotes are outnumbered by bacteria and archaea in the marine environment, this step removed only 142 contigs associated with eukaryotic viruses. From the remaining 16,124 contigs most likely to have originated from bacterial or archaeal viruses, the population study only used those longer than 10 kb in size—a total of 6322 contigs, which corresponded to an average of 4.04% of the assembled reads per sample (min: 0.98%, max: 9.97%).

These 6322 contigs were then clustered into populations if they shared more than 80% of their genes at >95% nucleotide identity; a threshold derived from naturally occurring T4-like cyanophages (*11*). This resulted in 5476 populations from the 6322 contigs, where as many

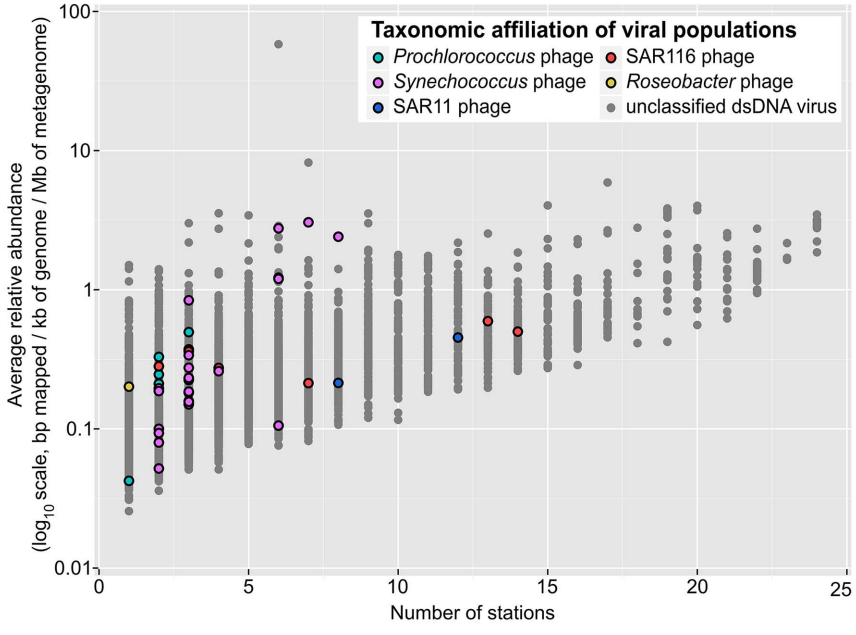


Fig. 6. Taxonomic affiliation of TOV viral populations sorted by distribution and average abundance. A population was considered as similar to a known virus when fewer than half of its reference contig genes were uncharacterized, and all characterized genes had taxonomic affiliations to the same reference genome. As in Fig. 4, the relative abundance (*y* axis) is computed for each sample as the number of bp mapped to a contig per kb of contig per Mb of metagenome sequenced. Here, the relative abundance of a population is defined as the average abundance of its reference contig across all samples.

Table 1. Relations between viral community structure and metadata. Relations between viral community structure (based on viral morphology, populations, and PCs) and metadata by using NMDS analysis of all samples and the sample subset (all samples except for TARA_67_SRF, TARA_70_MESO, TARA_82_DCM, and TARA_85_DCM because of exceptional environmental conditions at these locations). Significant relations are bold.

Category	*N* and *n*	Viral morphology (qTEM)	Populations (contigs)	Protein clusters (PCs)
Depth category	All samples	$P = 0.354$ ($N = 41$)	$P = 0.362$ ($N = 43$)	**$P = 0.033$ ($N = 43$)**
	Sample subset	$P = 0.228$ ($n = 38$)	$P = 0.105$ ($n = 39$)	**$P = 0.011$ ($n = 39$)**
Province	All samples	$P = 0.098$ ($N = 41$)	**$P < 0.001$ ($N = 43$)**	**$P = 0.014$ ($N = 43$)**
	Sample subset	**$P = 0.029$ ($n = 38$)**	**$P < 0.001$ ($n = 39$)**	**$P = 0.008$ ($n = 39$)**
Biome	All samples	$P = 0.099$ ($N = 41$)	**$P < 0.001$ ($N = 43$)**	$P = 0.097$ ($N = 43$)
	Sample subset	$P = 0.120$ ($n = 38$)	**$P < 0.001$ ($n = 39$)**	$P = 0.543$ ($n = 39$)
Latitude	All samples	**$P = 0.003$ ($N = 41$)**	**$P < 0.001$ ($N = 43$)**	**$P = 0.002$ ($N = 43$)**
	Sample subset	**$P = 0.014$ ($n = 38$)**	**$P < 0.001$ ($n = 39$)**	**$P = 0.010$ ($n = 39$)**
Temperature	All samples	**$P = 0.001$ ($N = 41$)**	**$P < 0.001$ ($N = 43$)**	**$P < 0.001$ ($N = 43$)**
	Sample subset	**$P = 0.001$ ($n = 38$)**	**$P < 0.001$ ($n = 39$)**	**$P = 0.015$ ($n = 39$)**
Salinity	All samples	$P = 0.118$ ($N = 39$)	**$P = 0.035$ ($N = 41$)**	**$P = 0.029$ ($N = 41$)**
	Sample subset	$P = 0.138$ ($n = 36$)	$P = 0.075$ ($n = 37$)	**$P = 0.001$ ($n = 37$)**
Oxygen	All samples	**$P = 0.001$ ($N = 41$)**	**$P < 0.001$ ($N = 43$)**	**$P < 0.001$ ($N = 43$)**
	Sample subset	**$P = 0.005$ ($n = 38$)**	**$P < 0.001$ ($n = 39$)**	**$P < 0.001$ ($n = 39$)**
Chlorophyll	All samples	$P = 0.711$ ($N = 41$)	**$P < 0.001$ ($N = 43$)**	**$P = 0.001$ ($N = 39$)**
	Sample subset	$P = 0.738$ ($n = 38$)	$P = 0.412$ ($n = 39$)	$P = 0.059$ ($n = 39$)
Nitrite	All samples	$P = 0.951$ ($N = 39$)	$P = 0.648$ ($N = 41$)	$P = 0.828$ ($N = 41$)
	Sample subset	$P = 0.851$ ($n = 36$)	$P = 0.509$ ($n = 37$)	$P = 0.999$ ($n = 37$)
Phosphate	All samples	$P = 0.275$ ($N = 39$)	**$P < 0.001$ ($N = 41$)**	**$P < 0.001$ ($N = 41$)**
	Sample subset	$P = 0.411$ ($n = 36$)	**$P < 0.001$ ($n = 37$)**	$P = 0.583$ ($n = 37$)
Nitrite + Nitrate	All samples	**$P = 0.046$ ($N = 39$)**	**$P < 0.001$ ($N = 41$)**	**$P < 0.001$ ($N = 41$)**
	Sample subset	$P = 0.290$ ($n = 36$)	$P = 0.052$ ($n = 37$)	$P = 0.643$ ($n = 37$)
Silica	All samples	**$P = 0.008$ ($N = 39$)**	**$P = 0.002$ ($N = 41$)**	**$P = 0.008$ ($N = 41$)**
	Sample subset	$P = 0.255$ ($n = 36$)	$P = 0.285$ ($n = 37$)	$P = 0.191$ ($n = 37$)
Bacteria	All samples	$P = 0.579$ ($N = 39$)	**$P < 0.001$ ($N = 40$)**	$P = 0.119$ ($N = 40$)
	Sample subset	$P = 0.329$ ($n = 36$)	**$P = 0.003$ ($n = 36$)**	**$P = 0.007$ ($n = 36$)**
Low DNA bacteria	All samples	$P = 0.227$ ($N = 39$)	$P = 0.090$ ($N = 40$)	$P = 0.123$ ($N = 40$)
	Sample subset	$P = 0.468$ ($n = 36$)	**$P = 0.018$ ($n = 36$)**	**$P = 0.005$ ($n = 36$)**
High DNA bacteria	All samples	$P = 0.967$ ($N = 39$)	**$P < 0.001$ ($N = 40$)**	$P = 0.273$ ($N = 40$)
	Sample subset	$P = 0.174$ ($n = 36$)	**$P = 0.027$ ($n = 36$)**	**$P = 0.024$ ($n = 36$)**
Percentage of high-DNA bacteria	All samples	**$P = 0.007$ ($N = 39$)**	$P = 0.078$ ($N = 40$)	**$P = 0.009$ ($N = 40$)**
	Sample subset	**$P = 0.017$ ($n = 36$)**	$P = 0.059$ ($n = 36$)	**$P < 0.001$ ($n = 36$)**
Synechococcus	All samples	$P = 0.143$ ($N = 39$)	$P = 0.094$ ($N = 40$)	**$P = 0.041$ ($N = 40$)**
	Sample subset	$P = 0.142$ ($n = 36$)	**$P = 0.023$ ($n = 36$)**	**$P = 0.013$ ($n = 36$)**
Prochlorococcus	All samples	$P = 0.118$ ($N = 39$)	$P = 0.076$ ($N = 40$)	$P = 0.123$ ($N = 40$)
	Sample subset	$P = 0.249$ ($n = 37$)	$P = 0.161$ ($n = 37$)	$P = 0.140$ ($n = 37$)

as 12 contigs (average 1.15 contigs) were included per population. For each population, the longest contig was chosen as the seed sequence.

The relative abundance of each population was computed by mapping all quality-controlled reads to the set of 5476 nonredundant populations (considering only mapping quality scores greater than 1) with Bowtie 2 (*66*). For each sample-sequence pair, if more than 75% of the reference sequence was covered by virome reads, the relative abundance was computed as the number of base pairs recruited to the contig normalized to the total number of base pairs available in the virome and the contig length. Shannon diversity index (H') and Pielou's evenness (J) were calculated as done for PCs using the relative abundance of viral populations.

The sample containing the seed sequence (the longest contig in a population) was also considered the best estimate of that population's origin. We reasoned that this was because the longest contig in a population would derive most often from the sample with the highest coverage (a proxy for population abundance) and likely corresponded to the location with the greatest viral abundance for this population. This assumption was supported by the results showing that populations were most abundant in their original samples (Figs. 4 and 5B). Even though some individual cases could diverge from this rule, we expected to correctly identify most of these original locations and, hence, to get an accurate global signal.

The seed sequence was also used to assess taxonomic affiliation of the viral population. Cases where >50% of the genes were affiliated to a specific reference genome from RefSeq (based on a BLASTp comparison with thresholds of 50 for bit score and 10^{-5} for e-value) with an identity percentage of at least 75% (at the protein sequence level) were considered confident affiliations with the corresponding reference virus.

Finally, estimations of net viral population movement between samples were made on the basis of the relative abundance of populations in one sample compared with that of its neighboring samples (Fig. 4). For each neighboring sample pair, the average relative abundance of populations originating from sample A in sample B was compared with the relative abundance of populations originating from sample B in sample A. The origin of each population was defined as the sample in which the longest contig of the population was assembled. The magnitude of these differences was carried through the analysis to estimate the level of transport between each pair of samples (depicted as line width in Fig. 7) and the difference between these values was used to estimate the directionality of the transfer. For example, if sample B contains many populations from sample A, but very few populations from sample B are detected in sample A, we calculate that the net movement

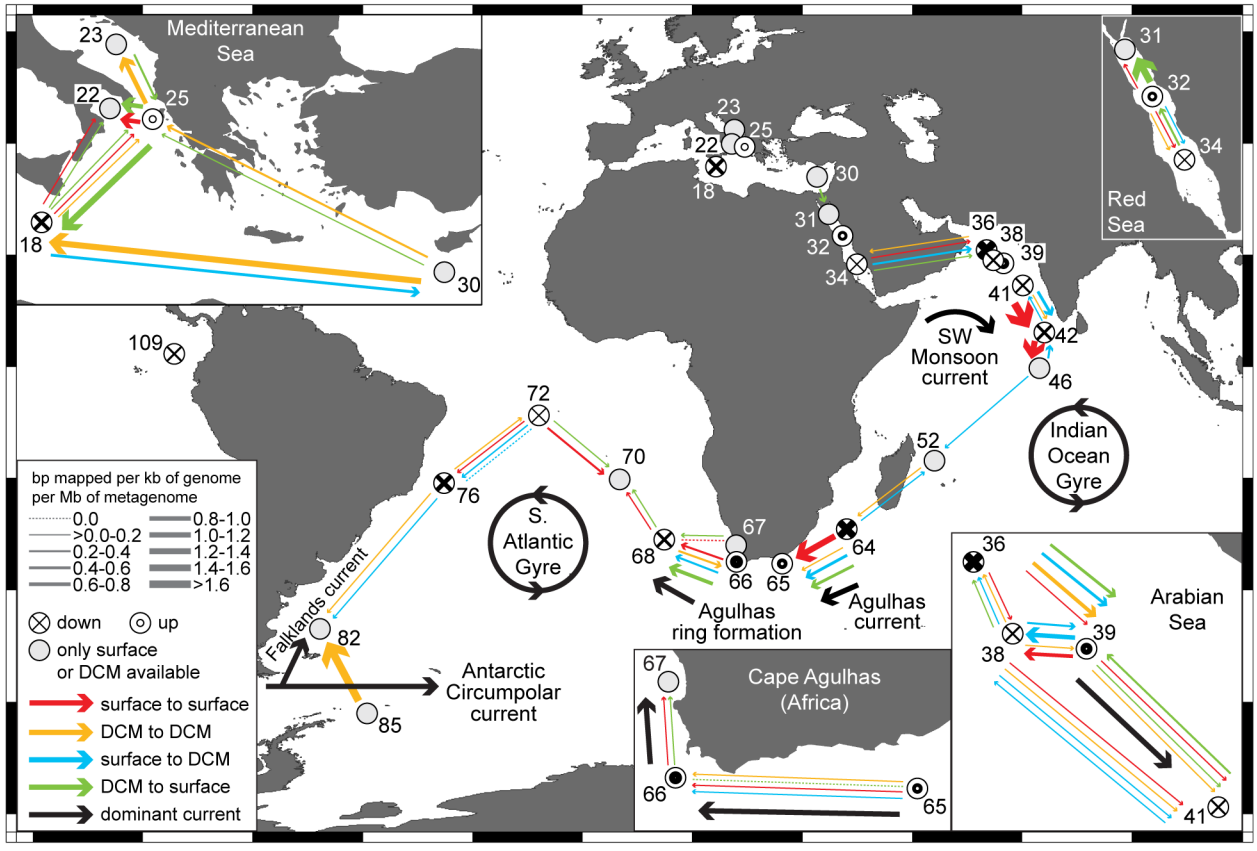


Fig. 7. Net movement of viral populations throughout the oceans. Calculations are based on reciprocal comparison of viral population abundances between neighboring samples (see Fig. 3 and Methods). For each sample pair, the average relative population abundances in one sample originating from a neighboring sample were calculated and compared (for example, relative abundance of populations from sample A found in sample B are compared with relative abundance of populations from sample B found in sample A). The sign of the relative abundance difference between neighboring samples was used to estimate the movement direction (arrowhead) and the absolute value of the difference was interpreted as reflecting the movement magnitude (line width). Stations are labeled with station number. "Down" and "up" refer to net vertical movement of viral populations between the surface and DCM samples at the same station.

is from sample A to sample B. Again, although the sampling of some populations may not be strong, the net movement was calculated as the average of all shared populations between neighboring sample pairs, which corresponded to 105 different populations on average (ranging from 2 to 412).

Statistical ordination of samples

Viral community composition based on capsid diameter distributions (from qTEM; using 7-nm histogram bin sizes), population abundances, and normalized PC read counts (using only PCs with more than 20 representatives) were compared by using nonmetric multidimensional scaling (NMDS) performed using the "metaMDS" function (default parameters) of the vegan package (*67*) in *R* version 2.15.2 (*68*). The influence of metadata on sample ordination was evaluated using the functions in the vegan package "envfit"—for factor variables including depth category, Longhurst province, and biome—and "ordisurf" for all linear variables (*67*, *69*). Several samples had exceptional environmental conditions (TARA_67_SUR, TARA_70_MESO, TARA_82_DCM, and TARA_85_DCM), thus all statistical ordination analyses were conducted with and without these samples (referred to as the "sample subset") to evaluate their influence.

REFERENCES AND NOTES

1. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998). doi: 10.1126/science.281.5374.237; pmid: 9657713
2. P. G. Falkowski, T. Fenchel, E. F. Delong, The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008). doi: 10.1126/science.1153213; pmid: 18497287
3. M. Breitbart, Marine viruses: Truth or dare. *Annu. Rev. Mar. Sci.* **4**, 425–448 (2012). doi: 10.1146/annurev-marine-120709-142805; pmid: 22457982
4. C. A. Suttle, Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007). doi: 10.1038/nrmicro1750; pmid: 17853907
5. E. Karsenti *et al.*, A holistic approach to marine eco-systems biology. *PLOS Biol.* **9**, e1001177 (2011). doi: 10.1371/journal.pbio.1001177; pmid: 22028628
6. S. A. Solonenko *et al.*, Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics* **14**, 320 (2013). doi: 10.1186/1471-2164-14-320; pmid: 23663384
7. J. R. Brum, R. O. Schenck, M. B. Sullivan, Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J.* **7**, 1738–1751 (2013). doi: 10.1038/ismej.2013.67; pmid: 23635867
8. P. Hingamp *et al.*, Exploring nucleo-cytoplasmic large DNA viruses in *Tara* Oceans microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013). doi: 10.1038/ismej.2013.59; pmid: 23575371
9. S. Yooseph *et al.*, The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PLOS Biol.* **5**, e16 (2007). pmid: 17355171
10. B. L. Hurwitz, M. B. Sullivan, The Pacific Ocean virome (POV): A marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLOS ONE* **8**, e57355 (2013). doi: 10.1371/journal.pone.0057355; pmid: 23468974
11. L. Deng *et al.*, Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature* **513**, 242–245 (2014). doi: 10.1038/nature13459; pmid: 25043051
12. M. B. Duhaime, M. B. Sullivan, Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* **434**, 181–186 (2012). doi: 10.1016/j.virol.2012.09.036; pmid: 23084423
13. M. B. D. Duhaime, L. Deng, B. T. Poulos, M. B. Sullivan, Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: A rigorous assessment and optimization of the linker amplification method. *Environ. Microbiol.* **14**, 2526–2537 (2012). doi: 10.1111/j.1462-2920.2012.02791.x; pmid: 22713159
14. B. L. Hurwitz, L. Deng, B. T. Poulos, M. B. Sullivan, Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.* **15**, 1428–1440 (2013). doi: 10.1111/j.1462-2920.2012.02836.x; pmid: 22845467
15. S. G. John *et al.*, A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.* **3**, 195–202 (2011). doi: 10.1111/j.1758-2229.2010.00208.x; pmid: 21572525
16. G. F. Steward *et al.*, Are we missing half of the viruses in the ocean? *ISME J.* **7**, 672–679 (2013). doi: 10.1038/ismej.2012.121; pmid: 23151645

17. K. Holmfeldt, D. Odić, M. B. Sullivan, M. Middelboe, L. Riemann, Cultivated single-stranded DNA phages that infect marine Bacteroidetes prove difficult to detect with DNA-binding stains. *Appl. Environ. Microbiol.* **78**, 892–894 (2012). doi: 10.1128/AEM.06580-11; pmid: 22138992
18. Y. Tomaru, K. Nagasaki, Flow cytometric detection and enumeration of DNA and RNA viruses infecting marine eukaryotic microalgae. *J. Oceanogr.* **63**, 215–221 (2007). doi: 10.1007/s10872-007-0023-8
19. C. P. D. Brussaard, D. Marie, G. Bratbak, Flow cytometric detection of viruses. *J. Virol. Methods* **85**, 175–182 (2000). doi: 10.1016/S0166-0934(99)00167-6; pmid: 10716350
20. Y. Bettarel, T. Sime-Ngando, C. Amblard, H. Laveran, A comparison of methods for counting viruses in aquatic systems. *Appl. Environ. Microbiol.* **66**, 2283–2289 (2000). doi: 10.1128/AEM.66.6.2283-2289.2000; pmid: 10831400
21. K. P. Hennes, C. A. Suttle, Direct counts of viruses in natural waters and laboratory cultures by epifluorescence microscopy. *Limnol. Oceanogr.* **40**, 1050–1055 (1995). doi: 10.4319/lo.1995.40.6.1050
22. M. G. Weinbauer, C. A. Suttle, Comparison of epifluorescence and transmission electron microscopy for counting viruses in natural marine waters. *Aquat. Microb. Ecol.* **13**, 225–232 (1997). doi: 10.3354/ame013225
23. R. T. Noble, J. A. Fuhrman, Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquat. Microb. Ecol.* **14**, 113–118 (1998). doi: 10.3354/ame014113
24. D. Marie, C. P. D. Brussaard, R. Thyrhaug, G. Bratbak, D. Vaulot, Enumeration of marine viruses in culture and natural samples by flow cytometry. *Appl. Environ. Microbiol.* **65**, 45–52 (1999).pmid: 9872758
25. S. Sunagawa, Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
26. D. M. Kristensen *et al*., Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J. Bacteriol.* **195**, 941–950 (2013). doi: 10.1128/JB.01801-12; pmid: 23222723
27. J. C. Ignacio-Espinoza, S. A. Solonenko, M. B. Sullivan, The global virome: Not as big as we thought? *Curr. Opin. Virol.* **3**, 566–571 (2013). doi: 10.1016/j.coviro.2013.07.004; pmid: 23896279
28. B. L. Hurwitz, J. R. Brum, M. B. Sullivan, Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome. *ISME J.* **9**, 472–484 (2015). doi: 10.1038/ismej.2014.143; pmid: 25093636
29. D. P. Tittensor *et al*., Global patterns and predictors of marine biodiversity across taxa. *Nature* **466**, 1098–1101 (2010). doi: 10.1038/nature09329; pmid: 20668450
30. W. J. Sul, T. A. Oliver, H. W. Ducklow, L. A. Amaral-Zettler, M. L. Sogin, Marine bacteria exhibit a bipolar distribution. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2342–2347 (2013). doi: 10.1073/pnas.1212424110; pmid: 23324742
31. D. I. Jarvis *et al*., A global perspective of the richness and evenness of traditional crop-variety diversity maintained by farming communities. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 5326–5331 (2008). doi: 10.1073/pnas.0800607105; pmid: 18362337
32. S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton Univ. Press, Princeton, NJ, 2001).
33. J. R. Brum, M. B. Sullivan, Rising to the challenge: Accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13**, 147–159 (2015). doi: 10.1038/nrmicro3404; pmid: 25639680
34. C. de Vargas *et al*., Global oceans eukaryotic plankton diversity. *Science* **348**, 1261605 (2015).
35. I. Kang, H.-M. Oh, D. Kang, J.-C. Cho, Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 12343–12348 (2013). doi: 10.1073/pnas.1219930110; pmid: 23798439
36. S. J. Labrie *et al*., Genomes of marine cyanopodoviruses reveal multiple origins of diversity. *Environ. Microbiol.* **15**, 1356–1376 (2013). doi: 10.1111/1462-2920.12053; pmid: 23320838
37. M. B. Sullivan *et al*., Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ. Microbiol.* **12**, 3035–3056 (2010). doi: 10.1111/j.1462-2920.2010.02280.x; pmid: 20662890
38. Y. Zhao *et al*., Abundant SAR11 viruses in the ocean. *Nature* **494**, 357–360 (2013). doi: 10.1038/nature11921; pmid: 23407494
39. F. Rohwer *et al*., The complete genomic sequence of the marine phage Roseaphage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr.* **45**, 408–418 (2000). doi: 10.4319/lo.2000.45.2.0408
40. S. Roux *et al*., Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife* **3**, e03125 (2014). doi: 10.7554/eLife.03125; pmid: 25171894
41. C. M. Mizuno, F. Rodriguez-Valera, N. E. Kimes, R. Ghai, Expanding the marine virosphere using metagenomics. *PLOS Genet.* **9**, e1003987 (2013). doi: 10.1371/journal.pgen.1003987; pmid: 24348267
42. B. L. Hurwitz, A. H. Westveld, J. R. Brum, M. B. Sullivan, Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 10714–10719 (2014). doi: 10.1073/pnas.1319778111; pmid: 25002514
43. P. Kukkaro, D. H. Bamford, Virus-host interactions in environments with a wide range of ionic strengths. *Environ. Microbiol. Rep.* **1**, 71–77 (2009). doi: 10.1111/j.1758-2229.2008.00007.x; pmid: 23765723
44. K. E. Wommack, R. R. Colwell, Virioplankton: Viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000). doi: 10.1128/MMBR.64.1.69-114.2000; pmid: 10704475
45. V. Dang, M. B. Sullivan, Emerging methods to study bacteriophage infection at the single-cell level. *Front. Microbiol.* **5**, 724 (2014).
46. L. D. Talley, G. L. Pickard, W. J. Emery, J. H. Swift, *Descriptive Physical Oceanography: An Introduction* (Elsevier, Boston, ed. 6, 2011).
47. D. B. Olson, R. H. Evans, Rings of the Agulhas current. *Deep Sea Res. Pt. I* **33**, 27–42 (1986). doi: 10.1016/0198-0149(86)90106-8
48. E. Villar, Dispersal and remodeling of plankton communities by Agulhas rings. *Science* **348**, 1261447 (2015).
49. M. Breitbart, F. Rohwer, Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13**, 278–284 (2005). doi: 10.1016/j.tim.2005.04.003; pmid: 15936660
50. G. Lima-Mendez *et al*., Top-down determinants of ocean microbial community structure. *Science* **348**, 1262073 (2015).
51. D. M. Needham *et al*., Short-term observations of marine bacterial and viral communities: Patterns, connections and resilience. *ISME J.* **7**, 1274–1285 (2013). doi: 10.1038/ismej.2013.19; pmid: 23446831
52. C.-E. T. Chow, D. Y. Kim, R. Sachdeva, D. A. Caron, J. A. Fuhrman, Top-down controls on bacterial community structure: Microbial network analysis of bacteria, T4-like viruses and protists. *ISME J.* **8**, 816–829 (2014). doi: 10.1038/ismej.2013.199; pmid: 24196323
53. A. Longhurst, *Ecological Geography of the Sea* (Elsevier, London, 2007).
54. S. Pesant *et al*., Open science resources for the discovery and analysis of *Tara* Oceans data. http://biorxiv.org/content/early/2015/05/08/019117 (2015).
55. A. Aminot, R. Kerouel, S. C. Coverly, in *Practical Guidelines for the Analysis of Seawater*, O. Wurl, Eds. (CRC Press, Boca Raton, FL, 2009), pp. 143–178.
56. J. Ras, H. Claustre, J. Uitz, Spatial variability of phytoplankton pigment distributions in the Subtropical South Pacific Ocean: Comparison between *in situ* and predicted data. *Biogeosciences* **5**, 353–369 (2008). doi: 10.5194/bg-5-353-2008
57. L. Van Heukelem, C. S. Thomas, Computer-assisted high-performance liquid chromatography method development with applications to the isolation and analysis of phytoplankton pigments. *J. Chromatogr. A* **910**, 31–49 (2001). doi: 10.1016/S0378-4347(00)00603-4; pmid: 11263574
58. J. M. Gasol, P. A. del Giorgio, Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. *Sci. Mar.* **64**, 197–224 (2000).
59. M. D. Abramoff, P. J. Magalhaes, S. J. Ram, Image processing with ImageJ. *Biophot. Int.* **11**, 36–42 (2004).
60. S. Schloissnig *et al*., Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013). doi: 10.1038/nature11711; pmid: 23222524
61. R. Luo *et al*., SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18 (2012). doi: 10.1186/2047-217X-1-18; pmid: 23587118
62. J. R. Kultima *et al*., MOCAT: A metagenomics assembly and gene prediction toolkit. *PLOS ONE* **7**, e47656 (2012). doi: 10.1371/journal.pone.0047656; pmid: 23082188
63. D. Hyatt *et al*., Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010). doi: 10.1186/1471-2105-11-119; pmid: 20211023
64. S. Roux, M. Krupovic, D. Debroas, P. Forterre, F. Enault, Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* **3**, 130160 (2013). doi: 10.1098/rsob.130160; pmid: 24335607
65. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39** (suppl), W29–W37 (2011). doi: 10.1093/nar/gkr367; pmid: 21593126
66. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi: 10.1038/nmeth.1923; pmid: 22388286
67. J. Oksanen *et al*., vegan: Community Ecology Package, R package version 2.1-27/r2451 (R Core Team, Vienna, 2013).
68. R Core Team, *R: A Language and Environment for Statistical Computing*, version 2.15.2 (R Core Team, Vienna, 2012).
69. S. N. Wood, Fast stable restricted maximum and marginal likellihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc .Series B Stat. Methodol.* **73**, 3–36 (2011). doi: 10.1111/j.1467-9868.2010.00749.x

ACKNOWLEDGMENTS

We thank J. Czekanski-Moir for advice on statistics and L. Coppola for assistance with validating nutrient data. We thank the commitment of the following people and sponsors: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB KU Leuven, Stazione Zoologica Anton Dohrn, University of Milano-Bicocca, Fund for Scientific Research—Flanders, Rega Institute, KU Leuven, The French Ministry of Research, the French Government "Investissements d'Avenir" programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), Paris Sciences et Lettres (PSL) Research University (ANR-11-IDEX-0001-02), ANR (projects POSEIDON/ANR-09-BLAN-0348, PHYTBACK/ANR-2010-1709-01, PROMETHEUS/ANR-09-PCS-GENM-217, TARA-GIRUS/ANR-09-PCS-GENM-218), European Union 7th Framework Programme (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376), European Research Council (ERC) advanced grant award to C.B. (Diatomite: 294823), Gordon and Betty Moore Foundation grant (3790) to M.B.S., Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to S.G.A., TANIT (CONES 2010-0036) from the Agència de Gestió d´Ajusts Universitaris i Reserca to S.G.A., Japan Society for the Promotion of Science KAKENHI grant 26430184 to H.O., The Research Foundation—Flanders (FWO), BIO5, Biosphere 2 to M.B.S., and the Gordon and Betty Moore Foundation through grants GBMF2631 and GBMF3790 to M.B.S. We also thank the support and commitment of Agnès B. and Etienne Bourgois, the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the Education for Development Foundation, Foundation for Biodiversity Research (FRB), the Prince Albert II de Monaco Foundation, the *Tara* schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge excellent assistance from the European Bioinformatics Institute (EBI), in particular G. Cochrane and P. ten Hoopen, as well as the EMBL Advanced Light Microscopy Facility (ALMF), in particular R. Pepperkok. The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. Data described herein is available at EBI (project identifiers PRJEB402 and PRJEB7988) and PANGAEA (see table S1), and the data release policy regarding future public release of *Tara* Oceans data is described in Pesant *et al*. (*54*). We also acknowledge support from University of Arizona (UA) High Performance Computing and High Throughput Computing; the foundation for France-American Cultural Exchange, Partner University Fund program, awarded to Ecole Normale Supérieure and UA; and a grant to UA Ecosystem Genomics Institute through the UA Technology and Research Initiative Fund

and the Water, Environmental, and Energy Solutions Initiative. All authors approved the final manuscript. This article is contribution number 23 of the *Tara* Oceans Expedition. Supplement contains additional data.

***Tara* Oceans Coordinators**
Silvia G. Acinas,[1] Peer Bork,[2,3] Emmanuel Boss,[4] Chris Bowler,[5] Colomban de Vargas,[6,7] Michael Follows,[8] Gabriel Gorsky,[9,31] Nigel Grimsley,[10,11] Pascal Hingamp,[12] Daniele Iudicone,[13] Olivier Jaillon,[14,15,16] Stefanie Kandels-Lewis,[2,17] Lee Karp-Boss,[18] Eric Karsenti,[5,17] Uros Krzic,[19] Fabrice Not,[6,7] Hiroyuki Ogata,[20] Stephane Pesant,[21,22] Jeroen Raes,[23,24,25] Emmanuel G. Reynaud,[26] Christian Sardet,[27,28] Mike Sieracki,[29]† Sabrina Speich,[30]‡ Lars Stemmann,[9,31] Matthew B. Sullivan,[32] Shinichi Sunagawa,[2] Didier Velayoudon,[33] Jean Weissenbach,[14,15,16] Patrick Wincker[14,15,16]

[1]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Pg. Marítim de la Barceloneta 37-49, E08003 Barcelona, Spain. [2]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [3]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. [4]School of Marine Sciences, University of Maine, Orono, ME 04469, USA. [5]Institut de Biologie de l' Ecole Normale Supérieure (IBENS), and INSERM U1024, and CNRS UMR 8197, 75005 Paris , France. [6]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Sorbonne Universités, Uiversité Pierre et Marie Curie, Univ Paris 06, and UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [8]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [9]CNRS, UMR 7093, Laboratoire d'océanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-mer, France. [10]CNRS UMR 7232, Biologie Intégrative des Organismes Marins (BIOM), Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [11]Observatoire Océanologique de Banyuls (OOB), Sorbonne Universités, Pierre et Marie Curie Paris 06, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [12]Aix Marseille Université, CNRS IGS UMR 7256 13288 Marseille, France. [13]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [14] Genoscope, Commissariat à l'Energie Atomique (CEA)—Institut de Génomique, 2 rue Gaston Crémieux, 91057 Evry, France. [15]CNRS, UMR 8030, CP5706, 91057 Evry, France. [16]Université d'Evry, UMR 8030, CP5706, 91057 Evry, France. [17]Directors' Research, European Molecular Biology Laboratory Meyerhofstrasse 1 69117 Heidelberg, Germany. [18]School of Marine Sciences, University of Maine, Orono, ME 04469,USA. [19]Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [20]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0001, Japan. [21]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, 28359 Bremen, Germany. [22]MARUM, Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany. [23]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [24]Center for the Biology of Disease, VIB KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [25]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [26]The Rlab, University College Dublin, Belfield, Dublin, Ireland. [27]Biodev, Observatoire Océanologique, CNRS, UMR 7009, 06230 Villefranche-sur-mer, France. [28]Sorbonne Universités, Uiversité Pierre et Marie Curie, Univ Paris 06, UMR 7009 Biodev, 06230 Observatoire Océanologique, Villefranche-sur-mer, France. [29]Bigelow Laboratory for Ocean Science, East Boothbay, ME 04544, USA. [30]Laboratoire de Physique des Océan, Institut Universitaire Européen de la Mer, Université de Bretagne Occidentale (UBO-IUEM), Place Copernic, 29820 Plouzané, France. [31]Sorbonne Universités, Uiversité Pierre et Marie Curie, Univ Paris 06, UMR 7093, Laboratoire d'océanographie de Villefranche (LOV), Observatoire Océanologique, 06230, Villefranche-sur-mer, France. [32]Department of Ecology and Evolutionary Biology, Depts Molecular and Cellular Biology and Soil, Water and Environmental Science, University of Arizona, Tucson, AZ 85721, USA. [33]DVIP Consulting, 92310 Sèvres, France. †Present address: National Science Foundation, Arlington, VA 22230, USA. ‡Present address: Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue L'homond 75231 Paris, Cedex 05, France. §Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/1261498/suppl/DC1
File S1
References
Databases

19 September 2014; accepted 25 February 2015
10.1126/science.1261498

OCEAN PLANKTON

Eukaryotic plankton diversity in the sunlit ocean

Colomban de Vargas,[1,2]*† Stéphane Audic,[1,2]† Nicolas Henry,[1,2]† Johan Decelle,[1,2]† Frédéric Mahé,[3,1,2]† Ramiro Logares,[4] Enrique Lara,[5] Cédric Berney,[1,2] Noan Le Bescot,[1,2] Ian Probert,[6,7] Margaux Carmichael,[1,2,8] Julie Poulain,[9] Sarah Romac,[1,2] Sébastien Colin,[1,2,8] Jean-Marc Aury,[9] Lucie Bittner,[10,11,8,1,2] Samuel Chaffron,[12,13,14] Micah Dunthorn,[3] Stefan Engelen,[9] Olga Flegontova,[15,16] Lionel Guidi,[17,18] Aleš Horák,[15,16] Olivier Jaillon,[9,19,20] Gipsi Lima-Mendez,[12,13,14] Julius Lukeš,[15,16,21] Shruti Malviya,[8] Raphael Morard,[22,1,2] Matthieu Mulot,[5] Eleonora Scalco,[23] Raffaele Siano,[24] Flora Vincent,[13,8] Adriana Zingone,[23] Céline Dimier,[1,2,8] Marc Picheral,[17,18] Sarah Searson,[17,18] Stefanie Kandels-Lewis,[25,26] *Tara* Oceans Coordinators‡ Silvia G. Acinas,[4] Peer Bork,[25,27] Chris Bowler,[8] Gabriel Gorsky,[17,18] Nigel Grimsley,[28,29] Pascal Hingamp,[30] Daniele Iudicone,[23] Fabrice Not,[1,2] Hiroyuki Ogata,[31] Stephane Pesant,[32,22] Jeroen Raes,[12,13,14] Michael E. Sieracki,[33,34] Sabrina Speich,[35,36] Lars Stemmann,[17,18] Shinichi Sunagawa,[25] Jean Weissenbach,[9,19,20] Patrick Wincker,[9,19,20]* Eric Karsenti[26,8]*

Marine plankton support global biological and geochemical processes. Surveys of their biodiversity have hitherto been geographically restricted and have not accounted for the full range of plankton size. We assessed eukaryotic diversity from 334 size-fractionated photic-zone plankton communities collected across tropical and temperate oceans during the circumglobal *Tara* Oceans expedition. We analyzed 18*S* ribosomal DNA sequences across the intermediate plankton-size spectrum from the smallest unicellular eukaryotes (protists, >0.8 micrometers) to small animals of a few millimeters. Eukaryotic ribosomal diversity saturated at ~150,000 operational taxonomic units, about one-third of which could not be assigned to known eukaryotic groups. Diversity emerged at all taxonomic levels, both within the groups comprising the ~11,200 cataloged morphospecies of eukaryotic plankton and among twice as many other deep-branching lineages of unappreciated importance in plankton ecology studies. Most eukaryotic plankton biodiversity belonged to heterotrophic protistan groups, particularly those known to be parasites or symbiotic hosts.

The sunlit surface layer of the world's oceans functions as a giant biogeochemical membrane between the atmosphere and the ocean interior (*1*). This biome includes plankton communities that fix CO_2 and other elements into biological matter, which then enters the food web. This biological matter can be remineralized or exported to the deeper ocean, where it may be sequestered over ecological to geological time scales. Studies of this biome have typically focused on either conspicuous phyto- or zooplankton at the larger end of the organismal size spectrum or microbes (prokaryotes and viruses) at the smaller end. In this work, we studied the taxonomic and ecological diversity of the intermediate size spectrum (from 0.8 μm to a few millimeters), which includes all unicellular eukaryotes (protists) and ranges from the smallest protistan cells to small animals (*2*). The ecological biodiversity of marine planktonic protists has been analyzed using Sanger (*3*–*5*) and high-throughput (*6*, *7*) sequencing of mainly ribosomal DNA (rDNA) gene markers, on relatively small taxonomic and/or geographical scales, unveiling key new groups of phagotrophs (*8*), parasites (*9*), and phototrophs (*10*). We sequenced 18*S* rDNA metabarcodes up to local and global saturations from size-fractionated plankton communities sampled systematically across the world tropical and temperate sunlit oceans.

A global metabarcoding approach

To explore patterns of photic-zone eukaryotic plankton biodiversity, we generated ~766 million raw rDNA sequence reads from 334 plankton samples collected during the circumglobal *Tara* Oceans expedition (*11*). At each of 47 stations, plankton communities were sampled at two water-column depths corresponding to the main hydrographic structures of the photic zone: subsurface mixed-layer waters and the deep chlorophyll maximum (DCM) at the top of the thermocline. A low-shear, nonintrusive peristaltic pump and plankton nets of various mesh sizes were used on board *Tara* to sample and concentrate appropriate volumes of seawater to theoretically recover complete local eukaryotic biodiversity from four major organismal size fractions: piconanoplankton (0.8 to 5 μm), nanoplankton (5 to 20 μm), microplankton (20 to 180 μm), and mesoplankton (180 to 2000 μm) [see (*12*) for detailed *Tara* Oceans field sampling strategy and protocols].

We extracted total DNA from all samples, polymerase chain reaction (PCR)–amplified the hypervariable V9 region of the nuclear gene that encodes 18*S* rRNA (*13*), and generated an average of 1.73 ± 0.65 million sequence reads (paired-end Illumina) per sample (*11*). Strict bioinformatic quality control led to a final data set of 580 million reads, of which ~2.3 million were distinct,

[1]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [2]Sorbonne Universités, Université Pierre et Marie Curie (UPMC) Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [3]Department of Ecology, University of Kaiserslautern, Erwin-Schroedinger Street, 67663 Kaiserslautern, Germany. [4]Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)–Consejo Superior de Investigaciones Científicas (CSIC), Passeig Marítim de la Barceloneta 37-49, Barcelona E08003, Spain. [5]Laboratory of Soil Biology, University of Neuchâtel, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland. [6]CNRS, FR2424, Roscoff Culture Collection, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Sorbonne Universités, UPMC Paris 06, FR 2424, Roscoff Culture Collection, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [8]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, Paris, F-75005 France. [9]Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91000 Evry, France. [10]CNRS FR3631, Institut de Biologie Paris-Seine, F-75005, Paris, France. [11]Sorbonne Universités, UPMC Paris 06, Institut de Biologie Paris-Seine, F-75005, Paris, France. [12]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [13]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [14]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [15]Institute of Parasitology, Biology Centre, Czech Academy of Sciences, Branišovská 31, 37005 České Budějovice, Czech Republic. [16]Faculty of Science, University of South Bohemia, Branišovská 31, 37005 České Budějovice, Czech Republic. [17]CNRS, UMR 7093, Laboratoire d'Océanographie de Villefranche-sur-Mer (LOV), Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [18]Sorbonne Universités, UPMC Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [19]CNRS, UMR 8030, CP5706, Evry, France. [20]Université d'Evry, UMR 8030, CP5706, Evry, France. [21]Canadian Institute for Advanced Research, 180 Dundas Street West, Suite 1400, Toronto, Ontario M5G 1Z8, Canada. [22]MARUM, Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany. [23]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [24]Ifremer, Centre de Brest, DYNECO/Pelagos CS 10070, 29280 Plouzané, France. [25]Structural and Computational Biology, European Molecular Biology Laboratory (EMBL), Meyerhofstraße 1, 69117 Heidelberg, Germany. [26]Directors' Research, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [27]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. [28]CNRS UMR 7232, Biologie Intégrative des Organismes Marins (BIOM), Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [29]Sorbonne Universités Paris 06, Observatoire Océanologique de Banyuls (OOB) UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [30]Aix Marseille Université, CNRS IGS UMR 7256, 13288 Marseille, France. [31]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. [32]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [33]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA. [34]National Science Foundation, Arlington, VA 22230, USA. [35]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France. [36]Laboratoire de Physique des Océans, Université de Bretagne Occidentale (UBO)–Institut Universitaire Européen de la Mer (IUEM), Place Copernic, 29820 Plouzané, France.

***Corresponding author. E-mail: vargas@sb-roscoff.fr (C.d.V.); pwincker@genoscope.cns.fr (P.W.); karsenti@embl.de (E.K.)**

†These authors contributed equally to this work. ‡*Tara* Oceans Coordinators and affiliations appear at the end of this paper.

hereafter denoted "metabarcodes." We then clustered metabarcodes into biologically meaningful operational taxonomic units (OTUs) (*14*) and assigned a eukaryotic taxonomic path to all metabarcodes and OTUs by global similarity analysis with 77,449 reference, Sanger-sequenced V9 rDNA barcodes covering the known diversity of eukaryotes and assembled into an in-house database called *V9_PR2* (*15*). Beyond taxonomic assignation, we inferred basic trophic and symbiotic ecological modes (photo- versus heterotrophy; parasitism, commensalism, mutualism for both hosts and symbionts) to *Tara* Oceans reads and OTUs on the basis of their genetic affiliation to large monophyletic and monofunctional groups of reference barcodes. We finally inferred large-scale ecological patterns of eukaryotic biodiversity across geography, taxonomy, and organismal size fractions based on rDNA abundance data and community similarity analyses and compared them to current knowledge extracted from the literature.

The extent of eukaryotic plankton diversity in the photic zone of the world ocean

Sequencing of ~1.7 million V9 rDNA reads from each of the 334 size-fractionated plankton samples was sufficient to approach saturation of eukaryotic richness at both local and global scales (Fig. 1, A and B). Local richness represented, on average, 9.7 ± 4% of global richness, the latter approaching saturation at ~2 million eukaryotic metabarcodes or ~110,000 OTUs (*16*). The global pool of OTUs displayed a good fit to the truncated Preston log-normal distribution (*17*), which, by extrapolation, suggests a total photic-zone eukaryotic plankton richness of ~150,000 OTUs, of which ~40,000 were not found in our survey (Fig. 1C). Thus, we estimate that our survey unveiled ~75% of eukaryotic ribosomal diversity in the globally distributed water masses analyzed. The extrapolated ~150,000 total OTUs is much higher than the ~11,200 formally described species of marine eukaryotic plankton (see below) and probably represents a highly conservative, lower-boundary estimate of the true number of eukaryotic species in this biome, given the relatively limited taxonomic resolution power of the 18*S* rDNA gene. Our data indicate that eukaryotic taxonomic diversity is higher in smaller organismal size fractions, with a peak in the piconanoplankton (Fig. 1A), highlighting the richness of tiny organisms that are poorly characterized in terms of morphotaxonomy and physiology (*18*). A first-order, supergroup-level classification of all *Tara* Oceans OTUs demonstrated the prevalence (at the biome scale and across the >four orders of size magnitude sampled) of protist rDNA biodiversity with respect to that of classical multicellular eukaryotes, i.e., animals, plants, and fungi (Fig. 2A). Protists accounted for >85% of total eukaryotic ribosomal diversity, a ratio that may well hold true for other marine, freshwater, and terrestrial oxygenic ecosystems (*19*). The latest estimates of total marine eukaryotic biodiversity based on statistical extrapolations from classical taxonomic knowledge predict the existence of 0.5 to 2.2 million species [including all benthic and planktonic systems from reefs to deep-sea vents (*20*, *21*)] but do not take into account the protistan knowledge gap highlighted here. Simple application of our animal–to–other eukaryotes ratio of ~13% to the robust prediction of the total number of metazoan species from (*20*) would imply that 16.5 million and 60 million eukaryotic species potentially inhabit the oceans and Earth, respectively.

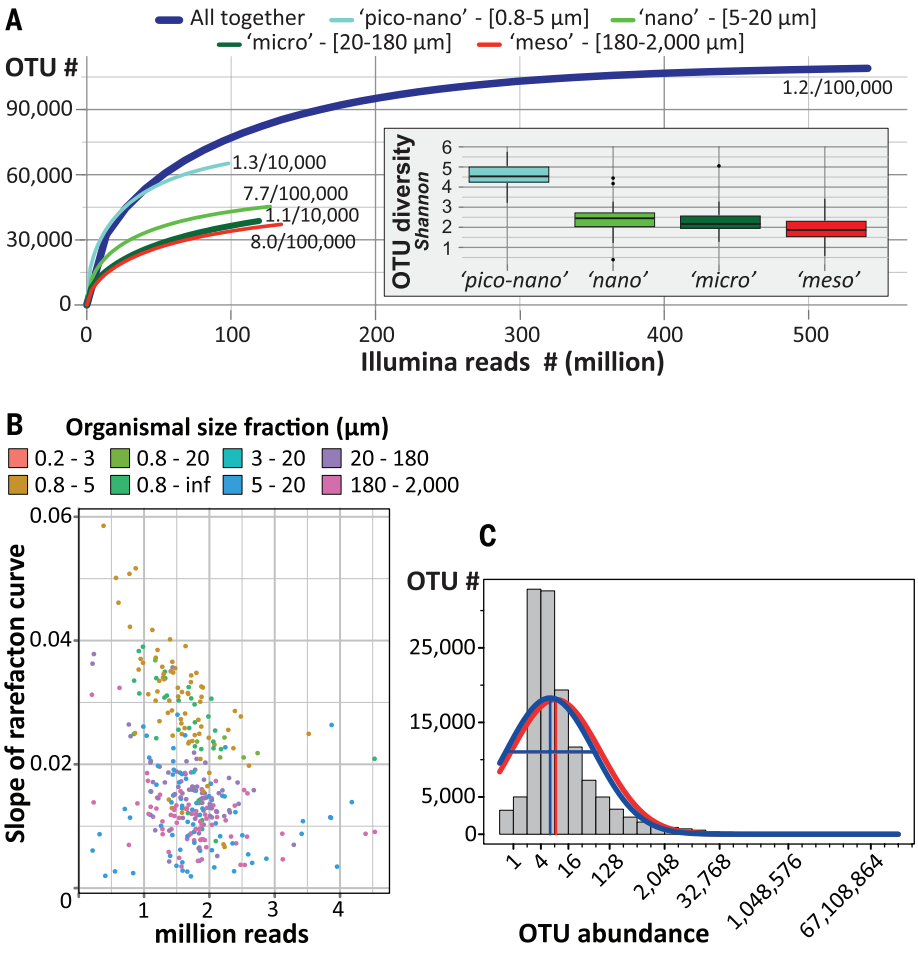


Fig. 1. Photic-zone eukaryotic plankton ribosomal diversity. (**A**) V9 rDNA OTUs rarefaction curves and overall diversity (Shannon index, inset) for each plankton organismal size fraction. Proximity to saturation is indicated by weak slopes at the end of each rarefaction curve (e.g., 1.2/100,000 means 1.2 novel metabarcodes obtained every 100,000 rDNA reads sequenced). (**B**) Saturation slope versus number of V9 rDNA reads for all of the 334 samples (dots) analyzed herein. A slope of 0.02 indicates that two novel barcodes can be recovered if 100 new reads are sequenced. Samples are colored according to size fraction. (**C**) Global OTU abundance distribution and fit to the Preston log-normal model. Most OTUs in our data set were represented by 3 to 16 reads, whereas fewer OTUs presented less or more abundances. Quasi-Poisson fit to octaves (red curve) and maximized likelihood to $\log_2$ abundances (blue curve) approximations were used to fit the OTU abundance distribution to the Preston log-normal model. Overall, the global (A) and local (B) saturation values indicate that our extensive sampling effort (in terms of spatiotemporal coverage and sequencing depth) uncovered the majority of eukaryotic ribosomal diversity within the photic layer of the world's tropical to temperate oceans. Calculation of the Preston veil, which infers the number of OTUs that we missed (or were veiled) during our sampling (~40,000), confirmed that we captured most of the protistan richness, thus allowing extraction of holistic and general patterns of eukaryotic plankton biodiversity from our data set.

Phylogenetic breakdown of photic-zone eukaryotic biodiversity

About one-third of eukaryotic ribosomal diversity in our data set did not match any reference barcode in the extensive *V9_PR2* database ("unassigned" category in Fig. 2A). This unassignable diversity represented only a small proportion (2.6%) of total reads and increased in both richness and abundance in smaller organismal size fractions, suggesting that it corresponds mostly to rare and minute taxa that have escaped previous characterization. Some may also correspond to divergent rDNA pseudogenes, known to exist in eukaryotes (*22*, *23*) or sequencing artefacts (*24*), although both of these would be expected to be present in equal proportion in all

size fractions [details in (*16*)]. The remaining ~87,000 assignable OTUs were classified into 97 deep-branching lineages covering the full spectrum of cataloged eukaryotic diversity amongst the seven recognized supergroups and multiple lineages of uncertain placement (*15*) whose origins go back to the primary radiation of eukaryotic life in the Neoproterozoic. Although highly represented in the *V9_PR2* reference database, several well-known lineages adapted to terrestrial, marine benthic, or anaerobic habitats (e.g., Embryophyta; apicomplexan and trypanosome parasites of land plants and animals; amoeboflagellate Breviatea; and several lineages of Amoebozoa, Excavata, and Cercozoa) were not detected in our metabarcoding data set, suggesting the absence of contamination during the PCR and sequencing steps on land and reducing the number of deep branches of eukaryotic plankton to 85 (Fig. 3).

We then extracted the metabarcodes assigned to morphologically well-known planktonic eukaryotic taxa from our data set and compared them with the conventional, 150 year-old morphological view of marine eukaryotic plankton that includes ~11,200 cataloged species divided into three broad categories: ~4350 species of phytoplankton (microalgae), ~1350 species of protozooplankton (relatively large, often biomineralized, heterotrophic protists), and ~5500 species of metazooplankton (holoplanktonic animals) (*25*–*27*). A congruent picture of the distribution of morphogenetic diversity among and within these organismal categories emerged from our data set (Fig. 2B), but typically, three to eight times more rDNA OTUs were found than described morphospecies in the best-known lineages within these categories. This is within the range of the number of cryptic species typically detected in globally-distributed pelagic taxa using molecular data (*28*, *29*). The general congruency between genetic and morphological data in the cataloged compartment of eukaryotic plankton suggests that the protocols used, from plankton sampling to DNA sequencing, recovered the known eukaryotic biodiversity without major qualitative or quantitative biases. However, OTUs related to morphologically described taxa represented only a minor part of the total eukaryotic plankton ribosomal and phylogenetic diversity. Overall, <1% of OTUs were strictly identical to reference sequences, and OTUs were, on average, only ~86% similar to any V9 reference sequence (Fig. 3F) (*16*). This shows that most photic-zone eukaryotic plankton V9 rDNA diversity had not been previously sequenced from cultured strains, single-cell isolates, or even environmental clone library surveys. The *Tara* Oceans metabarcode data set added considerable phylogenetic information to previous protistan rDNA knowledge, with an estimated mean tree-length increase of 453%, reaching >100% in 43 lineages (*16*). Even in the best-referenced groups such as the diatoms (1232 reference sequences) (Fig. 3B), we identified many new rDNA sequences, both within known groups and forming new clades (*16*).

Eleven "hyperdiverse" lineages each contained >1000 OTUs, together representing ~88 and

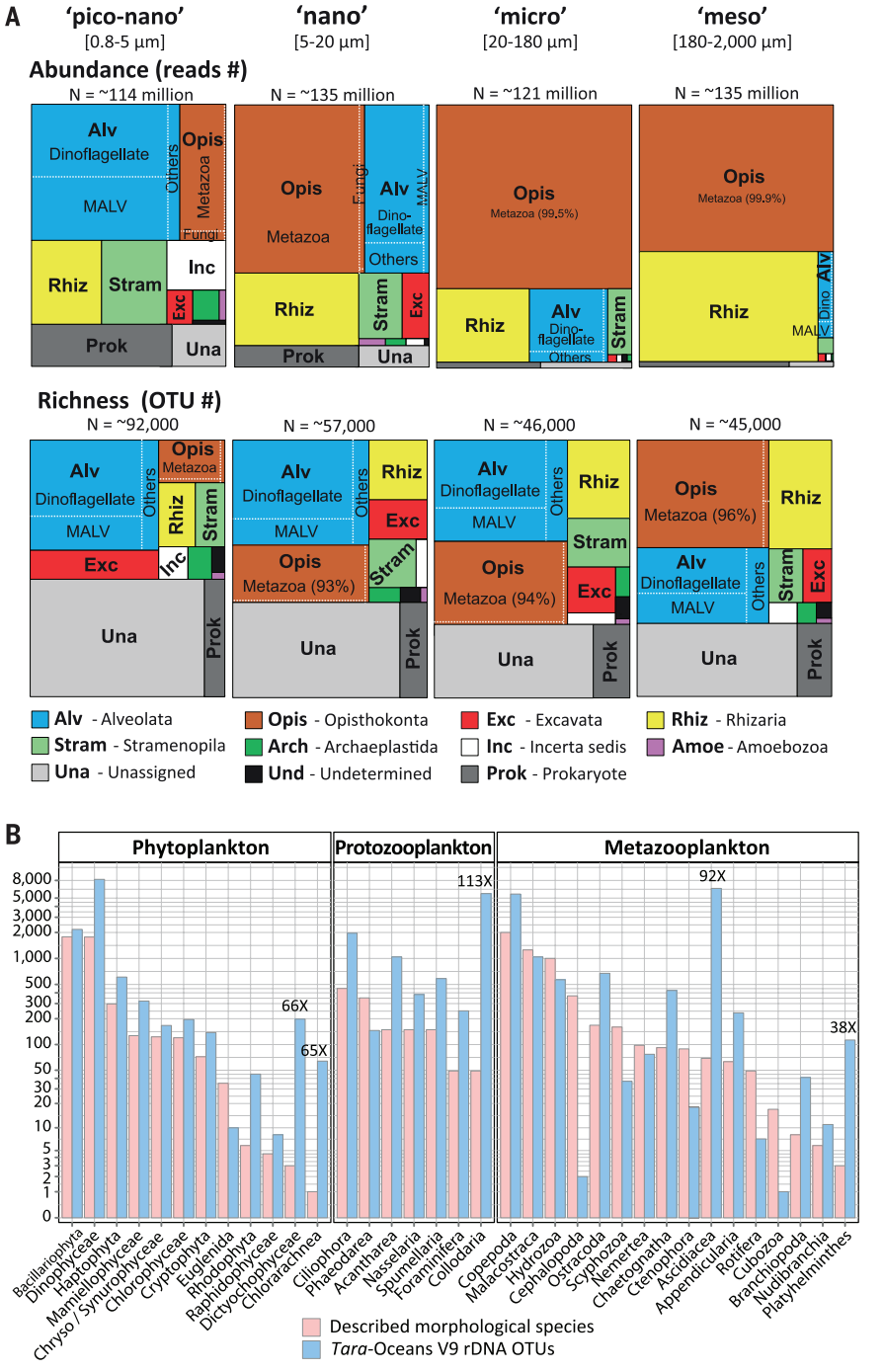


Fig. 2. Unknown and known components of eukaryotic plankton biodiversity. (**A**) Phylogenetic breakdown of the entire metabarcoding data set at the eukaryotic supergroup level. All *Tara* Oceans V9 rDNA reads and OTUs were classified among the seven recognized eukaryotic supergroups plus the known but unclassified deep-branching lineages (incertae sedis). The tree maps display the relative abundance (upper part) and richness (lower part) of the different eukaryotic supergroups in each organismal size fraction. Note that ~5% of barcodes were assigned to prokaryotes, essentially in the piconano fraction, witnessing the universality of the eukaryotic primers used. Barcodes are "unassigned" when sequence similarity to a reference sequence is <80% and "undetermined" when eukaryotic supergroups could not be discriminated (at similarity >80%). (**B**) Ribosomal DNA diversity associated with the morphologically known and cataloged part of eukaryotic plankton. The total number of morphologically described species in the literature [red bars, based on (*25*–*27*)] and the corresponding total number of *Tara* Oceans V9 rDNA OTUs (blue bars) are indicated for each of the 35 classical lineages of eukaryotic phyto-, protozoo-, and metazooplankton. The five classical groups that were found to be substantially more diverse than previously thought (from 38- to 113-fold more OTUs than morphospecies) are highlighted. Note that in the classical morphological view, phyto- and metazooplankton comprise ~88% of total eukaryotic plankton diversity.

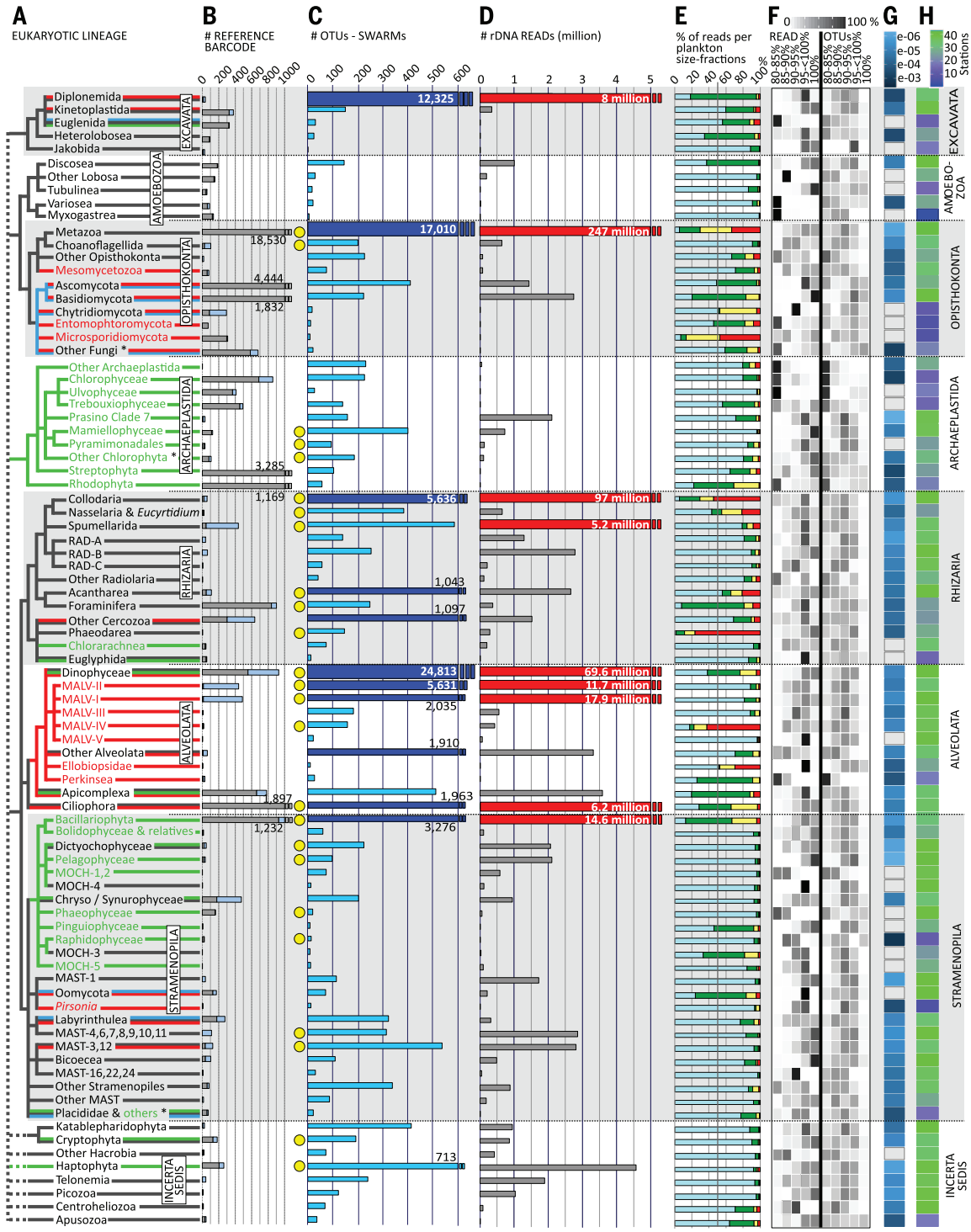


Fig. 3. Phylogenetic distribution of the assignable component of eukaryotic plankton ribosomal diversity. (**A**) Schematic phylogeny of the 85 deep-branching eukaryotic lineages represented in our global oceans metabarcoding data set, with broad ecological traits based on current knowledge: red, parasitic; green, photoautotrophic; blue, osmo- or saprotrophic; black, mostly phagotrophic lineages. Lineages known only from environmental sequence data were colored in black by default. For simplicity, three branches (denoted by asterisks) artificially group a few distinct lineages [details in (*15*)]. (**B**) Number of reference V9 rDNA barcodes used to annotate the metabarcoding data set (gray, with known taxonomy at the genus and/or species level; light blue, from previous 18S rDNA environmental clone libraries). (**C**) *Tara* Oceans V9 rDNA OTU richness. Dark blue thicker bars indicate the 11 hyperdiverse lineages containing >1000 OTUs. Yellow circles highlight the 25 lineages that have been recognized as important in previous marine plankton biodiversity and ecology studies using morphological and/or molecular data [see also (*15*)]. (**D**) Eukaryotic plankton abundance expressed as numbers of rDNA reads (the red bars indicate the nine most abundant lineages with >5 million reads). (**E**) Proportion of rDNA reads per organismal size fraction. Light blue, piconano-; green, nano-; yellow, micro-; red, mesoplankton. (**F**) Percentage of reads and OTUs with 80 to 85%, 85 to 90%, 90 to 95%, 95 to <100%, and 100% sequence similarity to a reference sequence. (**G**) Slope of OTU rarefaction curves. (**H**) Mean geographic occupancy (average number of stations in which OTUs were observed, weighted by OTU abundance).

~90% of all OTUs and reads, respectively (Fig. 3C). Among these, the only permanently phototrophic taxa were diatoms (Fig. 4A) and about one-third of dinoflagellates (Fig. 4, B to F), together comprising ~15 and ~13% of hyperdiverse OTUs and reads, respectively (*30*). Most hyperdiverse photic-zone plankton belonged to three supergroups—the Alveolata, Rhizaria, and Excavata—about which we have limited biological or ecological information. The Alveolata, which consist mostly of parasitic [marine alveolates (MALVs)] (Fig. 4F) and phagotrophic (ciliates and most dinoflagellates) taxa, were by far the most diverse supergroup, comprising ~42% of all assignable OTUs. The Rhizaria are a group of amoeboid heterotrophic protists with active pseudopods displaying a broad spectrum of ecological behavior, from phagotrophy to parasitism and mutualism (symbioses) (*31*). Rhizarian diversity peaked in the Retaria (Fig. 4, C and D) a subgroup including giant protists that build complex skeletons of silicate (Polycystinea), strontium sulfate (Acantharia) (Fig. 4C), or calcium carbonate (Foraminifera) and thus comprise key microfossils for paleoceanography. Unsuspected rDNA diversity was recorded within the Collodaria (5636 OTUs), polycystines that are mostly colonial, poorly silicified, or naked and live in obligatory symbiosis with photosynthetic dinoflagellates (Fig. 4D) (*32*, *33*). Arguably, the most surprising component of novel biodiversity was the >12,300 OTUs related to reference sequences of diplonemids, an excavate lineage that has only two described genera of flagellate grazers, one of which parasitizes diatoms and crustaceans (*34*, *35*). Their ribosomal diversity was not only much higher than that observed in classical plankton groups such as foraminifers, ciliates, or diatoms (50-fold, 6-fold, and 3.8-fold higher, respectively) but was also far from richness saturation (Fig. 3E). Eukaryotic rDNA diversity peaked especially in the few lineages that extend across larger size fractions (i.e., metazoans, rhizarians, dinoflagellates, ciliates, diatoms) (Fig. 3E). Larger cells or colonies not only provide protection against predation via size-mediated avoidance and/or construction of composite skeletons but also provide support for complex and coevolving relationships with often specialized parasites or mutualistic symbionts.

Beyond this hyperdiverse, largely heterotrophic eukaryotic majority, our data set also highlighted the phylogenetic diversity of poorly known phagotrophic (e.g., 413 OTUs of Katablepharidophyta, 240 OTUs of Telonemia), osmotrophic (e.g., 410 OTUs of Ascomycota, 322 OTUs of Labyrinthulea), and parasitic (e.g., 384 OTUs of gregarine apicomplexans, 160 OTUs of Ascetosporea, 68

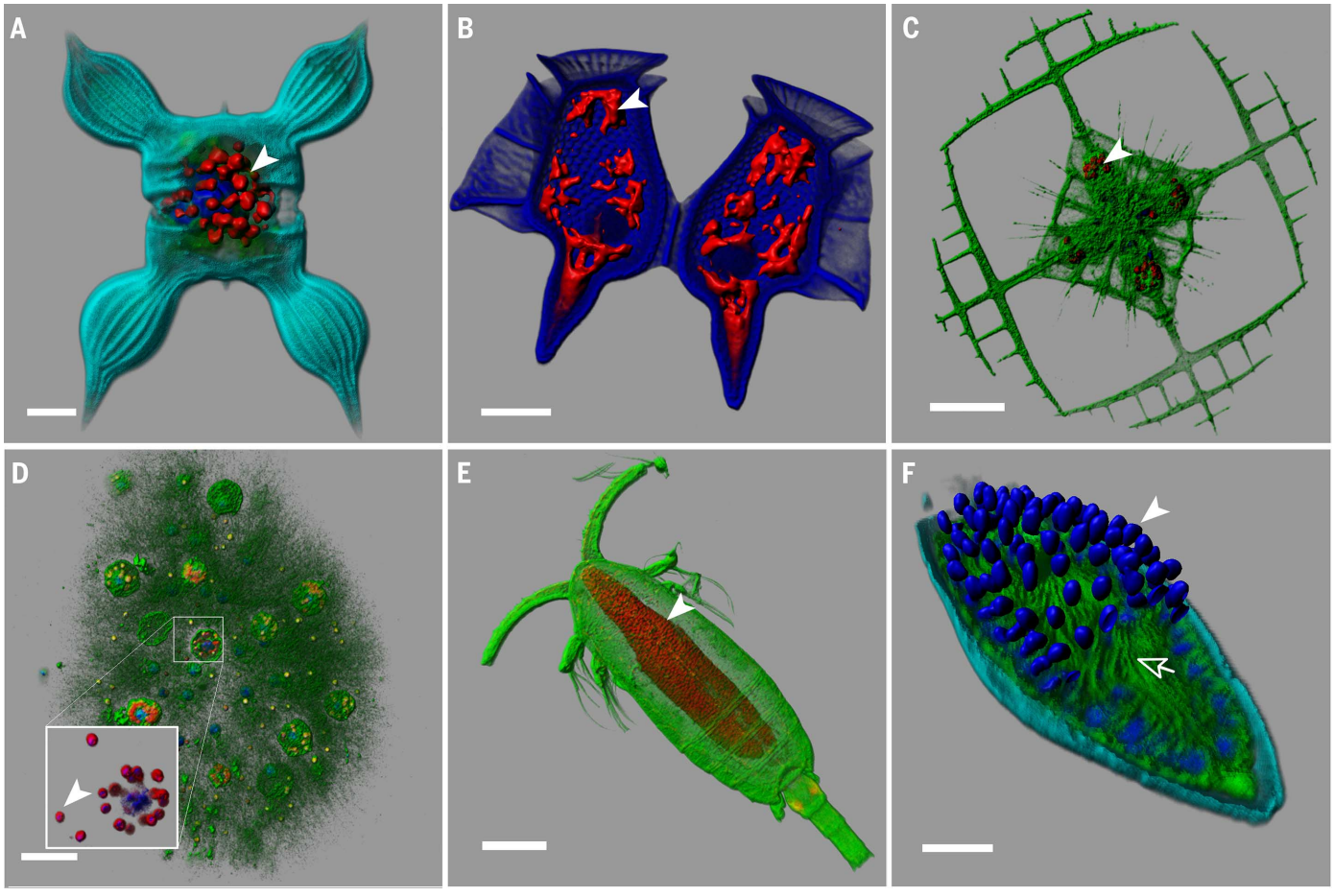


Fig. 4. Illustration of key eukaryotic plankton lineages. (**A**) Stramenopila; a phototrophic diatom *Chaetoceros bulbosus*, with its chloroplasts in red (arrowhead). Scale bar, 10 μm. (**B**) Alveolata; a heterotrophic dinoflagellate *Dinophysis caudata* harboring kleptoplasts [in red (arrowhead)]. Scale bar, 20 μm (*75*). (**C**) Rhizaria; an acantharian *Lithoptera* sp. with endosymbiotic haptophyte cells from the genus *Phaeocystis* [in red (arrowhead)]. Scale bar, 50 μm (*41*). (**D**) Rhizaria; inside a colonial network of Collodaria, a cell surrounded by several captive dinoflagellate symbionts of the genus *Brandtodinium* (arrowhead). Scale bar, 50 μm (*33*). (**E**) Opisthokonta; a copepod whose gut is colonized by the parasitic dinoflagellate *Blastodinium* [red area shows nuclei (arrowhead)]. Scale bar, 100 μm (*51*). (**F**) Alveolata; a cross-sectioned, dinoflagellate cell infected by the parasitoid alveolate *Amoebophrya* (MALV-II). Each blue spot (arrowhead) is the nucleus of future free-living dinospores; their flagella are visible in green inside the mastigocoel cavity (arrow). Scale bar, 5 μm. The cellular membranes were stained with DiOC6 (green); DNA and nuclei were stained with Hoechst (blue) [the dinoflagellate theca in (B) was also stained by this dye]. Chlorophyll autofluorescence is shown in red [except for in (E)]. An unspecific fluorescent painting of the cell surface (light blue) was used to reveal cell shape for (A) and (F). All specimens come from *Tara* Oceans samples preserved for confocal laser scanning fluorescent microscopy. Images were three-dimensionally reconstructed with Imaris (Bitplane).

OTUs of Ichthyosporea) protist groups. Amongst the 85 major lineages presented in the phylogenetic framework of Fig. 3, less than one-third (~25) have been recognized as important in previous marine plankton biodiversity and ecology studies using morphological and/or molecular data (Fig. 3C) (*15*). The remaining ~60 branches had either never been observed in marine plankton or were detected through morphological description of one or a few species and/or the presence of environmental sequences in geographically restricted clone library surveys (*15*). This understudied diversity represents ~25% of all taxonomically assignable OTUs (>21,500) and covers broad taxonomic and geographic scales, thus representing a wealth of new actors to integrate into future plankton systems biology studies.

Insights into photic-zone eukaryotic plankton ecology

Functional annotation of taxonomically assigned V9 rDNA metabarcodes was used as a first attempt to explore ecological patterns of eukaryotic diversity across broad spatial scales and organismal size fractions, focusing on fundamental trophic modes (photo- versus heterotrophy) and symbiotic interactions (parasitism to mutualism). Heterotroph (protists and metazoans) V9 rDNA metabarcodes were substantially more diverse (63%) and abundant (62%) than phototroph metabarcodes that represented <20% of OTUs and reads across all size fractions and geographic sites, with an increasing heterotroph-to-phototroph ratio in the micro- and mesoplankton (Fig. 5A, confirmed in 17 non–size-fractionated samples (*30*). These results challenge the classical morphological view of plankton diversity, biased by a terrestrial ecology approach, whereby phyto- and metazooplankton (the plant-animal paradigm) are thought to comprise ~88% of eukaryotic plankton diversity (Fig. 2B) and heterotrophic protists are typically reduced in food-web modeling to a single entity, often idealized as ciliate grazers.

An unsuspected richness and abundance of metabarcodes assigned to monophyletic groups of heterotrophic protists that cannot survive without endosymbiotic microalgae was found in larger size fractions ("photosymbiotic hosts" in Fig. 5A). Their abundance and even diversity were sometimes greater than those of all metazoan metabarcodes, including those from copepods. Most of these cosmopolitan photosymbiotic hosts were found within the hyperdiverse radiolarians Acantharia (1043 OTUs) and Collodaria (5636 OTUs) (Figs. 3, 4B, and 5D), which have often been overlooked in traditional morphological surveys of plankton-net–collected material because of their delicate gelatinous and/or easily dissolved structures but are known to be very abundant from microscope-based and in situ imaging studies (*36*–*38*). All 95 known colonial collodarian species described since the 19th century (*39*) harbor intracellular symbiotic microalgae, and these key players for plankton ecology are protistan analogs of photosymbiotic corals in tropical coastal reef ecosystems with no equivalent in terrestrial ecology. In addition to their contribution to total primary production (*36*, *38*), these diverse, biologically complex, often biomineralized, and relatively long-lived giant mixotrophic protists stabilize carbon in larger size fractions and probably increase its flux to the ocean interior (*38*). Conversely, the microalgae that are known obligate intracellular partners in open-ocean photosymbioses (*33*, *40*–*42*) (Fig. 5B) were neither very diverse nor highly abundant and occurred evenly across organismal size fractions (Fig. 5C). However, their relative contribution was greatest in the mesoplankton category (10%) (Fig. 5C), where the known photosymbionts of pelagic rhizarians were found (together with their hosts) (Fig. 5B). The stable and systematic abundance of photosymbiotic microalgae across size fractions [a pattern not shown by nonphotosymbiotic microalgae (*30*)] suggests that pelagic photosymbionts maintain free-living and potentially actively growing populations in the piconano- and nanoplankton, representing an accessible pool for recruitment by their heterotrophic hosts. This appears to contrast with photosymbioses in coral reefs and terrestrial systems, where symbiotic microalgal populations mainly occur within their multicellular hosts (*43*).

On the other end of the spectrum of biological interactions, rDNA metabarcodes affiliated to groups of known parasites were ~90 times more diverse than photosymbionts in the piconanoplankton, where they represented ~59% of total heterotrophic protistan ribosomal richness and ~53% of abundance (Figs. 4 and 5C), although this latter value may be inflated by a hypothetically higher rDNA copy number in some marine alveolate lineages (*18*). Parasites in this size fraction were mostly (89% of diversity and 88% of abundance across all stations) within the MALV-I and -II Syndiniales (*30*), which are known exclusively as parasitoid species that kill their hosts and release hundreds of small (2 to 10 μm), nonphagotrophic dinospores (*9*, *44*) that survive for only a few days in the water column (*45*). Abundant parasite-assigned metabarcodes in small size fractions (Fig. 5, B and C) suggest the existence of a large and diverse pool of free-living parasites in photic-zone piconanoplankton, mirroring phage ecology (*46*) and reflecting the extreme diversity and abundance of their known main hosts: radiolarians, ciliates, and dinoflagellates (Fig. 3) (*9*, *47*–*49*). Contrasting with the pattern observed for metabarcodes affiliated to purely phagotrophic taxa, the relative abundance and richness of putative parasite metabarcodes decreased in the nano- and microplanktonic size fractions but increased again in the mesoplankton (Fig. 5C), where parasites are most likely in their infectious stage within larger-sized host organisms. This putative in hospite parasites richness, equivalent to only 23% of that in the piconanoplankton, consisted mostly of a variety of alveolate taxa known to infect crustaceans: MALV-IV such as *Haematodinium* and *Syndinium*; dinoflagellates such as *Blastodinium* (Fig. 4E); and apicomplexan gregarines, mainly *Cephaloidophoroidea* (Fig. 5B) (*9*, *50*, *51*). This pattern contrasts with terrestrial systems where most parasites live within their hosts and are typically transmitted either vertically or through vectors because they generally do not survive outside their hosts (*52*). In the pelagic realm, free-living parasitic spores, like phages, are protected from dessication and dispersed by water diffusion and are apparently massively produced, which likely increases horizontal transmission rate.

Community structuring of photic-zone eukaryotic plankton

Clustering of communities by their compositional similarity revealed the primary influence of organism size ($P = 10^{-3}$, $r^2 = 0.73$) on community structuring, with piconanoplankton displaying stronger cohesiveness than larger organismal size fractions (Fig. 6A). Filtered size-fraction–specific communities separated by thousands of kilometers were more similar in composition than they were to communities from other size fractions at the same location. This was emphasized by the fact that ~36% of all OTUs were restricted to a single size category (*53*). Further analyses within each organismal size fraction indicated that geography plays a role in community structuring, with samples being partially structured according to basin of origin, a pattern that was stronger in larger organismal size fractions ($P = 0.001$ in all cases, $r^2 = 0.255$ for piconanoplankton, 0.371 for nanoplankton, 0.473 for microplankton, and 0.570 for mesoplankton) (Fig. 6B). Mantel correlograms comparing Bray-Curtis community similarity to geographic distances between all samples indicated significant positive correlations in all organismal size fractions over the first ~6000 km, the correlation breaking down at larger geographic distances (*54*). This positive correlation between community dissimilarity and geographic distance, expected under neutral biodiversity dynamics (*55*), challenges the classical niche model for photic-zone eukaryotic plankton biogeography (*56*). The significantly stronger community differentiation by ocean basin in larger organismal size fractions (Fig. 6B) suggests increasing dispersal limitation from piconano- to nano-, micro-, and mesoplankton. Thus, larger-sized eukaryotic plankton communities, containing the highest abundance and diversity of metazoans (Figs. 2A and 5B), were spatially more heterogeneous in terms of both taxonomic (Fig. 6) and functional (Fig. 5A) composition and abundance. The complex life cycle and behaviors of metazooplankton, including temporal reproductive and growth cycles and vertical migrations, together with putative rapid adaptive evolution processes to mesoscale oceanographic features (*57*), may explain the stronger geographic differentiation of mesoplanktonic communities. By contrast, eukaryotic communities in the piconanoplankton were richer (Fig. 1A) and more homogeneous in taxonomic composition (Fig. 6), representing a stable compartment across the world's oceans (*58*).

Even though protistan communities were diverse, the proportions of abundant (>1%) and

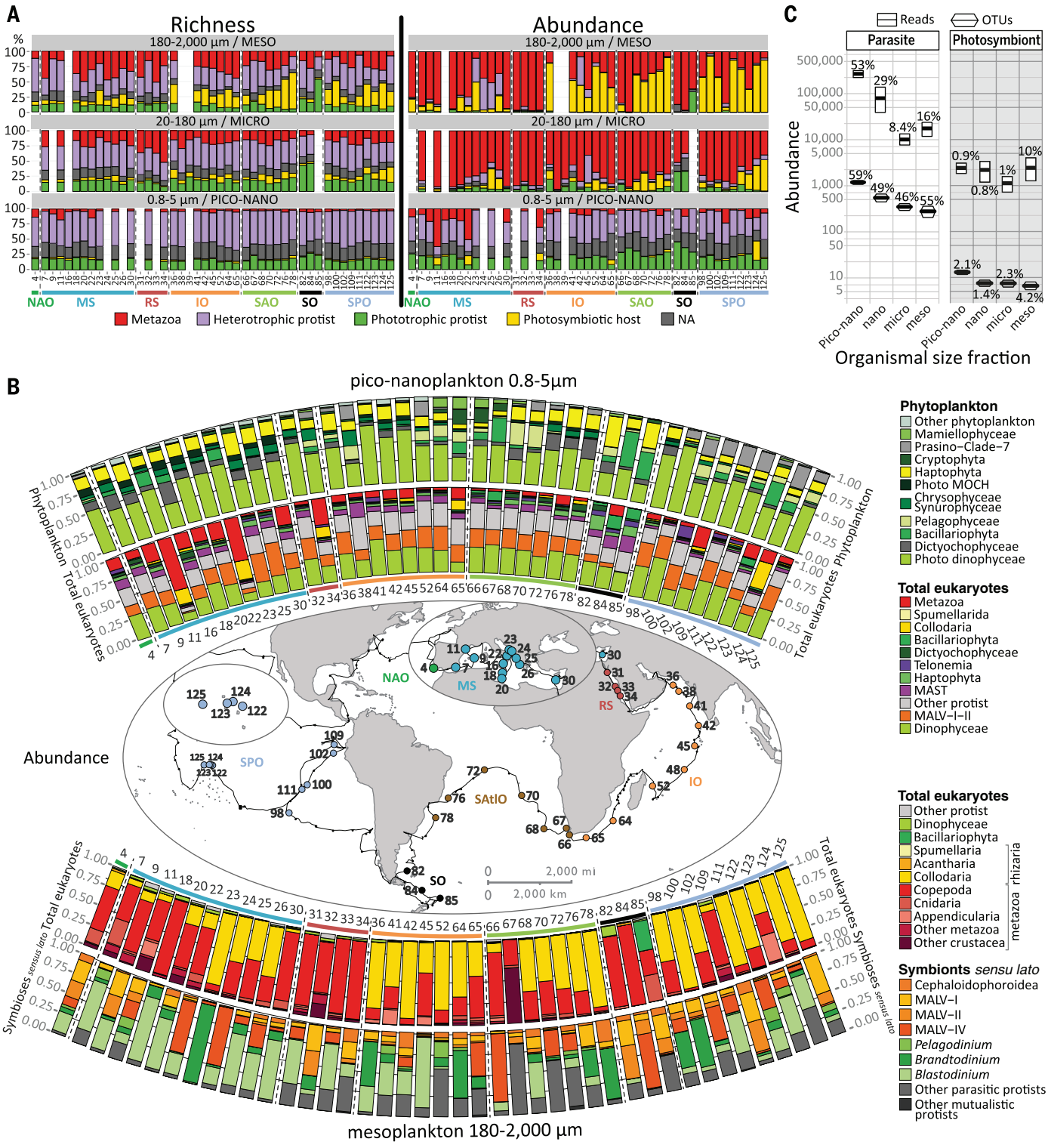


Fig. 5. Metabarcoding inference of trophic and symbiotic ecological diversity of photic-zone eukaryotic plankton. (**A**) Richness (OTU number) and abundance (read number) of rDNA metabarcodes assigned to various trophic taxo-groups across plankton organismal size fractions and stations. Note that the nano size fraction did not contain enough data to be used in this biogeographical analysis [for all size-fraction data, see (*30*)]. NA, not applicable. (**B**) Relative abundance of major eukaryotic taxa across *Tara* Oceans stations for (i) phytoplankton and all eukaryotes in piconanoplankton (above the map) and (ii) all eukaryotes and protistan symbionts (sensu lato) in mesoplankton (below the map). Note the pattern of inverted relative abundance between collodarian colonies (Fig. 4) and copepods in, respectively, the oligotrophic and eutrophic and mesotrophic systems. The dinoflagellates *Brandtodinium* and *Pelagodinium* are endophotosymbionts in Collodaria (*33*) and Foraminifera (*40*, *42*), respectively. (**C**) Richness and abundance of parasitic and photosymbiotic (microalgae) protists across organismal size fractions. The relative contributions (percent) of parasites to total heterotrophic protists and of photosymbionts to total phytoplankton are indicated above each symbol.

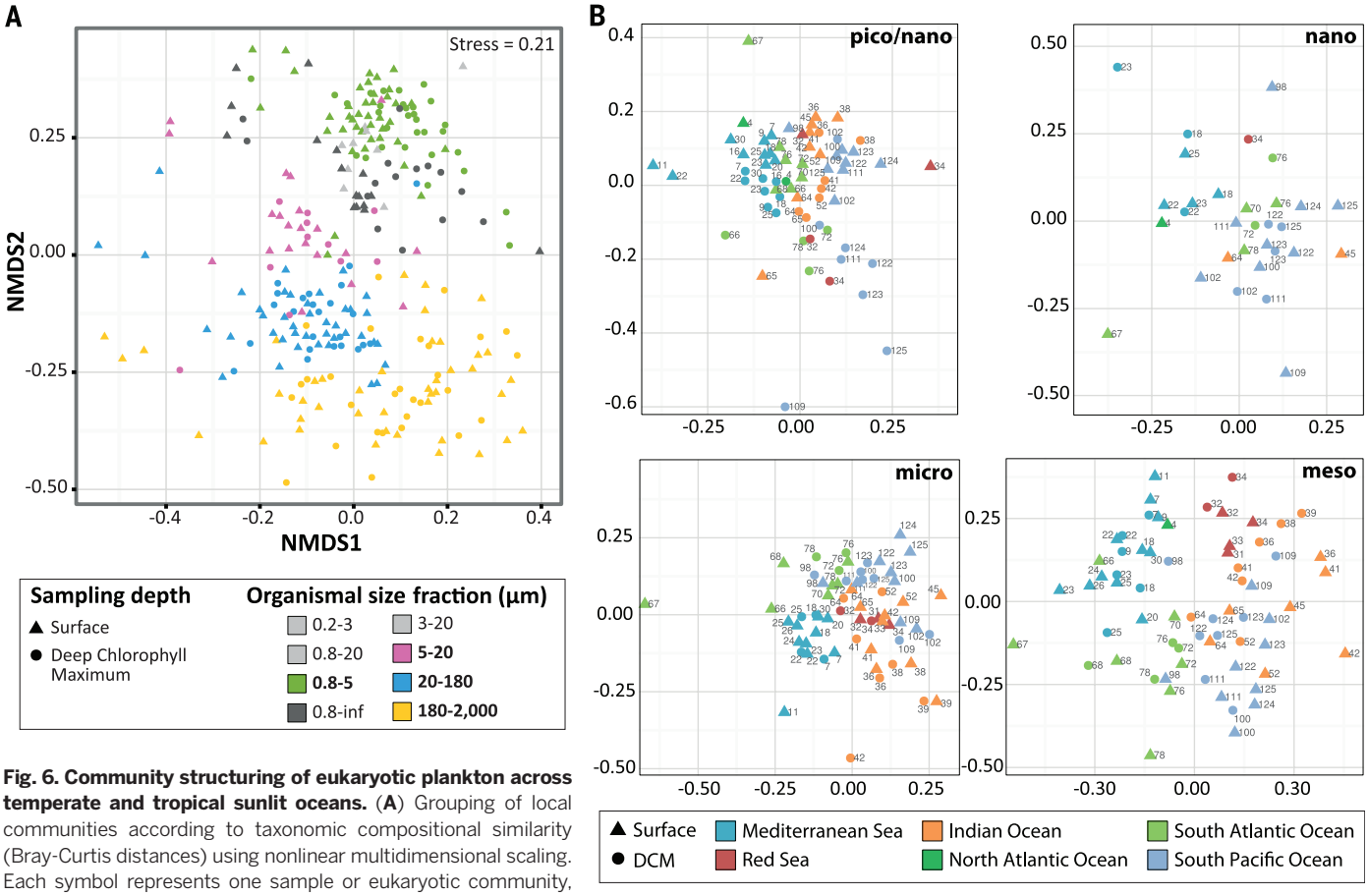


Fig. 6. Community structuring of eukaryotic plankton across temperate and tropical sunlit oceans. (**A**) Grouping of local communities according to taxonomic compositional similarity (Bray-Curtis distances) using nonlinear multidimensional scaling. Each symbol represents one sample or eukaryotic community, corresponding to a particular depth (shape) and organismal size fraction (color). (**B**) Same as in (A), but the different plankton organismal size fractions were analyzed independently, and communities are distinguished by depth (shape) and ocean basins' origin (color). An increasing geographic community differentiation along increasing organismal size fractions is visible and confirmed by the Mantel test [$P = 10^{-3}$, R_m = 0.36, 0.49, 0.50, and 0.51 for the highest piconano- to mesoplankton correlations in Mantel correlograms; see also (*54*)]. In addition, samples from the piconanoplankton only were discriminated by depth (surface versus DCM; $P = 0.001$, $r^2 = 0.2$). The higher diversity and abundance of eukaryotic phototrophs in this fraction (Fig. 5A) may explain overall community structuring by light and, thus, depth.

rare (<0.01%) OTUs were more or less constant across communities, as has been observed in coastal waters (*6*). Only 2 to 17 OTUs (i.e., 0.2 to 8% of total OTUs per and across sample) dominated each community (*54*), suggesting that a small proportion of eukaryotic taxa are key for local plankton ecosystem function. On a worldwide scale, an occurrence-versus-abundance analysis of all ~110,000 *Tara* Oceans OTUs revealed the hyperdominance of cosmopolitan taxa (Fig. 7A). The 381 (0.35% of the total) cosmopolitan OTUs represented ~68% of the total number of reads in the data set. Of these, 269 (71%) OTUs had >100,000 reads and accounted for nearly half (48%) of all rDNA reads (Fig. 7A), a pattern reminiscent of hyperdominance in the largest forest ecosystem on Earth, where only 227 tree species out of an estimated total of 16,000 account for half of all trees in Amazonia (*59*). The cosmopolitan OTUs belonged mainly (314 of 381) to the 11 hyperdiverse eukaryotic planktonic lineages (Fig. 3C) and were essentially phagotrophic (40%) or parasitic (21%), with relatively few (15%) phytoplanktonic taxa (*54*). Of the cosmopolitan OTUs, which represent organisms that are likely among the most abundant eukaryotes on Earth, 25% had poor identity (<95%) to reference taxa, and 11 of these OTUs could not even be affiliated to any available reference sequence (Fig. 7B) (*54*).

Conclusions and perspectives

We used rDNA sequence data to explore the taxonomic and ecological structure of total eukaryotic plankton from the photic oceanic biome, and we integrated these data with existing morphological knowledge. We found that eukaryotic plankton are more diverse than previously thought, especially heterotrophic protists, which may display a wide range of trophic modes (*60*) and include an unsuspected diversity of parasites and photosymbiotic taxa. Dominance of unicellular heterotrophs in plankton ecosystems likely emerged at the dawn of the radiation of eukaryotic cells, together with arguably their most important innovation: phagocytosis. The onset of eukaryophagy in the Neoproterozoic (*61*) probably led to adaptive radiation in heterotrophic eukaryotes through specialization of trophic modes and symbioses, opening novel serial biotic ecological niches. The extensive codiversification of relatively large heterotrophic eukaryotes and their associated parasites supports the idea that biotic interactions, rather than competition for resources and space (*62*), are the primary forces driving organismal diversification in marine plankton systems. Based on rDNA, heterotrophic protists may be even more diverse than prokaryotes in the planktonic ecosystem (*63*). Given that organisms in highly diverse and abundant groups, such as the alveolates and rhizarians, can have genomes more complex than those of humans (*64*), eukaryotic plankton may contain a vast reservoir of unknown marine planktonic genes (*65*). Insights are developing into how heterotrophic protists contribute to a multilayered and integrated ecosystem. The protistan parasites and mutualistic symbionts increase connectivity and complexity of pelagic food webs (*66*, *67*) while contributing to the carbon quota of their larger, longer-lived, and often biomineralized symbiotic hosts, which themselves contribute to carbon export when they die. Decoding the ecological and evolutionary rules governing plankton diversity remains essential for understanding how the

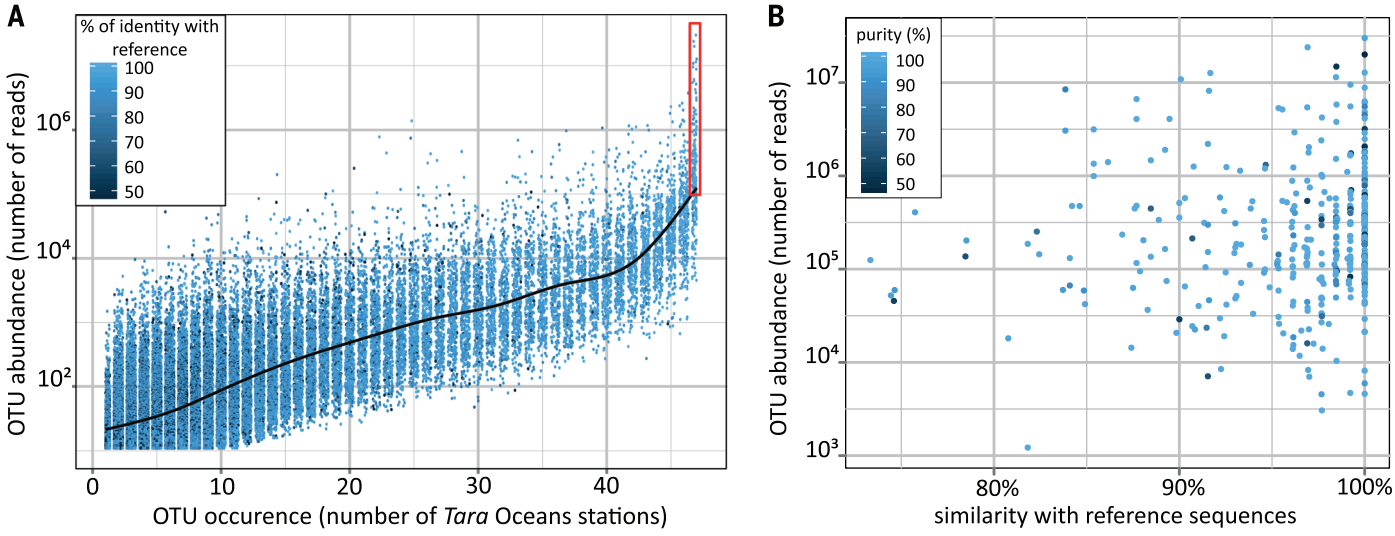


Fig. 7. Cosmopolitanism and abundance of eukaryotic marine plankton. (**A**) Occurrence-versus-abundance plot including the ~110,000 *Tara* Oceans V9 rDNA OTUs. OTUs are colored according to their identity with a reference sequence, and a fitted curve indicates the median OTU size value for each OTU geographic occurrence value. The red rectangle encloses the cosmopolitan and hyperdominant (>10^5 reads) OTUs. (**B**) Similarity to reference barcode and taxonomic purity [a measure of taxonomic assignment consistency defined as the percentage of reads within an OTU assigned to the same taxon; see (*13*)] of the 381 cosmopolitan OTUs, along their abundance (*y* axis).

critical ocean biomes contribute to the functioning of the Earth system.

Materials and methods

V9-18S rDNA for eukaryotic metabarcoding

We used universal eukaryotic primers (*68*) to PCR-amplify (25 cycles in triplicate) the V9-18*S* rDNA genes from all *Tara* Oceans samples. This barcode presents a combination of advantages for addressing general questions of eukaryotic biodiversity over extensive taxonomic and ecological scales: (i) It is universally conserved in length (130 ± 4 base pairs) and simple in secondary structure, thus allowing relatively unbiased PCR amplification across eukaryotic lineages followed by Illumina sequencing. (ii) It includes both stable and highly variable nucleotide positions over evolutionary time frames, allowing discrimination of taxa over a substantial phylogenetic depth. (iii) It is extensively represented in public reference databases across the eukaryotic tree of life, allowing taxonomic assignment among all known eukaryotic lineages (*13*).

Biodiversity analyses

Our bioinformatic pipeline included quality checking (Phred score filtering, elimination of reads without perfect forward and reverse primers, and chimera removal) and conservative filtering (removal of metabarcodes present in less than three reads and two distinct samples). The ~2.3 million metabarcodes (distinct reads) were clustered using an agglomerative, unsupervised single-linkage clustering algorithm, allowing OTUs to reach their natural limits while avoiding arbitrary global clustering thresholds (*13*, *14*). This clustering limited overestimation of biodiversity due to errors in PCR amplification or DNA sequencing, as well as intragenomic polymorphism of rDNA gene copies (*13*). *Tara* Oceans metabarcodes and OTUs were taxonomically assigned by comparison to the 77,449 reference barcodes included in our *V9_PR2* database (*15*). This database derives from the Protist Ribosomal Reference (PR2) database (*69*) but focuses on the V9 region of the gene and includes the following reorganizations: (i) extension of the number of ranks for groups with finer taxonomy (e.g., animals), (ii) expert curation of the taxonomy and renaming in novel environmental groups and dinoflagellates, (iii) resolution of all taxonomic conflicts and inclusion of environmental sequences only if they provide additional phylogenetic information, and (iv) annotation of basic trophic and/or symbiotic modes for all reference barcodes assigned to the genus level [see (*53*) and (*15*) for details]. The *V9_PR2* reference barcodes represent 24,435 species and 13,432 genera from all known major lineages of the tree of eukaryotic life (*15*). Metabarcodes with ≥80% identity to a reference V9 rDNA barcode were considered assignable. Below this threshold it is not possible to discriminate between eukaryotic supergroups, given the short length of V9 rDNA sequences and the relatively fast rate accumulation of substitution mutations in the DNA. In addition to assignment at the finest-possible taxonomic resolution, all assignable metabarcodes were classified into a reference taxonomic framework consisting of 97 major monophyletic groups comprising all known high-rank eukaryotic diversity. This framework, primarily based on a synthesis of protistan biodiversity (*19*), also included all key but still unnamed planktonic clades revealed by previous environmental rDNA clone library surveys (*70*) [e.g., marine alveolates (MALV), marine stramenopiles (MAST), marine ochrophytes (MOCH), and radiolarians (RAD)] (*15*). Details of molecular and bioinformatics methods are available on a companion Web site at http://taraoceans.sb-roscoff.fr/EukDiv/ (*53*). We compiled our data into two databases including the taxonomy, abundance, and size fraction and biogeography information associated with each metabarcode and OTU (*71*).

Ecological inferences

From our *Tara* Oceans metabarcoding data set, we inferred patterns of eukaryotic plankton functional ecology. Based on a literature survey, all reference barcodes assigned to at least the genus level that recruited *Tara* Oceans metabarcodes were associated to basic trophic and symbiotic modes of the organism they come from (*15*) and used for a taxo-functional annotation of our entire metabarcoding data set with the same set of rules used for taxonomic assignation (*53*). False positives were minimized by (i) assigning ecological modes to all individual reference barcodes in *V9_PR2*; (ii) inferring ecological modes to metabarcodes related to monomodal reference barcode(s) (otherwise transferring them to a "NA, nonapplicable" category); and (iii) exploring broad and complex trophic and symbiotic modes that involve fundamental reorganization of the cell structure and metabolism, emerged relatively rarely in the evolutionary history of eukaryotes, and most often concern all known species within monophyletic and ancient groups [see (*15*) for details]. In case of photo- versus heterotrophy, >75% of the major, deep-branching eukaryotic lineages considered (Fig. 3) are monomodal and recruit ~87 and ~69% of all *Tara* Oceans V9 rDNA reads and OTUs, respectively. For parasitism, ~91% of *Tara* Oceans metabarcodes are falling within monophyletic and major groups containing exclusively parasitic species (essentially within the major MALVs groups). Although biases could arise in functional annotation of metabarcodes

relatively distant from reference barcodes in the few complex polymodal groups (e.g., the dinoflagellates that can be phototrophic, heterotrophic, parasitic, or photosymbiotic), a conservative analysis of the trophic and symbiotic ecological patterns presented in Fig. 3, using a ≥99% assignation threshold, shows that these are stable across organismal size fractions and space, independently of the similarity cutoff (80 or 99%), demonstrating their robustness across evolutionary times (*30*).

Note that rDNA gene copy number varies from one to thousands in single eukaryotic genomes (*72*, *73*), precluding direct translation of rDNA read number into abundance of individual organisms. However, the number of rDNA copies per genome correlates positively to the size (*73*) and particularly to the biovolume (*72*) of the eukaryotic cell it represents. We compiled published data from the last ~20 years, confirming the positive correlation between eukaryotic cell size and rDNA copy number across a wide taxonomic and organismal size range [see (*74*); note, however, the ~one order of magnitude of cell size variation for a given rDNA copy number]. To verify whether our molecular ecology protocol preserved this empirical correlation, light microscopy counts of phytoplankton belonging to different eukaryotic supergroups (coccolithophores, diatoms, and dinoflagellates) were performed from nine *Tara* Oceans stations from the Indian, Atlantic, and Southern oceans; transformed into biomass and biovolume data; and then compared with the relative number of V9 rDNA reads found for the identified taxa in the same samples (*74*). Results confirmed the correlation between biovolume and V9 rDNA abundance data ($r^2 = 0.97$, $P = 1 \times 10^{-16}$), although we cannot rule out the possibility that some eukaryotic taxa may not follow the general trend.

REFERENCES AND NOTES

1. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998). doi: 10.1126/science.281.5374.237; pmid: 9657713
2. D. A. Caron, P. D. Countway, A. C. Jones, D. Y. Kim, A. Schnetzer, Marine protistan diversity. *Annu. Rev. Mar. Sci.* **4**, 467–493 (2012). doi: 10.1146/annurev-marine-120709-142802; pmid: 22457984
3. P. López-García, F. Rodríguez-Valera, C. Pedrós-Alió, D. Moreira, Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**, 603–607 (2001). doi: 10.1038/35054537; pmid: 11214316
4. S. Y. Moon-van der Staay, R. De Wachter, D. Vaulot, Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610 (2001). doi: 10.1038/35054541; pmid: 11214317
5. B. Díez, C. Pedrós-Alió, R. Massana, Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl. Environ. Microbiol.* **67**, 2932–2941 (2001). doi: 10.1128/AEM.67.7.2932-2941.2001; pmid: 11425705
6. R. Logares *et al*., Patterns of rare and abundant marine microbial eukaryotes. *Curr. Biol.* **24**, 813–821 (2014). doi: 10.1016/j.cub.2014.02.050; pmid: 24704080
7. V. Edgcomb *et al*., Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J.* **5**, 1344–1356 (2011). doi: 10.1038/ismej.2011.6; pmid: 21390079
8. R. Massana *et al*., Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl. Environ. Microbiol.* **70**, 3528–3534 (2004). doi: 10.1128/AEM.70.6.3528-3534.2004; pmid: 15184153
9. L. Guillou *et al*., Widespread occurrence and genetic diversity of marine parasitoids belonging to *Syndiniales* (*Alveolata*). *Environ. Microbiol.* **10**, 3349–3365 (2008). doi: 10.1111/j.1462-2920.2008.01731.x; pmid: 18771501
10. H. Liu *et al*., Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12803–12808 (2009). doi: 10.1073/pnas.0905841106; pmid: 19622724
11. Companion Web site: Figure W1 and Database W1 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
12. S. Pesant *et al*., Open science resources for the discovery and analysis of *Tara* Oceans data. http://biorxiv.org/content/early/2015/05/08/019117 (2015).
13. Companion Web site: Text W1 and Figure W2 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
14. F. Mahé, T. Rognes, C. Quince, C. de Vargas, M. Dunthorn, Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593 (2014). doi: 10.7717/peerj.593; pmid: 25276506
15. Companion Web site: Database W2, Database W3, and Database W6 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
16. Companion Web site: Text W3, Text W4, Text W5, Figure W4, Figure W5, Figure W6, and Figure W7 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
17. F. W. Preston, The commonness, and rarity, of species. *Ecology* **29**, 254–283 (1948). doi: 10.2307/1930989
18. R. Massana, Eukaryotic picoplankton in surface oceans. *Annu. Rev. Microbiol.* **65**, 91–110 (2011). doi: 10.1146/annurev-micro-090110-102903; pmid: 21639789
19. J. Pawlowski *et al*., CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLOS Biol.* **10**, e1001419 (2012). doi: 10.1371/journal.pbio.1001419; pmid: 23139639
20. C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, B. Worm, How many species are there on Earth and in the ocean? *PLOS Biol.* **9**, e1001127 (2011). doi: 10.1371/journal.pbio.1001127; pmid: 21886479
21. W. Appeltans *et al*., The magnitude of global marine species diversity. *Curr. Biol.* **22**, 2189–2202 (2012). doi: 10.1016/j.cub.2012.09.036; pmid: 23159596
22. L. M. Márquez, D. J. Miller, J. B. MacKenzie, M. J. H. Van Oppen, Pseudogenes contribute to the extreme diversity of nuclear ribosomal DNA in the hard coral *Acropora*. *Mol. Biol. Evol.* **20**, 1077–1086 (2003). doi: 10.1093/molbev/msg122; pmid: 12777522
23. S. R. Santos, R. A. Kinzie III, K. Sakai, M. A. Coffroth, Molecular characterization of nuclear small subunit (18S)-rDNA pseudogenes in a symbiotic dinoflagellate (*Symbiodinium*, Dinophyta). *J. Eukaryot. Microbiol.* **50**, 417–421 (2003). doi: 10.1111/j.1550-7408.2003.tb00264.x; pmid: 14733432
24. J. Decelle, S. Romac, E. Sasaki, F. Not, F. Mahé, Intracellular diversity of the V4 and V9 regions of the 18S rRNA in marine protists (radiolarians) assessed by high-throughput sequencing. *PLOS ONE* **9**, e104297 (2014). doi: 10.1371/journal.pone.0104297; pmid: 25090095
25. A. Sournia, M.-J. Chrétiennot-Dinet, M. Ricard, Marine phytoplankton: How many species in the world ocean? *J. Plankton Res.* **13**, 1093–1099 (1991). doi: 10.1093/plankt/13.5.1093
26. P. H. Wiebe *et al*., Deep-sea sampling on CMarZ cruises in the Atlantic Ocean – An introduction. *Deep-Sea Res. Part II* **57**, 2157–2166 (2010). doi: 10.1016/j.dsr2.2010.09.018
27. D. Boltovskoy, Diversity and endemism in cold waters of the South Atlantic: Contrasting patterns in the plankton and the benthos. *Sci. Mar.* **69**, 17–26 (2005).
28. C. de Vargas, R. Norris, L. Zaninetti, S. W. Gibb, J. Pawlowski, Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 2864–2868 (1999). doi: 10.1073/pnas.96.6.2864; pmid: 10077602
29. K. M. K. Halbert, E. Goetze, D. B. Carlon, High cryptic diversity across the global range of the migratory planktonic copepods *Pleuromamma piseki* and *P. gracilis*. *PLOS ONE* **8**, e77011 (2013). doi: 10.1371/journal.pone.0077011; pmid: 24167556
30. Companion Web site: Figure W8, Figure W9, Figure W10, and Figure W14 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
31. F. Burki, P. J. Keeling, Rhizaria. *Curr. Biol.* **24**, R103–R107 (2014). doi: 10.1016/j.cub.2013.12.025; pmid: 24502779
32. N. R. Swanberg, thesis, Massachussetts Institute of Technology (1974).
33. I. Probert *et al*., *Brandtodinium* gen. nov. and *B. nutricula* comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. *J. Phycol.* **50**, 388–399 (2014). doi: 10.1111/jpy.12174
34. S. von der Heyden, E. E. Chao, K. Vickerman, T. Cavalier-Smith, Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of euglenozoa. *J. Eukaryot. Microbiol.* **51**, 402–416 (2004). doi: 10.1111/j.1550-7408.2004.tb00387.x; pmid: 15352322
35. E. Schnepf, Light and electron microscopical observations in *Rhynchopus coscinodiscivorus* spec. nov., a Colorless, phagotrophic euglenozoon with concealed flagella. *Arch. Protistenkd.* **144**, 63–74 (1994). doi: 10.1016/S0003-9365(11)80225-3
36. M. R. Dennett, Video plankton recorder reveals high abundances of colonial Radiolaria in surface waters of the central North Pacific. *J. Plankton Res.* **24**, 797–805 (2002). doi: 10.1093/plankt/24.8.797
37. L. Stemmann *et al*., Global zoogeography of fragile macrozooplankton in the upper 100-1000 m inferred from the underwater video profiler. *ICES J. Mar. Sci.* **65**, 433–442 (2008). doi: 10.1093/icesjms/fsn010
38. A. F. Michaels, D. A. Caron, N. R. Swanberg, F. A. Howse, C. M. Michaels, Planktonic sarcodines (Acantharia, Radiolaria, Foraminifera) in surface waters near Bermuda: Abundance, biomass and vertical flux. *J. Plankton Res.* **17**, 131–163 (1995). doi: 10.1093/plankt/17.1.131
39. E. Haeckel, "Report on the Radiolaria collected by H.M.S. Challenger during the years 1873-1876" in *Report on the Scientific Results of the Voyage of* H.M.S. Challenger *During the Years 1873-76. Zoology.* (Neill, Edinburgh, 1887).
40. R. Siano, M. Montresor, I. Probert, F. Not, C. de Vargas, *Pelagodinium* gen. nov. and *P. béii* comb. nov., a dinoflagellate symbiont of planktonic foraminifera. *Protist* **161**, 385–399 (2010). doi: 10.1016/j.protis.2010.01.002; pmid: 20149979
41. J. Decelle *et al*., An original mode of symbiosis in open ocean plankton. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 18000–18005 (2012). doi: 10.1073/pnas.1212303109; pmid: 23071304
42. Y. Shaked, C. de Vargas, Pelagic photosymbiosis: rDNA assessment of diversity and evolution of dinoflagellate symbionts and planktonic foraminiferal hosts. *Mar. Ecol. Prog. Ser.* **325**, 59–71 (2006). doi: 10.3354/meps325059
43. J. Decelle, New perspectives on the functioning and evolution of photosymbiosis in plankton: Mutualism or parasitism? *Commun. Integr. Biol.* **6**, e24560 (2013). doi: 10.4161/cib.24560; pmid: 23986805
44. R. Siano *et al*., Distribution and host diversity of Amoebophryidae parasites across oligotrophic waters of the Mediterranean Sea. *Biogeosciences* **8**, 267–278 (2011). doi: 10.5194/bg-8-267-2011
45. D. Coats, M. Park, Parasitism of photosynthetic dinoflagellates by three strains of *Amoebophyra* (Dinophyta): Parasite survival, infectivity, generation time, and host specificity. *J. Phycol.* **528**, 520–528 (2002). doi: 10.1046/j.1529-8817.2002.01200.x
46. K. E. Wommack, R. R. Colwell, Virioplankton: Viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000). doi: 10.1128/MMBR.64.1.69-114.2000; pmid: 10704475
47. A. Skovgaard, Dirty tricks in the plankton: Diversity and role of marine parasitic protists. *Acta Protozool.* **53**, 51–62 (2014).
48. J. Bråte *et al*., Radiolaria associated with large diversity of marine alveolates. *Protist* **163**, 767–777 (2012). doi: 10.1016/j.protis.2012.04.004; pmid: 22658831
49. T. R. Bachvaroff, S. Kim, L. Guillou, C. F. Delwiche, D. W. Coats, Molecular diversity of the syndinean genus *Eudubosquella* based on single-cell PCR analysis. *Appl. Environ. Microbiol.* **78**, 334–345 (2012). doi: 10.1128/AEM.06678-11; pmid: 22081578
50. S. Rueckert, T. G. Simdyanov, V. V. Aleoshin, B. S. Leander, Identification of a divergent environmental DNA sequence clade using the phylogeny of gregarine parasites (Apicomplexa) from crustacean hosts. *PLOS ONE* **6**, e18163 (2011). doi: 10.1371/journal.pone.0018163; pmid: 21483868
51. A. Skovgaard, S. A. Karpov, L. Guillou, The parasitic dinoflagellates *Blastodinium* spp. inhabiting the gut of marine, planktonic copepods: Morphology, ecology, and unrecognized species diversity. *Front. Microbiol.* **3**, 305 (2012). doi: 10.3389/fmicb.2012.00305; pmid: 22973263
52. H. McCallum *et al*., Does terrestrial epidemiology apply to marine systems? *Trends Ecol. Evol.* **19**, 585–591 (2004). doi: 10.1016/j.tree.2004.08.009

53. Companion Web site: detailed Material and Methods, Database W9, and Figure W11 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
54. Companion Web site: Figure W12, Figure W13, Database W7, and Database W8 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
55. M. Holyoak, M. A. Leibold, R. D. Holt, Eds., *Metacommunities: Spatial Dynamics and Ecological Communities* (University of Chicago Press, Chicago, 2005).
56. L. G. M. Baas Becking, *Geobiologie of Inleiding tot de Milieukunde* (W. P. Van Stockum and Zoon, The Hague, Netherlands, 1934).
57. K. T. C. A. Peijnenburg, E. Goetze, High evolutionary potential of marine zooplankton. *Ecol. Evol.* **3**, 2765–2781 (2013). doi: 10.1002/ece3.644; pmid: 24567838
58. V. Smetacek, Microbial food webs. The ocean's veil. *Nature* **419**, 565 (2002). doi: 10.1038/419565a; pmid: 12374956
59. H. ter Steege *et al*., Hyperdominance in the Amazonian tree flora. *Science* **342**, 1243092 (2013). doi: 10.1126/science.1243092; pmid: 24136971
60. D. Vaulot, K. Romari, F. Not, Are autotrophs less diverse than heterotrophs in marine picoplankton? **10**, 266–267 (2002).
61. A. H. Knoll, Paleobiological perspectives on early eukaryotic evolution. *Cold Spring Harb. Perspect. Biol.* **6**, 1–14 (2014). doi: 10.1101/cshperspect.a016121; pmid: 24384569
62. V. Smetacek, A watery arms race. *Nature* **411**, 745 (2001). doi: 10.1038/35081210; pmid: 11459035
63. S. Sunagawa *et al*., Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
64. M. J. Oliver, D. Petrov, D. Ackerly, P. Falkowski, O. M. Schofield, The mode and tempo of genome size evolution in eukaryotes. *Genome Res.* **17**, 594–601 (2007). doi: 10.1101/gr.6096207; pmid: 17420184
65. H. Abida *et al*., Bioprospecting marine plankton. *Mar. Drugs* **11**, 4594–4611 (2013). doi: 10.3390/md11114594; pmid: 24240981
66. G. Lima-Mendez *et al*., Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).
67. K. D. Lafferty, A. P. Dobson, A. M. Kuris, Parasites dominate food web links. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11211–11216 (2006). doi: 10.1073/pnas.0604755103; pmid: 16844774
68. L. A. Amaral-Zettler, E. A. McCliment, H. W. Ducklow, S. M. Huse, A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLOS ONE* **4**, e6372 (2009). doi: 10.1371/journal.pone.0006372; pmid: 19633714
69. L. Guillou *et al*., The Protist Ribosomal Reference database (PR^2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, D597–D604 (2013). doi: 10.1093/nar/gks1160; pmid: 23193267
70. R. Massana, J. del Campo, M. E. Sieracki, S. Audic, R. Logares, Exploring the uncultured microeukaryote majority in the oceans: Reevaluation of ribogroups within stramenopiles. *ISME J.* **8**, 854–866 (2014). doi: 10.1038/ismej.2013.204; pmid: 24196325
71. Companion Web site: Database W4 and Database W5 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
72. A. Godhe *et al*., Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl. Environ. Microbiol.* **74**, 7174–7182 (2008). doi: 10.1128/AEM.01298-08; pmid: 18849462
73. F. Zhu, R. Massana, F. Not, D. Marie, D. Vaulot, Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol. Ecol.* **52**, 79–92 (2005). doi: 10.1016/j.femsec.2004.10.006; pmid: 16329895
74. Companion Web site: Text W2 and Figure W3 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
75. M. Kim, S. Nam, W. Shin, D. W. Coats, M. Park, *Dinophysis caudata* (dinophyceae) sequesters and retains plastids from the mixotrophic ciliate prey *Mesodinium rubrum*. *J. Phycol.* **48**, 569–579 (2012). doi: 10.1111/j.1529-8817.2012.01150.x

ACKNOWLEDGMENTS

We thank the following people and sponsors for their commitment: CNRS (in particular, the GDR3280); EMBL; Genoscope/CEA; UPMC; VIB; Stazione Zoologica Anton Dohrn; UNIMIB; Rega Institute; KU Leuven; Fund for Scientific Research – The French Ministry of Research, the French Government "Investissements d'Avenir" programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), and MEMO LIFE (ANR-10-LABX-54); PSL* Research University (ANR-11-IDEX-0001-02); ANR (projects POSEIDON/ANR-09-BLAN-0348, PROMETHEUS/ANR-09-PCS-GENM-217, PHYTBACK/ANR-2010-1709-01, and TARA-GIRUS/ANR-09-PCS-GENM-218); EU FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376); European Research Council Advanced Grant Awards to C. Bowler (Diatomite:294823); Gordon and Betty Moore Foundation grant 3790 to M.B.S.; Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS and TANIT (CONES 2010-0036) grant from the Agency for Administration of University and Research Grants (AGAUR) to S.G.A.; and Japan Society for the Promotion of Science KAKENHI grant 26430184 to H.O. We also thank the following for their support and commitment: A. Bourgois, E. Bourgois, R. Troublé, Région Bretagne, G. Ricono, the Veolia Environment Foundation, Lorient Agglomération, World Courier, Illumina, the Electricité de France Foundation, Fondation pour la Recherche sur la Biodiversité, the Prince Albert II de Monaco Foundation, and the *Tara* schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge assistance from European Bioinformatics Institute (EBI) (in particular, G. Cochrane and P. ten Hoopen) as well as the EMBL Advanced Light Microscopy Facility (in particular, R. Pepperkok). We thank F. Gaill, B. Kloareg, F. Lallier, D. Boltovskoy, A. Knoll, D. Richter, and E. Médard for help and advice on the manuscript. We declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. Data described herein are available at http://taraoceans.sb-roscoff.fr/EukDiv/, at EBI under the project IDs PRJEB402 and PRJEB6610, and at PANGAEA (see table S1). The data release policy regarding future public release of *Tara* Oceans data is described in (*12*). All authors approved the final manuscript. This article is contribution number 24 of *Tara* Oceans. The supplementary materials contain additional data.

Tara Oceans Coordinators

Silvia G. Acinas,[1] Peer Bork,[2] Emmanuel Boss,[3] Chris Bowler,[4] Colomban de Vargas,[5,6] Michael Follows,[7] Gabriel Gorsky,[8,9] Nigel Grimsley,[10,11] Pascal Hingamp,[12] Daniele Iudicone,[13] Olivier Jaillon,[14,15,16] Stefanie Kandels-Lewis,[2,17] Lee Karp-Boss,[3] Eric Karsenti,[4,17] Uros Krzic,[18] Fabrice Not,[5,6] Hiroyuki Ogata,[19] Stephane Pesant,[20,21] Jeroen Raes,[22,23,24] Emmanuel G. Reynaud,[25] Christian Sardet,[26,27] Mike Sieracki,[28] Sabrina Speich,[29,30] Lars Stemmann,[8] Matthew B. Sullivan,[31]* Shinichi Sunagawa,[2] Didier Velayoudon,[32] Jean Weissenbach,[14,15,16] Patrick Wincker[14,15,16]

[1]Department of Marine Biology and Oceanography, ICM-CSIC, Passeig Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [2]Structural and Computational Biology, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [3]School of Marine Sciences, University of Maine, Orono, ME 04469, USA. [4]Ecole Normale Supérieure, IBENS, and Inserm U1024, and CNRS UMR 8197, Paris, F-75005 France. [5]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [6]Sorbonne Universités, UPMC Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [8]CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [9]Sorbonne Universités, UPMC Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [10]CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [11]Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [12]Aix Marseille Université, CNRS IGS, UMR 7256, 13288 Marseille, France. [13]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [14]CEA, Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [15]CNRS, UMR 8030, CP5706 Evry, France. [16]Université d'Evry, UMR 8030, CP5706 Evry, France. [17]Directors' Research, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [18]Cell Biology and Biophysics, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [19]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. [20]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [21]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [22]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [23]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [24]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [25]Earth Institute, University College Dublin, Dublin, Ireland. [26]CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [27]Sorbonne Universités, UPMC Univ Paris 06, UMR 7009 Biodev, F-06230 Observatoire Océanologique, Villefranche-sur-Mer, France. [28]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA. [29]Department of Geosciences, LMD, Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris, Cedex 05, France. [30]Laboratoire de Physique des Océan UBO-IUEM Place Copernic 29820 Plouzané, France. [31]Department of Ecology and Evolutionary Biology, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA. [32]DVIP Consulting, Sèvres, France.
*Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/1261605/suppl/DC1
Table S1
Appendix S1

23 September 2014; accepted 27 February 2015
10.1126/science.1261605

OCEAN PLANKTON

Determinants of community structure in the global plankton interactome

Gipsi Lima-Mendez,[1,2,3]* Karoline Faust,[1,2,3]* Nicolas Henry,[4,5]* Johan Decelle,[4,5] Sébastien Colin,[4,5,6] Fabrizio Carcillo,[1,2,3,7] Samuel Chaffron,[1,2,3] J. Cesar Ignacio-Espinosa,[8]† Simon Roux,[8]† Flora Vincent,[2,6] Lucie Bittner,[4,5,6,9] Youssef Darzi,[2,3] Jun Wang,[1,2] Stéphane Audic,[4,5] Léo Berline,[10,11] Gianluca Bontempi,[7] Ana M. Cabello,[12] Laurent Coppola,[10,11] Francisco M. Cornejo-Castillo,[12] Francesco d'Ovidio,[13] Luc De Meester,[14] Isabel Ferrera,[12] Marie-José Garet-Delmas,[4,5] Lionel Guidi,[10,11] Elena Lara,[12] Stéphane Pesant,[15,16] Marta Royo-Llonch,[12] Guillem Salazar,[12] Pablo Sánchez,[12] Marta Sebastian,[12] Caroline Souffreau,[14] Céline Dimier,[4,5,6] Marc Picheral,[10,11] Sarah Searson,[10,11] Stefanie Kandels-Lewis,[17,18] Tara Oceans coordinators‡ Gabriel Gorsky,[10,11] Fabrice Not,[4,5] Hiroyuki Ogata,[19] Sabrina Speich,[20,21] Lars Stemmann,[10,11] Jean Weissenbach,[22,23,24] Patrick Wincker,[22,23,24] Silvia G. Acinas,[12] Shinichi Sunagawa,[17] Peer Bork,[17,25] Matthew B. Sullivan,[8]† Eric Karsenti,[6,18]§ Chris Bowler,[6]§ Colomban de Vargas,[4,5]§ Jeroen Raes[1,2,3]§

Species interaction networks are shaped by abiotic and biotic factors. Here, as part of the *Tara* Oceans project, we studied the photic zone interactome using environmental factors and organismal abundance profiles and found that environmental factors are incomplete predictors of community structure. We found associations across plankton functional types and phylogenetic groups to be nonrandomly distributed on the network and driven by both local and global patterns. We identified interactions among grazers, primary producers, viruses, and (mainly parasitic) symbionts and validated network-generated hypotheses using microscopy to confirm symbiotic relationships. We have thus provided a resource to support further research on ocean food webs and integrating biological components into ocean models.

The structure of oceanic ecosystems results from the complex interplay between resident organisms and their environment. In the world's largest ecosystem, oceanic plankton (composed of viruses, prokaryotes, microbial eukaryotes, phytoplankton, and zooplankton) form trophic and symbiotic interaction networks (*1*–*4*) that are influenced by environmental conditions. Ecosystem structure and composition are governed by abiotic as well as biotic factors. The former include environmental conditions and nutrient availability (*5*), whereas the latter include grazing, pathogenicity, and parasitism (*6*, *7*). Historically, abiotic factors have been considered to have a stronger effect, but recently, appreciation for biotic factors is growing (*8*, *9*). We sought to develop a quantitative understanding of biotic and abiotic interactions in natural systems in which the organisms are taxonomically and trophically diverse (*10*). We used sequencing technologies to profile communities across trophic levels, organismal sizes, and geographic ranges and to predict organismal interactions across biomes based on co-occurrence patterns (*11*). Previous efforts addressing these issues have provided insights on the structure (*12*, *13*) and dynamics of microbial communities (*14*–*16*).

We analyzed data from 313 plankton samples the *Tara* Oceans expedition (*17*) derived from seven size-fractions covering collectively 68 stations at two depths across eight oceanic provinces (table S1). The plankton samples spanned sizes that include organisms from viruses to small metazoans. We derived viral, prokaryotic, and eukaryotic abundance profiles from clusters of metagenomic contigs, Illumina-sequenced metagenomes ($_{mi}$tags), and 18*S* ribosomal DNA (rDNA) V9 sequences, respectively (table S1) (*10*, *18*, *19*) and collected environmental data from on-site and satellite measurements (*17*, *20*, *21*). We used network inference methods and machine-learning techniques so as to disentangle biotic and abiotic signals shaping ocean plankton communities and to construct an interactome that described the network of interactions among photic zone plankton groups. We used the interactome to focus on specific relationships, which we validated through microscopic analysis of symbiont pairs and in silico analysis of phage-host pairings.

Evaluating the effect of abiotic and biotic factors on community structure

We first reassessed the effects of environment and geography on community structure. Using variation partitioning (*22*), we found that on average, the percentage of variation in community composition explained by environment alone was 18%, by environment combined with geography 13%, and by geography alone only 3% (*23*, *24*). In addition, we built random forest-based models (*25*) in order to predict abundance profiles of the Operational Taxonomic Units (OTU) using (i) OTUs alone, (ii) environmental variables alone, and (iii) OTUs and environmental variables combined and tested for each OTU whether one of the three approaches outcompeted the other. These analyses revealed that 95% of the OTU-only models are more accurate in predicting OTU abundances than environmental variable models, and that combined models were no better than the OTU-only models (*26*, *27*). This suggests that abiotic factors have a more limited effect on community structure than previously assumed (*8*).

To study the role of biotic interactions, we developed a method with which to identify robust species associations in the context of environmental conditions. Twenty-three taxon-taxon and taxon-environment co-occurrence networks were constructed based on 9292 taxa, representing the combinations of two depths, seven organismal

[1]Department of Microbiology and Immunology, Rega Institute KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [2]VIB Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [3]Department of Applied Biological Sciences (DBIT), Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [4]Station Biologique de Roscoff, CNRS, UMR 7144, Place Georges Teissier, 29680 Roscoff, France. [5]Sorbonne Universités, Université Pierre et Marie Curie (UPMC) Université Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [6]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), Inserm U1024, CNRS UMR 8197, Paris, F-75005 France. [7]Interuniversity Institute of Bioinformatics in Brussels (IB)^2, ULB Machine Learning Group, Computer Science Department, Université Libre de Bruxelles (ULB), Brussels, Belgium. [8]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721, USA. [9]Institut de Biologie Paris-Seine, CNRS FR3631, F-75005, Paris, France. [10]CNRS, UMR 7093, Laboratoire d'Océanographie de Villefranche (LOV), Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [11]Sorbonne Universités, UPMC Paris 06, UMR 7093, Laboratoire d'Océanographie de Villefranche (LOV), Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [12]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)–Consejo Superior de Investigaciones Científicas (CSIC), Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [13]Sorbonne Universités, UPMC, Université Paris 06, CNRS–Institut pour la Recherche et le Développement–Muséum National d'Histoire Naturelle, Laboratoire d'Océanographie et du Climat: Expérimentations et Approches Numériques (LOCEAN) Laboratory, 4 Place Jussieu, 75005, Paris, France. [14]KU Leuven, Laboratory of Aquatic Ecology, Evolution and Conservation, Charles Deberiotstraat 32, 3000 Leuven. [15]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Hochschulring 18, 28359 Bremen, Germany. [16]MARUM, Center for Marine Environmental Sciences, University of Bremen, Hochschulring 18, 28359 Bremen, Germany. [17]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [18]Directors' Research, European Molecular Biology Laboratory, Heidelberg, Germany. [19]Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011 Kyoto, Japan. [20]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France. [21]Laboratoire de Physique des Océan, Université de Bretagne Occidentale (UBO)–Institut Universaire Européen de la Mer (IUEM), Palce Copernic, 29820 Polouzané, France. [22]Commissariat à l'Énergie Atomique (CEA), Genoscope, 2 rue Gaston Crémieux, 91000 Evry, France. [23]CNRS, UMR 8030, 2 rue Gaston Crémieux, 91000 Evry, France. [24]Université d'Evry, UMR 8030, CP5706 Evry, France. [25]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany.

*These authors contributed equally to this work. †Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA. ‡*Tara* Oceans coordinators and affiliations are listed at the end of this manuscript. **§Corresponding author. E-mail: jeroen.raes@vib-kuleuven.be (J.R.); vargas@sb-roscoff.fr (C.d.V.); cbowler@biologie.ens.fr (C.B.); karsenti@embl.de (E.K.)**

size ranges, and four organismal domains (Bacteria, Archaea, Eukarya, and viruses) (28). To reduce noise and thus false-positive predictions, we restricted our analysis to taxa present in at least 20% of the samples and used conservative statistical cutoffs. We merged the individual networks into a global network, which features a total of 127,995 distinct edges, of which 92,633 are taxon-taxon edges and 35,362 are taxon-environment edges (Table 1). Node degree does not depend on the abundance of the node (28). As such, this network represents a resource with which to examine species associations in the global oceans (28–31).

Next, we assessed how many of the taxon links represented "niche effects" driven by geography or environment (such as when taxa respond similarly to a common environmental condition). We examined motifs consisting of two correlated taxa that also correlate with at least one common environmental parameter ("environmental triplets" to identify associations that were driven by environment) using three approaches [interaction information, sign pattern analysis, and network deconvolution (32)]. We identified 29,912 taxon-taxon-environment associations (32.3% of total). Among environmental factors, we found that PO_4, temperature, NO_2, and mixed-layer depth were frequent drivers of network connections (Fig. 1A). Although the three methodologies pinpoint indirect associations, only interaction information directly identifies synergistic effects in these biotic-abiotic triplets. Exploiting this property, we disentangled the 29,912 environment-affected associations into 11,043 edges driven solely by abiotic factors (excluded from the network for the remainder of the study) (31, 33) and 18,869 edges whose dependencies result from biotic-abiotic synergistic effects. Thus, we find that a minority of associations can be explained by an environmental factor.

Evaluation of predicted interactions

Co-occurrence techniques have heretofore mainly been applied to bacteria. We detected eukaryotic interactions on the basis of analysis of sequences at the V9 hypervariable region of the 18*S* ribosomal RNA (rRNA) gene. We built a literature-curated collection (34) of 574 known symbiotic interactions (including both parasitism and mutualism) in marine eukaryotic plankton (30, 35). From 43 genus-level interactions represented by OTUs in the abundance preprocessed input matrices, we found 42% (18 genus pairs; 47% when limiting to parasitic interactions) represented in our reference list. The probability of having found each of these interactions by chance alone was <0.01 (Fisher exact test, average $P = 4^{-3}$, median $P = 5e^{-7}$). On the basis of this sensitivity and a false discovery rate averaging to 9% (computed from null models), we estimate the number of interactions among eukaryotes present in our filtered input matrices to be between 53,000 and 139,000. Most of the false-negative interactions were due to the strict filtering rules we used to avoid false positives; this hampers detection when, for example, interactions are facultative or when interaction partners may vary among closely related groups depending on oceanic region (4). False positives could represent indirect interactions between species (bystander effects) or environmental effects caused by factors not captured in this study (36, 37).

Biotic interactions within and across kingdoms

The integrated network contained 81,590 predicted biotic interactions (30) that were nonrandomly distributed within and between size fractions (Fig. 1, B and C) (38). Positive associations outnumbered mutual exclusions (72% versus 28%), and we observed a nonrandom edge distribution with regard to phylogeny (Fig. 2A), with most associations derived from syndiniales and other dinoflagellates (examples are shown in

Table 1. Properties of the merged taxon network. The positive subset of the network was clustered with the leading eigen vector algorithm (91).

Nodes	Edges	Positive edges (%)	Negative edges	Average clustering coefficient	Average path length	Diameter	Average betweenness	Modularity of positive network	Number of modules in positive network
9169	92,633	68,856 (74.33)	23,777	0.229	3.43	12	11024	0.51	51

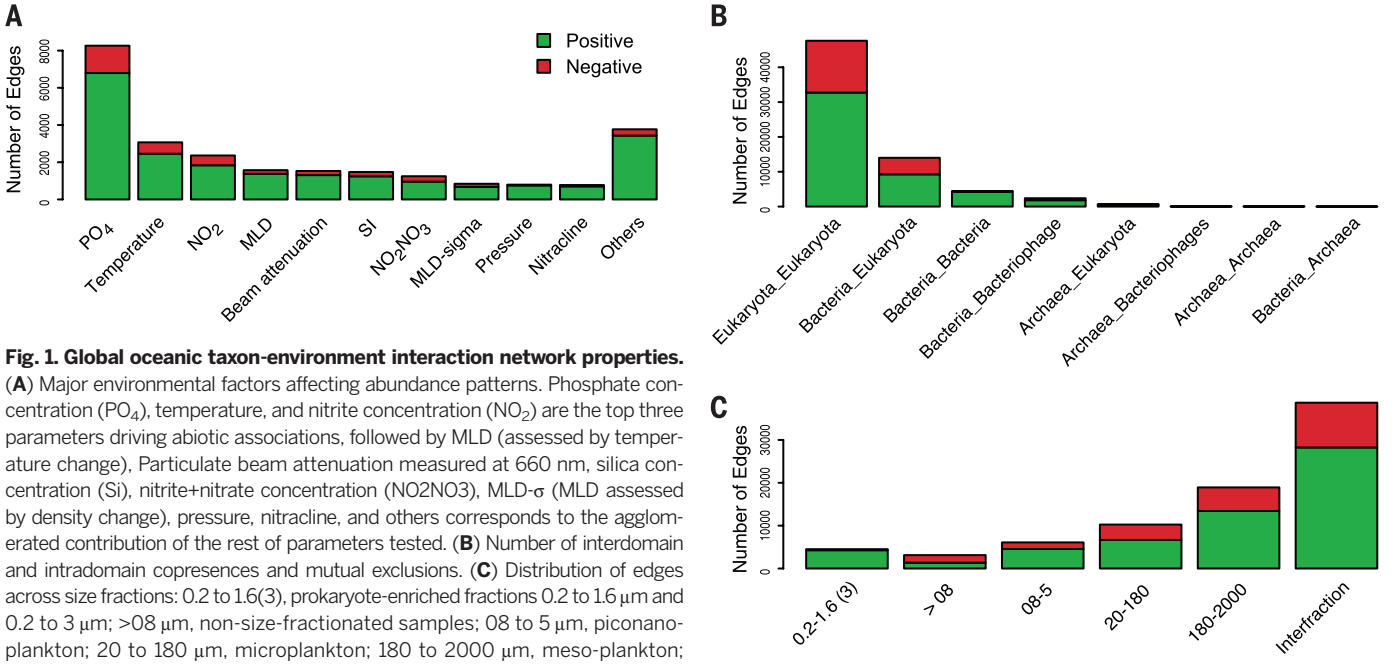


Fig. 1. Global oceanic taxon-environment interaction network properties. (**A**) Major environmental factors affecting abundance patterns. Phosphate concentration (PO_4), temperature, and nitrite concentration (NO_2) are the top three parameters driving abiotic associations, followed by MLD (assessed by temperature change), Particulate beam attenuation measured at 660 nm, silica concentration (Si), nitrite+nitrate concentration (NO2NO3), MLD-σ (MLD assessed by density change), pressure, nitracline, and others corresponds to the agglomerated contribution of the rest of parameters tested. (**B**) Number of interdomain and intradomain copresences and mutual exclusions. (**C**) Distribution of edges across size fractions: 0.2 to 1.6(3), prokaryote-enriched fractions 0.2 to 1.6 μm and 0.2 to 3 μm; >08 μm, non-size-fractionated samples; 08 to 5 μm, piconanoplankton; 20 to 180 μm, microplankton; 180 to 2000 μm, meso-plankton; interfrac, includes interfraction networks 08 to 5 μm versus 20 to 180 μm, 08 to 5 μm versus 180 to 2000 μm, 20 to 180 μm versus 180 to 2000 μm, and 0.2 to 1.6(3) μm versus ≤ 0.2 μm (virus-enriched fraction).

Fig. 3A), and exclusions involving arthropods. Certain combinations of phylogenetic groups are overrepresented (*39*). For instance, we found a clade of syndiniales [the MALV-II Clade 1 belonging to *Amoebophrya* (*3*)] enriched in positive associations with tintinnids ($P = 2^{-4}$), which are among the most abundant ciliates in marine plankton (*40*). The tintinnid *Xystonella lohmani* was described in 1964 to be infected by *Amoebophrya tintinnis* (*41*), and tintinnids can feed on *Amoebophrya* free-living stages (*42*). Other found host-parasite associations included the copepod parasites *Blastodinium*, *Ellobiopsis*, and *Vampyrophrya* (*41*, *43*–*45*).

On the other hand, *Maxillopoda*, *Bacillariophyceae*, and collodarians, three groups of relatively large sized organisms whose biomass can dominate planktonic ecosystems, are rich in negative associations among them (*33*). Collodarians and copepods are abundant in, respectively, the oligotrophic tropical and eutrophic and mesotrophic temperate systems (*10*, *46*). The decoupling of phyto- and zooplankton in open oceans by diatoms anticorrelating to copepods (*47*, *48*) is attributed to growth rate differences and to the diatom production of compounds harmful to their grazers (*49*). The combination of these

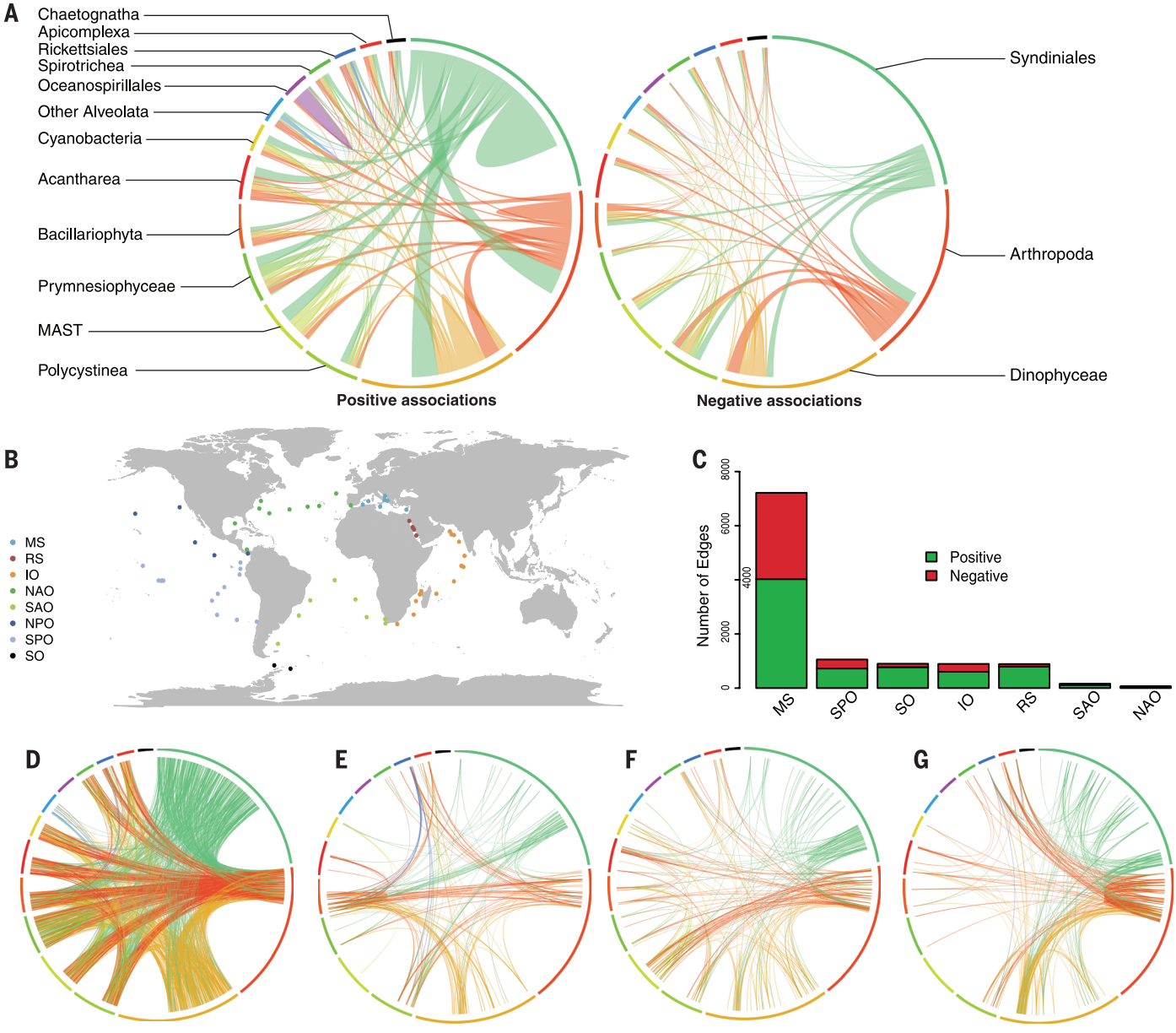


Fig. 2. Taxonomic and geographic patterns within the co-occurrence network. (**A**) Top 15 interacting taxon groups depicted as colored segments in a CIRCOS plot, in which ribbons connecting two segments indicate co-presence and exclusion links, on the left and right, respectively. Size of the ribbon is proportional to the number of links (copresences and exclusions) between the OTUs assigned to the respective segments, and color is segment (of the two involved) with the more total links. Links are dominated by the obligate parasites syndiniales and by Arthropoda and Dinophyceae. (**B**) *Tara* Oceans sampling stations grouped by oceanic provinces. (**C**) Frequency of local co-occurrence patterns across the oceanic provinces, showing that most local patterns are located in MS. (**D** to **G**) Taxonomic patterns of co-occurrences across MS (D), SPO (E), IO (F), and RS (G). Edges are represented as ribbons between barcodes grouped into their taxonomic order as in (A). Links sharing the same segment are affiliated to the same taxon (Order), showing that the connectivity patterns across taxa are conserved at high taxonomic ranks. The local specificity of interactions at higher resolution (OTUs) is apparent by thin ribbons (edge resolution), with different starts, and end positions (different OTUs) within the shared (taxon) segment, section color, and ordering correspond to those in (A). SO-specific associations are mainly driven by bacterial interactions (*53*).

effects could lie at the basis of this observation, which contrasts with other free-living autotrophs represented in the network (cyanobacteria and prymnesiophytes), which display primarily positive associations (Fig. 2A).

Cross-kingdom associations between Bacteria and Archaea were limited to 24 mutual exclusions. Within Archaea, Thermoplasmatales (Marine Group II) co-occur with several phytoplankton clades. Links between Bacteria and protists recovered five out of eight recently discovered interactions from protist single-cell sequencing (*50*). Associations between Diatoms and Flavobacteria agreed with their described symbioses (*51*). We also observed co-ocurrence of uncultured dinoflagellates with members of Rhodobacterales (*Ruegeria*), which is in agreement with a symbiosis between *Ruegeria* sp. TM1040 and *Pfiesteria piscicida* around the ability of *Ruegeria* to metabolize dinoflaggelate-produced dimethylsulfoniopropionate (*52*).

Global versus local associations

We further investigated whether our network was driven by global trends or is defined by local signals. To this aim, we divided our set of samples into seven main regions—Mediterranean

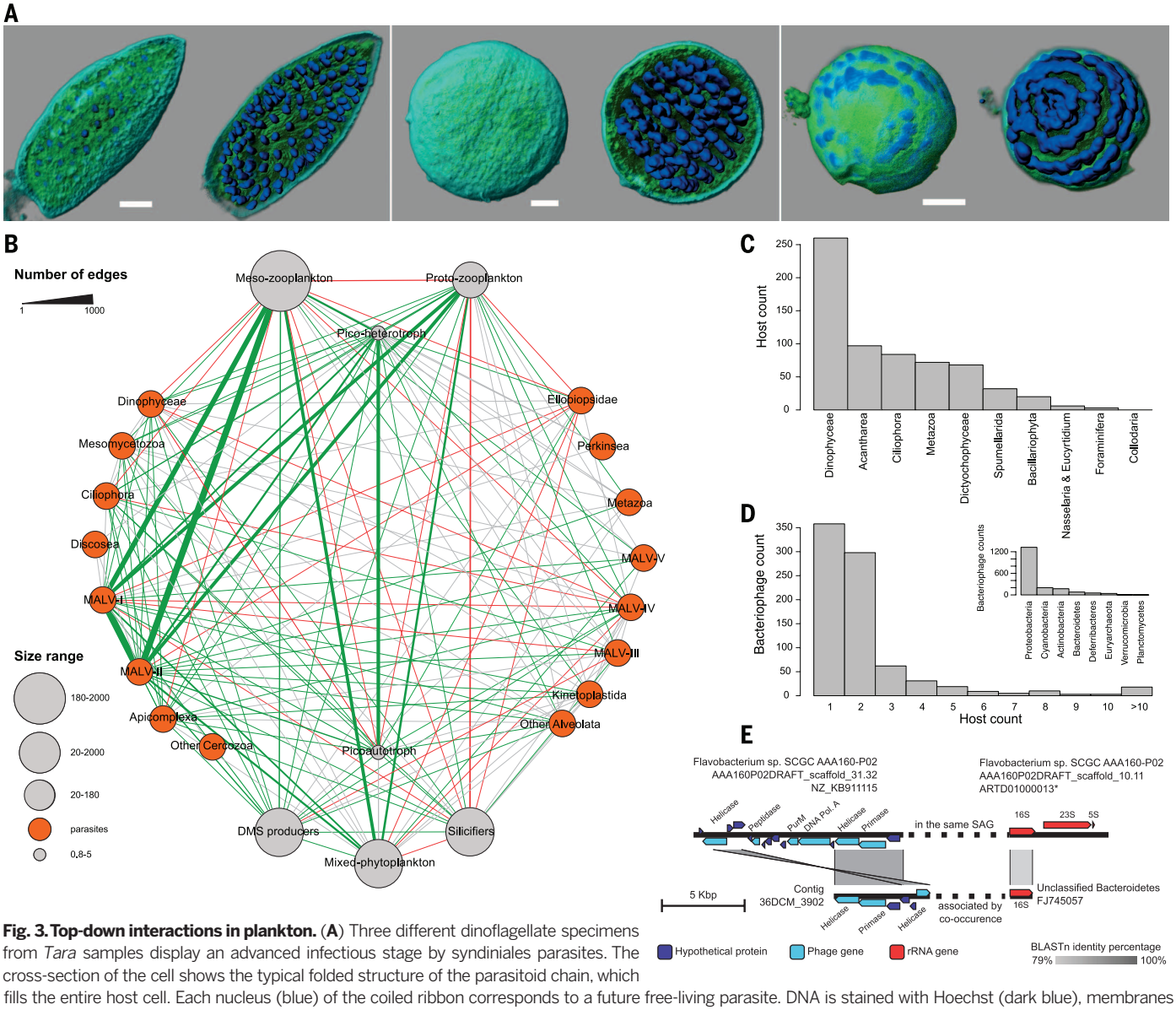


Fig. 3. Top-down interactions in plankton. (**A**) Three different dinoflagellate specimens from *Tara* samples display an advanced infectious stage by syndiniales parasites. The cross-section of the cell shows the typical folded structure of the parasitoid chain, which fills the entire host cell. Each nucleus (blue) of the coiled ribbon corresponds to a future free-living parasite. DNA is stained with Hoechst (dark blue), membranes are stained with DiOC6 (green), and specimen surface is light blue. Scale bar, 5 μm. (**B**) Subnetwork of metanodes that encapsulate barcodes affiliated to parasites or PFTs. The PFTs mapped onto the network are: phytoplankton DMS producers, mixed phytoplankton, phytoplankton silicifiers, pico-eukaryotic heterotrophs, proto-zooplankton and meso-zooplankton. Edge width reflects the number of edges in the taxon graph between the corresponding metanodes. Over-represented links (multiple-test corrected $P < 0.05$) are colored in green if they represent copresences and in red if they represent exclusions; gray means non-overrepresented combinations. When both copresences and exclusions were significant, the edge is shown as copresence. (**C**) Parasite connections within micro- and zooplankton groups. (**D**) Number of hosts per phage. (Inset) Phage associations to bacterial (target) phyla. (**E**) Putative Bacteroidetes viruses detected with co-occurence and detection in a single-cell genome (SAG). On the left are viral sequences from a Flavobacterium SAG (top) and *Tara* Oceans virome (bottom), displaying an average of 89% nucleotide identity. On the right is the correspondence between the ribosomal genes detected in the same SAG (top) and the 16*S* sequence associated to the *Tara* Oceans contig based on co-occurence (79% nucleotide identity). For clarity, a subset of contig ARTD0100013 only (from 10,000 to 16,000 nucleotides) is displayed. This sequence was also reverse-complemented. PurM, phosphoribosylaminoimidazole synthetase; DNA Pol. A, DNA polymerase A.

Sea (MS), Red Sea (RS), Indian Ocean (IO), South Atlantic (SAO), Southern Ocean (SO), South Pacific Ocean (SPO) and North Atlantic Ocean (NAO)—and assessed the "locality" of associations by comparing the score with or without that region. We found that association patterns were mostly driven by global trends because only 14% of edges were identified as local (Fig. 2, B and C). Approximately two thirds of local associations occur in MS (7215), followed by SPO (1058), whereas the rest are contributed by SO (901), IO (894), RS (889), SAO (163), and NAO (60) (Fig. 2, C to G). MS was the region with most sampling sites, which allowed us to recover more local patterns. Nevertheless, Fig. 2, C to G, shows that although the same major groups (order level) interact in both the global and local networks, each local site has its own specific interaction profile ($P < 1^{-8}$) (*33*, *39*, *53*).

Parasite impact on plankton functional types

Parasitic interactions are the most abundant pattern present in the network, which is also eminent by repeated microscopic observation of parasitic interactions from the *Tara* samples (Fig. 3A). We focused on predicted parasitic interactions and assessed their potential impact on biogeochemical processes by exploring a functional subnetwork (21,572 edges) of known and previously unidentified plankton parasites (*10*) together with classical "plankton functional types" (PFTs) (*54*). PFTs group taxa by trophic strategy (for example, autotrophs versus heterotrophs) and role in ocean biogeochemistry (Fig. 3A) (*55*). The relationship between the different PFTs (network density of 0.65) highlights strong dependencies between phytoplankton and grazers. We found that all PFTs are associated with parasites, but not always to the same extent. Most links involve syndiniales MALV-I and MALV-II clades associated to zooplankton and, to a lesser extent, to microphytoplankton (excluding diatoms). This emphasizes the role of alveolate parasitoids as top-down effectors of zooplankton and microphytoplankton population structure and functioning (*3*), although the latter group is also affected by grazing (*1*). The meso-planktonic networks contain known syndiniales targets (Dinophyceae, Ciliophora, Acantharia, and Metazoa) (Fig. 3B) (*56*). In large size fractions, we found interactions between known parasites and groups of organisms that in theory are too small to be their hosts (*57*); 32% of these associations involved the abundant and diverse marine stramenopiles (MASTs) and diplonemids (other Discoba and Diplonema) (*10*). Ecophysiology studies (*58*, *59*) suggest a parasitic role for these lineages. The association of these groups with other parasites would be explained by putative co-infection of the same hosts. Contrasting with the above observations, we found phytoplankton silicifiers (diatoms) displaying a variety of mutual exclusions. One possible interpretation of this is that diatom silicate exoskeletons (*60*) and toxic compound production (*49*) could act as efficient barriers against top-down pressures (*61*).

Phage-microbe associations

We investigated phage-microbe interactions, another major top-down process affecting global bacterial/archaeal community structure (*7*). Here, surface (SRF) and deep chlorophyll maximum (DCM) virus-bacteria networks revealed 1869 positive associations between viral populations and 7 of the 54 known bacterial phyla (specifically, Proteobacteria, Cyanobacteria, Actinobacteria, Bacteroidetes, Deferribacteres, Verrucomicrobia, and Planctomycetes), and one archaeal phylum (Euryarchaeota). These eight phyla represent most of abundant bacterial/archaeal groups across 37 investigated samples (Fig. 3D), suggesting that the networks are detecting abundant virus-host interactions. Additionally, these interactions include phyla of microbes lacking viral genomes in RefSeq databases including Verrucomicrobia, and nonextremophile Euryarchaeota, hinting at genomic sequences for understudied viral taxa (Fig. 3E) (*39*, *62*, *63*). Among the phage populations in the network, we found eight corresponding to phage sequences available in GenBank (>50% of genes with a >50% amino acid identity match). In all eight cases, the predicted host from the network corresponded to the annotated host family in the GenBank record, which is significantly higher than expected by chance ($P = 0.001$) (*62*).

Next, we evaluated viral host range, which is fundamental for predictive modeling and thus far largely limited to observations of cultured virus-host systems that insufficiently map complex community interactions (*64*). Our virus-host interaction data suggest that viruses are very host-specific: ~43% of the phage populations interact with only a single host OTU, and the remaining 57% interact with only a few, often closely related OTUs (Fig. 3D). These networks are modular at large scales (*65*), suggesting that viruses are host range–limited across large sections of host space. Nestedness analysis showed inconsistent results across algorithms.

Microscopic validation of predicted interactions

Our data predicted a photosymbiotic interaction between an acoel flatworm (*Symsagittifera* sp.) and a green microalga (*Tetraselmis* sp.). We validated this by means of laser scanning confocal microscopy (LSCM), three-dimensional (3D) reconstruction, and reverse molecular identification on flatworm specimens isolated from *Tara* Oceans preserved morphological samples. We observed microalgal cells (5 to 10 μm in diameter) within each of the 15 isolated acoel specimens (Fig. 4) (*66*). The 18*S* sequence from several sorted holobionts matched the metabarcode pair identified in the co-occurrence global network. Thus, molecular ecology, bioinformatics, and microscopic analysis can enable the discovery of marine symbioses.

Conclusions

The global ocean interactome can be used to predict the dynamics and structure of ocean ecosystems. The interactome reported here spans all three organismal domains and viruses. The analyses presented emphasize the role of top-down biotic interactions in the epipelagic zone. This data will inform future research to understand how symbionts, pathogens, predators, and parasites interact with their target organisms and will ultimately help elucidate the structure of the global food webs that drive nutrient and energy flow in the ocean.

Methods

Sampling

The sampling strategy used in the *Tara* Oceans expedition is described in (*67*), and samples used in the present study are listed in table S1 and http://doi.pangaea.de/10.1594/PANGAEA.840721. The *Tara* Oceans nucleotide sequences are available at the European Nucleotide Archive (ENA) under projects PRJEB402 and PRJEB6610.

Physical and environmental measurements

Physical and environmental measurements were carried out with a vertical profile sampling system (CTD-rosette) and data collected from Niskin bottles. We measured temperature, salinity, chlorophyll, CDOM fluorescence (fluorescence of the colored dissolved organic matter), particles abundance, nitrate concentration, and particle size distribution (using an underwater vision profiler). In addition, mean mixed-layer depth (MLD), maximum fluorescence, vertical maximum of the Brünt-Väisälä Frequency N ($s - 1$), vertical range of dissolved oxygen, and depth of nitracline were determined. Satellite altimetry provided the Okubo-Weiss parameter, Lyapunov exponent, mesoscale eddie retention, and sea-surface temperature (SST) gradients at eddie fronts (*19*). Data are available at http://www.pangaea.de (http://doi.pangaea.de/10.1594/PANGAEA.840718).

Abundance table construction

Prokaryotic 16*S* rDNA metagenomic reads were identified, annotated, and quantified from $_{mi}$tags) as described in (*68*) by using the SILVA v.115 database (*19*, *69*, *70*). The abundance table was normalized by using the summed read count per sample (*19*, *71*). Quality-checked V9 rDNA metabarcodes were clustered into swarms as in (*10*, *72*) and annotated by using the V9 PR2 database (*73*). PR2 barcodes were associated to fundamental trophic modes (auto- or heterotrophy) and symbiotic interactions (parasitism and mutualism) according to literature (Taxonomic and trophic mode annotations are available at http://doi.pangaea.de/10.1594/PANGAEA.843018 and http://doi.pangaea.de/10.1594/PANGAEA.843022). Swarm abundance and normalization was performed as in (*10*, *72*). Bacteriophage metagenomes were obtained from the < 0.2-μm fractions for 48 samples, and contigs were annotated and quantified as in (*18*). The abundance matrix was normalized by means of total sample read count and contig length.

In all cases, only OTUs with relative abundance $> 1^{-8}$ and detected in at least 20% of samples were retained. Because sample number in the input tables ranged from 17 to 63, prevalence thresholds varied (from 22 to 40%). The sum of all filtered OTU relative abundances was kept in the tables to preserve proportions. Abundance tables

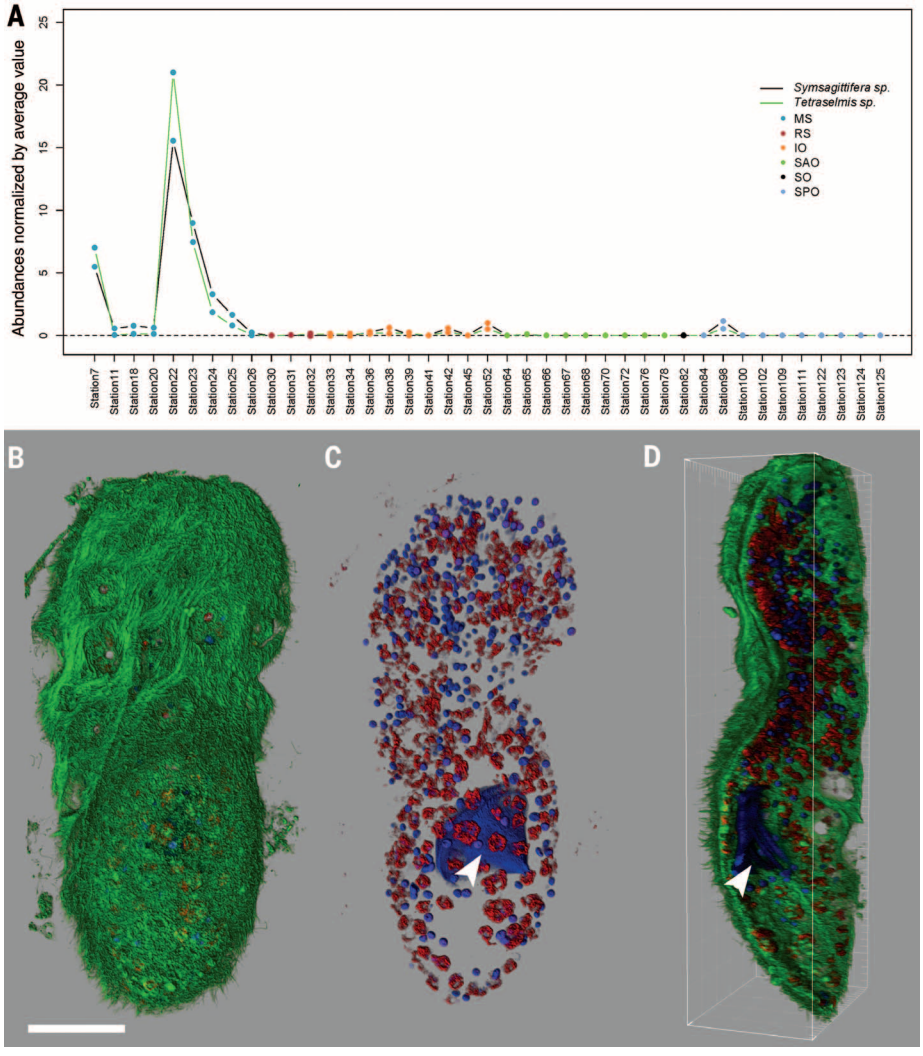


Fig. 4. Experimental validation of network-predicted interaction (photosymbiosis). Guided by the predictions from the co-occurrence network and abundance patterns, acoel flatworms (*Symsagittifera* sp.) together with their photosynthetic green microalgal endosymbionts (*Tetraselmis* sp.) were collected in microplankton samples from *Tara* Oceans Station 22 in the Mediterranean Sea. Pictures show a 3D reconstructed specimen from LSCM images [green channel, cellular membranes (DiOC6); blue channel, DNA and the nuclei (Hoechst33342); red channel: chlorophyll autofluorescence]. (**A**) Co-occurrence plot of *Symsagittifera*- and *Tetraselmis*-related OTUs along *Tara* Oceans stations, showing the relatively high abundance of the holobiont at Station 22. (**B**) Dorsal view of the entire acoel flatworm specimen (~300 μm). The epidermis (green) is completely covered with cilia and displays some pore holes. (**C**) The removal of the green channel reveals the widespread distribution of small unicellular algae (red areas) inside the acoel body. The worm's nuclei display a clear signal (compact round blue shapes), whereas the algal nuclei are dimmer. A dinoflagellate theca (arrowhead) is located in the central syncytium, likely indicating predation. (**D**) Cross-section along a *z*-*y* plane allows localization of the algae, beneath the epidermis in the parenchyma. Only the external cell layer (green signal) from the dorsal view is visible because of the thickness and opacity of the worm. Scale bar, 50 μm.

are available at www.raeslab.org/companion/ocean-interactome.html.

Random forest-based models

Eukaryotic, prokaryotic, and environmental matrices were merged into two matrices [deep chlorophyll maximum layer (DCM) and surface water layer (SRF)]. For each of the three models [OTU versus other OTUs (M_{OTU}), environmental factors (M_{ENV}) or combined ($M_{OTU+ENV}$)], regressions were perfomed with OTU abundance as dependent and the abundances of other OTUs or environmental factors as independent variables. For each regression, up to 20 independent variables were selected by using the minimum Redundancy Maximum Relevance (mRMR) filter-ranking algorithm. Random forest regression (*25*) was followed by a leave-one-out cross-validation. The variable subset with the minimum leave-one-out NMSE (normalized mean square error) was selected. To identify the best model for a given target OTU, the significance of the NMSE difference was tested on the absolute error values [paired Wilcoxon test adjusted by Benjamini-Hochberg false discovery rate (FDR) estimation (*74*)]. NMSE computed on random data are larger than those from original data. In addition, M_{ENV} outperformed M_{OTU} when OTU abundances were randomized.

Variance partitioning

Environmental variables were *z* score–transformed; spatial variables (MEM eigenvectors) were calculated based on latitude and longitude (*75*). Forward selection (*76*) was carried out with function forward.sel in R-package packfor. Significance of the selected variables was assessed with 1000 permutations by using functions rda and anova.cca in vegan. Variance partitioning (*77*) was performed by using function varpart in vegan on Hellinger-transformed abundance data, the forward-selected environmental variables, and the forward-selected spatial variables and tested for significance with 1000 permutations.

Network inference

Taxon-taxon co-occurrence networks were constructed as in (*78*), selecting Spearman and Kullback-Leibler dissimilarity measures. To compute *P* values, we first generated permutation and bootstrap distributions, with 1000 iterations each, by shuffling taxon abundances and resampling from samples with replacement, respectively. The measure-specific *P* value was then obtained as the probability of the null value (represented by the mean of the permutation distribution) under a Gauss curve fitted to the mean and standard deviation of the bootstrap distribution. Permutations computed for Spearman included a renormalization step, which mitigates compositionality bias (ReBoot). Measure-specific *P* values were merged by using Brown's method (*79*) and multiple-testing-corrected with Benjamini-Hochberg (*74*). Last, edges with an adjusted *P* value above 0.05, with a score below the thresholds (*30*) or not supported by both measures after assessment of significance, were discarded.

Taxon-environment networks were computed with the same procedure, starting with 8000 initial positive and negative edges, each supported by both methods. For computational efficiency, we computed 23 taxon-taxon and taxon-environment networks separately, for two depths (DCM and SRF), four eukaryotic size fractions (0.8 to 5 μm, >0.8 μm, 20 to 180 μm, and 180 to 2000 μm) and their combinations, the prokaryotic size fraction (0.2 to 1.6 μm and 0.2 to 3.0 μm) and its combination with each of the eukaryotic and virus (<0.2 μm) size fractions. We then generated 23 taxon-environment union networks for environmental triplet detection and merged the taxon-taxon networks into a global network with 92,633 edges.

Estimation of false discovery rate

We estimated the FDR of network construction with two null models. The first shuffles counts while preserving overall taxon proportions and total sample count sums, but removing any dependencies between taxa. For the second,

we fitted a Dirichlet-multinomial distribution to the input matrix using the dirmult package in R (*80*) and generated a null matrix by sampling from this distribution, preserving total sample count sums. Null matrices were generated from count matrices (0.8 to 5 μm, 20 to 180 μm, and 180 to 2000 μm eukaryotic and prokaryotic size fraction as well as bacteriophage-prokaryotic composite, SRF, and DCM). Network construction was performed with the 20 null matrices and thresholds applied to the original matrices (*28*). From edge numbers in the original and the null networks, we estimated an average FDR of 9% (*28*).

Indirect taxon edge detection

For each taxon-environment union network, node triplets consisting of two taxa and one environmental parameter were identified. For each triplet, interaction information II was computed as $II = CI(X, Y \mid Z) - I(X, Y)$, where *CI* is the conditional mutual information between taxa *X* and *Y* given environmental parameter *Z*, and *I* is the mutual information between *X* and *Y*. *CI* and *I* were estimated by using minet (*81*). Taxon edges in environmental triplets were considered indirect when $II < 0$ and within the 0.05 quantile of the random II distribution obtained by shuffling environmental vectors (500 iterations). If a taxon pair was part of more than one environmental triplet, the triplet with minimum interaction information was selected.

For each environmental triplet, we also checked whether its sign pattern (the combination of positive and/or negative correlations) was consistent with an indirect interaction. From eight possible patterns, four indicate indirect relationships (for example, two negatively correlated taxa correlated with opposite signs to an environmental factor).

Network deconvolution (*32*) was carried out with $\beta = 0.9$. We considered an environmental triplet as indirect according to network deconvolution if any of its edges were removed.

All (11,043) negative interaction information triplets were consistent with an indirect relationship according to their sign patterns, and a majority (8209) was also supported by network deconvolution.

Influence of ocean regions on co-occurrence patterns

Samples were divided into groups according to region membership. The impact of each sample group on the Spearman correlation of each edge in the network was assessed by dividing the (absolute) omission score (OS) (Spearman correlation without these samples) by the absolute original Spearman score. To account for group size, the OS was computed repeatedly for random, same-sized sample sets. Nonparametric *P* values were calculated as the number of times random OSs were smaller than the sample group OS, divided by number of random OS (500 for each taxon pair). Edges were classified as region-specific when the ratio of OS and absolute original score was below 1 and multiple-testing-corrected *P* values (Benjamini-Hochberg) were below 0.05.

Overrepresentation analysis

Significance of taxon–taxon counts at high taxonomic ranks was assessed with the hypergeometric distribution implemented in the R function phyper. Mutual exclusion versus copresence analysis was performed by using the binomial distribution implemented in the R function pbinom, with the background probability estimated by the frequency of edges in the network.

Oceanic region analysis was also assessed by use of R's pbinom function, with the background probability estimated by dividing total ocean-specific edge number by total edge number. The *P* value was computed as the probability of obtaining the observed number of ocean-specific edges among the edges of a taxon pair. The same procedure was repeated for each oceanic region separately, with region-specific success probabilities. Edges classified as indirect were discarded before the analysis.

In all tests, *P* values were adjusted for multiple testing according to Benjamini, Hochberg, and Yekutieli (BY), implemented in the R function p.adjust.

Extracting functional groups from the global plankton interactome

Functional groups consist of a mix of major monophyletic lineages of parasites, together with classical polyphyletic PFTs, as defined in (*10*, *54*, *55*). Metabarcodes in the network were sorted into 15 parasite groups and seven PFTs (*55*) according to their (i) taxonomical classification, (ii) membership in a given size fraction, (iii) trophic mode, and (iv) biogeochemical role in dimethyl sulfide (DMS) production or silicification. After mapping the metabarcodes and their edges onto PFTs and parasites, edges are weighted by the number of links they represent. Overrepresentation of the number of links included in each edge was assessed with the hypergeometric distribution.

Parasite links in large fractions may point to parasite-host connections. We extracted all edges in the large fractions (20 to 180 μm and 180 to 2000 μm) between barcodes annotated as parasites and nonparasitic barcodes. Partners of parasites comprised potential hosts (Fig. 3B) but also organisms that are either too small or without size information. The former may represent unknown parasites (for example, coinfecting a host with known parasites), whereas the latter may represent previously unknown hosts.

Nestedness and modularity analysis

The analysis was carried out for 1869 positively correlated phage-prokaryotic pairs. Modularity was computed with the LP (Label propagation) BRIM algorithm (*82*) in BiMAT (*83*) with 100 permutations. Nestedness of the host-phage network as quantified with the NODF (nestedness with overlap and decreasing fill) algorithm (*84*) in BiMAT with 100 permutations (preserving edge number and degree distribution) was significant, but not with the NTC algorithm (*85*). We also tested the impact of random removal or addition of 5, 10, 15, and 20% edges. After random addition/deletion of edges, modularity and nestedness (according to NODF) remained significant.

Confirmation of predicted viruses-host associations

Two different approaches were used to confirm virus-host associations predicted by the co-occurrence network. First, the network host prediction was compared with the "known" host for viral populations closely related to an isolated virus—populations with more than 50% of predicted genes affiliated to the same phage reference genome [based on a BLASTp against RefseqVirus, threshold of 10^{-03} on e-value and 50 on bit score (*18*)]. Known phages corresponded to viruses infecting SAR11, SAR116, and Cyanobacteria, so that a predicted host was considered correct if affiliated to Alphaproteobacteria, Alphaproteobacteria, and Cyanobacteria, respectively [the lowest rank for which there was taxonomic assignment for those bacterial OTUs (*69*)]. This procedure was repeated on 1000 randomized networks (with same-degree distribution) to calculate the significance of the results. Second, contigs of putative hosts predicted by co-occurrence analysis were compared with BLAST to a set of viral sequences detected in draft and single-cell genomes with VirSorter (https://pods.iplantcollaborative.org/wiki/display/DEapps/VIRSorter+1.0.2). One contig (36DCM_3902) (Fig. 3E) displayed significant sequence similarity (blastn *e*-value < 10^{-151} over two segments) to one contig detected in a single-cell genome (AA160P02DRAFT_ scaffold_31.32). In order to compare the putative host associated to each contig, rRNA genes were predicted in the single-cell amplified genome (SAG) contigs with meta-rRNA (*86*). Sequences were annotated based on BLAST against the nonredundant (nr) database, and the comparison plot was generated with Easyfig (*87*).

Literature-based evaluation of predicted protist interactions

A panel of four experts, two specialized in the study of planktonic mutualistic protists (C.d.V. and J.D.) and two specialized in the study of planktonic parasitic protists (C. Berney and N.H.), screened literature looking for symbiotic interactions occurring among eukaryotic plankton. From this search, they built a list of 574 known symbiotic interactions sensu lato (parasitism and mutualism, at least one protist partner) in marine eukaryotic plankton, covering 197 eukaryotic genera, described in 76 publications since 1971. The experts extracted only symbiotic interaction cases described either from direct observation of both interacting partners through microscope (45%), sequence from symbiont isolated from the observed host (14%), or both (41%). Direct observation of partners interacting (86%) provides high confidence for the interaction, and the symbiont sequence allows its taxonomic identification. The protocol to build the list was the following: (i) the experts manually screened 3170 publications associated to each PR2 db sequence http://ssu-rrna.org/pr2 (*73*); (ii) the experts screened 293 publications

retrieved from Web of Science with the following query: "TOPIC:(plankton* AND (marin* OR ocean*)) AND (parasit* OR symbios* OR mutualis*)"; (iii) the experts screened GenBank 18*S* rDNA sequences of symbionts for which the "host" field was known. They labeled these interactions as "Unpublished." Last, the experts discussed any observed discordance until agreement was reached. The final table of literature-curated interactions includes a column indicating the type of evidence gathered about the interaction: 1 for only getting symbiont sequence, 2 for direct observation, and 3 for both. Symbiont GenBank host field belongs to category 1.

Experimental validation of a predicted interaction

V9 pairs were searched for organisms of suitable size in order to allow its isolation from morphological samples. This way, we targeted a predicted photosymbiosis between an acoel flatworm [V9 rDNA metabarcode 83% similar to *Symsagittifera psammophila* (*88*)] and a photosynthetic microalga (*Tara* Oceans V9 metabarcode 100% similar to a *Tetraselmis* sp) (*89*).

Fifteen acoel specimens (hosts) were isolated from formaldehyde-4% microplankton samples of station 22 (A100000458), in which both partner OTUs displayed high abundances. Before imaging, specimens were rinsed with artificial seawater, then DNA and membrane structures were stained for 60 min with 10 μM Hoechst 33342 and 1.4 μM DiOC6(3) (Life Technologies, Grand Island, NY). Microscopy was conducted by using a Leica TCS SP8 (Leica Microsystems, Wetzlar, Germany) confocal laser scanning microscope and a HC PL APO 40x/1.10 W motCORR CS2 objective. The DiOC6 signal (ex488nm/em500-520nm) was collected simultaneously with the chlorophyll signal (ex488nm/em670-710nm), followed by the Hoechst signal (ex405 nm/em420-470nm). Images were processed with Fiji (*90*), and 3D specimens were reconstructed with Imaris (Bitplane, Belfast, UK).

To obtain the sequences of the metabarcodes of each partner, seven acoels were isolated from ethanol-preserved samples from station 22 (TARA_A100000451), individually rinsed in filtered seawater, and stored at −20°C in absolute ethanol. DNA was extracted with MasterPureTM DNA/RNA purification kit (Epicenter, Madison, WI) and polymerase chain reaction amplified by using the universal-eukaryote primers (forward 1389F and reverse 1510R) from (*10*). Chlorophyte-specific primers (Chloro2F: 5′- CGTATATTTAAGTT-GYTGCAG-3′ and Tetra2-rev 5′- CAGCAATGGGC-GGTGGC GAAC-3′) were designed to amplify the microalgae V9 rDNA as in (*4*). Purified amplicons were subjected to poly-A reaction and ligated in pCR®4-TOPO TA Cloning vector (Invitrogen, Carlsbad, CA), cloned by using chemically competent *Escherichia coli* cells, and Sanger-sequenced with the ABI-PRISM Big Dye Terminator Sequencing kit (Applied Biosystems, Foster City, CA) by using the 3130xl Genetic Analyzer (Applied Biosystems).

REFERENCES AND NOTES

1. F. Azam *et al*., The ecological role of water-column microbes in the sea. *Mar. Ecol. Prog. Ser.* **10**, 257–263 (1983). doi: 10.3354/meps010257
2. A. W. Thompson *et al*., Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science* **337**, 1546–1550 (2012). doi: 10.1126/science.1222700; pmid: 22997339
3. A. Chambouvet, P. Morin, D. Marie, L. Guillou, Control of toxic marine dinoflagellate blooms by serial parasitic killers. *Science* **322**, 1254–1257 (2008). doi: 10.1126/science.1164387; pmid: 19023082
4. J. Decelle *et al*., An original mode of symbiosis in open ocean plankton. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 18000–18005 (2012). doi: 10.1073/pnas.1212303109; pmid: 23071304
5. V. Smetacek, Making sense of ocean biota: How evolution and biodiversity of land organisms differ from that of the plankton. *J. Biosci.* **37**, 589–607 (2012). doi: 10.1007/s12038-012-9240-4; pmid: 22922185
6. J. L. Sabo, L. R. Gerber, "Trophic ecology," *AccessScience* (McGraw-Hill Education, 2014); available at www.accessscience.com/content/trophic-ecology/711650.
7. F. Rohwer, R. V. Thurber, Viruses manipulate the marine environment. *Nature* **459**, 207–212 (2009). doi: 10.1038/nature08060; pmid: 19444207
8. P. G. Verity, V. Smetacek, Organism life cycles, predation, and the structure of marine pelagic ecosystems. *Mar. Ecol. Prog. Ser.* **130**, 277–293 (1996). doi: 10.3354/meps130277
9. A. Z. Worden *et al*., Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science* **347**, 1257594 (2015). doi: 10.1126/science.1257594; pmid: 25678667
10. C. de Vargas *et al*., *Science* **348**, XXX–XXX (2014).
11. K. Faust, J. Raes, Microbial interactions: From networks to models. *Nat. Rev. Microbiol.* **10**, 538–550 (2012). doi: 10.1038/nrmicro2832; pmid: 22796884
12. S. Chaffron, H. Rehrauer, J. Pernthaler, C. von Mering, A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* **20**, 947–959 (2010). doi: 10.1101/gr.104521.109; pmid: 20458099
13. J. Raes, I. Letunic, T. Yamada, L. J. Jensen, P. Bork, Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Syst. Biol.* **7**, 473 (2011). doi: 10.1038/msb.2011.6; pmid: 21407210
14. J. A. Gilbert *et al*., Defining seasonal marine microbial community dynamics. *ISME J.* **6**, 298–308 (2012). doi: 10.1038/ismej.2011.107; pmid: 21850055
15. J. M. Beman, J. A. Steele, J. A. Fuhrman, Co-occurrence patterns for abundant marine archaeal and bacterial lineages in the deep chlorophyll maximum of coastal California. *ISME J.* **5**, 1077–1085 (2011). doi: 10.1038/ismej.2010.204; pmid: 21228895
16. C.-E. T. Chow, D. Y. Kim, R. Sachdeva, D. A. Caron, J. A. Fuhrman, Top-down controls on bacterial community structure: Microbial network analysis of bacteria, T4-like viruses and protists. *ISME J.* **8**, 816–829 (2014). doi: 10.1038/ismej.2013.199; pmid: 24196323
17. E. Karsenti *et al*., A holistic approach to marine eco-systems biology. *PLOS Biol.* **9**, e1001177 (2011). doi: 10.1371/journal.pbio.1001177; pmid: 22028628
18. J. R. Brum *et al*., *Science* **348**, XXX–XXX (2014).
19. S. Sunagawa *et al*., *Science* **348**, XXX–XXX (2014).
20. E. Villar *et al*., *Science* **348**, XXX–XXX (2014).
21. Companion web site table w2; available at www.raeslab.org/companion/ocean_interactome/tables/W2.xlsx.
22. A. Meot, P. Legendre, D. Borcard, *Environ. Ecol. Stat.* **5**, 1–27 (1998). doi: 10.1023/A:1009693501830
23. Companion web site table w3; available at www.raeslab.org/companion/ocean_interactome/tables/W3.xlsx.
24. Companion web site figure w1; available at www.raeslab.org/companion/ocean_interactome/figures//W1.pdf .
25. L. Breiman, *Mach. Learn.* **45**, 5–32 (2001). doi: 10.1023/A:1010933404324
26. Companion web site table w4; available at www.raeslab.org/companion/ocean_interactome/tables/W4.xlsx.
27. Companion web site figure w2; available at http://www.raeslab.org/companion/ocean_interactome/figures/W2.pdf.
28. Companion web site table w5; available at www.raeslab.org/companion/ocean_interactome/tables/W5.xlsx.
29. Companion web site table w6; available at http://www.raeslab.org/companion/ocean_interactome/tables/W6.xlsx
30. Companion web site table w7; available at www.raeslab.org/companion/ocean_interactome/tables/W7.xlsx.
31. Companion web site figure w3; available at http://www.raeslab.org/companion/ocean_interactome/figures/W3.pdf.
32. S. Feizi, D. Marbach, M. Médard, M. Kellis, Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.* **31**, 726–733 (2013). doi: 10.1038/nbt.2635; pmid: 23851448
33. Companion web site table w8 available at www.raeslab.org/companion/ocean_interactome/tables/W8.xlsx.
34. M. E. Cusick *et al*., Literature-curated protein interaction datasets. *Nat. Methods* **6**, 39–46 (2009). doi: 10.1038/nmeth.1284; pmid: 19116613
35. Companion web site table w9; available at www.raeslab.org/companion/ocean_interactome/tables/W9.xlsx.
36. J. A. Fuhrman, J. A. Cram, D. M. Needham, Marine microbial community dynamics and their ecological interpretation. *Nat. Rev. Microbiol.* **13**, 133–146 (2015). doi: 10.1038/nrmicro3417; pmid: 25659323
37. Companion web site additional material is available at www.raeslab.org/companion/ocean_interactome/Accompanying_Material.docx.
38. Companion web site figure w4; available at www.raeslab.org/companion/ocean_interactome/figures/W4.pdf.
39. Companion web site table w10; available at www.raeslab.org/companion/ocean_interactome/tables/W10.xlsx.
40. G. B. McManus, L. F. Santoferrara, *The Biology and Ecology of Tintinnid Ciliates* (John Wiley & Sons, New York, 2012.
41. J. Cachon, in *Ann sci nat b*. (Paris, 1964), vol. 6, p. 1.
42. P. S. Salomon, E. Granéli, M. H. C. B. Neves, E. G. Rodriguez, Infection by Amoebophrya spp. parasitoids of dinoflagellates in a tropical marine coastal area. *Aquat. Microb. Ecol.* **55**, 143–153 (2009). doi: 10.3354/ame01293
43. F. Gómez, P. López-García, A. Nowaczyk, D. Moreira, The crustacean parasites Ellobiopsis Caullery, 1910 and Thalassomyces Niezabitowski, 1913 form a monophyletic divergent clade within the Alveolata. *Syst. Parasitol.* **74**, 65–74 (2009). doi: 10.1007/s11230-009-9199-1; pmid: 19633933
44. S. Ohtsuka *et al*., Morphology and host-specificity of the apostome ciliate Vampyrophrya pelagica infecting pelagic copepods in the Seto Inland Sea, Japan. *Mar. Ecol. Prog. Ser.* **282**, 129–142 (2004). doi: 10.3354/meps282129
45. A. Skovgaard, S. A. Karpov, L. Guillou, The parasitic dinoflagellates Blastodinium spp. inhabiting the gut of marine, planktonic copepods: Morphology, ecology, and unrecognized species diversity. *Front. Microbiol.* **3**, 305 (2012). doi: 10.3389/fmicb.2012.00305; pmid: 22973263
46. L. Stemmann *et al*., Global zoogeography of fragile macrozooplankton in the upper 100-1000 m inferred from the underwater video profiler. *ICES J. Mar. Sci.* **65**, 433–442 (2008). doi: 10.1093/icesjms/fsn010
47. J. M. Gasol, P. A. Del Giorgio, C. M. Duarte, *Biomass Distribution in Marine Planktonic Communities* (American Society of Limnology and Oceanography, Waco, TX, 1997), vol. 42.
48. B. A. Ward, S. Dutkiewicz, M. J. Follows, Modelling spatial and temporal patterns in size-structured marine plankton communities: Top-down and bottom-up controls. *J. Plankton Res.* **36**, 31–47 (2014). doi: 10.1093/plankt/fbt097
49. A. Ianora *et al*., Aldehyde suppression of copepod recruitment in blooms of a ubiquitous planktonic diatom. *Nature* **429**, 403–407 (2004). doi: 10.1038/nature02526; pmid: 15164060
50. M. Martinez-Garcia *et al*., Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. *ISME J.* **6**, 703–707 (2012). doi: 10.1038/ismej.2011.126; pmid: 21938022
51. E. T. Jolley, A. K. Jones, The interaction between *Navicula muralis* grunow and an associated species of *Flavobacterium*. *Br. Phycol. J* **12**, 315–328 (1977). doi: 10.1080/00071617700650341
52. T. R. Miller, R. Belas, Dimethylsulfoniopropionate metabolism by Pfiesteria-associated Roseobacter spp. *Appl. Environ. Microbiol.* **70**, 3383–3391 (2004). doi: 10.1128/AEM.70.6.3383-3391.2004; pmid: 15184135
53. Companion web site figure w5; available at www.raeslab.org/companion/ocean_interactome/figures/W5.pdf.
54. C. Le Quere *et al*., *Glob. Change Biol.* **11**, 17 (2005).
55. Companion web site table w11; available at www.raeslab.org/companion/ocean_interactome/tables/W11.xlsx.
56. A. Skovgaard, *Acta Protozool.* **53**, 51 (2014).
57. Companion web site figure w6; available at www.raeslab.org/companion/ocean_interactome/figures/W6.pdf.
58. S. von der Heyden, E. E. Chao, K. Vickerman, T. Cavalier-Smith, Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of euglenozoa. *J. Eukaryot. Microbiol.* **51**, 402–416 (2004). doi: 10.1111/j.1550-7408.2004.tb00387.x; pmid: 15352322

59. F. Gómez, D. Moreira, K. Benzerara, P. López-García, Solenicola setigera is the first characterized member of the abundant and cosmopolitan uncultured marine stramenopile group MAST-3. *Environ. Microbiol.* **13**, 193–202 (2011). doi: 10.1111/j.1462-2920.2010.02320.x; pmid: 20722698
60. C. E. Hamm *et al*., Architecture and material properties of diatom shells provide effective mechanical protection. *Nature* **421**, 841–843 (2003). doi: 10.1038/nature01416; pmid: 12594512
61. P. Assmy, V. Smetacek, in *Encyclopedia of Microbiology*, M. Schaechter, Ed. (Elsevier, Oxford, UK, 2009), pp. 27–41.
62. Companion web site table w12; available at www.raeslab.org/companion/ocean_interactome/tables/W12.xlsx.
63. Companion web site figure w7; available at www.raeslab.org/companion/ocean_interactome/figures/W7.pdf.
64. J. S. Weitz *et al*., Phage-bacteria infection networks. *Trends Microbiol.* **21**, 82–91 (2013). doi: 10.1016/j.tim.2012.11.003; pmid: 23245704
65. Companion web site table w13; available at www.raeslab.org/companion/ocean_interactome/tablesW13.xlsx.
66. Companion web site figure w9; available at www.raeslab.org/companion/ocean_interactome/companion_figures/W9.pdf.
67. S. Pesant *et al*., Open science resources for the discovery and analysis of *Tara* Oceans data. http://biorxiv.org/content/early/2015/05/08/019117 (2015).
68. R. Logares *et al*., Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* **16**, 2659–2671 (2013).
69. E. Pruesse *et al*., SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**, 7188–7196 (2007). doi: 10.1093/nar/gkm864; pmid: 17947321
70. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010). doi: 10.1093/bioinformatics/btq461; pmid: 20709691
71. Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007). doi: 10.1128/AEM.00062-07; pmid: 17586664
72. F. Mahé, T. Rognes, C. Quince, C. de Vargas, M. Dunthorn, Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593 (2014). doi: 10.7717/peerj.593; pmid: 25276506
73. L. Guillou *et al*., The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41** (D1), D597–D604 (2013). doi: 10.1093/nar/gks1160; pmid: 23193267
74. Y. Benjamini, Y. Hochberg, *J. R. Stat. Soc., B* **57**, 289 (1995).
75. D. Borcard, P. Legendre, C. Avois-Jacquet, H. Tuomisto, DIssecting the spatial structure of ecological data at multiple scales. *Ecology* **85**, 1826–1832 (2004). doi: 10.1890/03-3111
76. F. G. Blanchet, P. Legendre, D. Borcard, Forward selection of explanatory variables. *Ecology* **89**, 2623–2632 (2008). doi: 10.1890/07-0986.1; pmid: 18831183
77. D. Borcard, P. Legendre, P. Drapeau, Partialling out the spatial component of ecological variation. *Ecology* **73**, 1045 (1992). doi: 10.2307/1940179
78. K. Faust *et al*., Microbial co-occurrence relationships in the human microbiome. *PLOS Comput. Biol.* **8**, e1002606 (2012). doi: 10.1371/journal.pcbi.1002606; pmid: 22807668
79. M. B. Brown, 400: A method for combining non-independent, one-sided tests of significance. *Biometrics* **31**, 987 (1975). doi: 10.2307/2529826
80. T. Tvedebrink, Overdispersion in allelic counts and θ-correction in forensic genetics. *Theor. Popul. Biol.* **78**, 200–210 (2010). doi: 10.1016/j.tpb.2010.07.002; pmid: 20633572
81. P. E. Meyer, F. Lafitte, G. Bontempi, minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics* **9**, 461 (2008). doi: 10.1186/1471-2105-9-461; pmid: 18959772
82. L. Xin, T. Murata, in *Web Intelligence and Intelligent Agent Technologies, 2009* (WI-IAT 09. IEEE/WIC/ACM International Joint Conferences, 2009), vol. 1, pp. 50.
83. C. O. Flores, T. Poisot, S. Valverde, J. S. Weitz, http://arxiv.org/abs/1406.6732 (2014)
84. M. J. Barber, Modularity and community detection in bipartite networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **76**, 066102 (2007). doi: 10.1103/PhysRevE.76.066102; pmid: 18233893
85. W. Atmar, B. D. Patterson, The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia* **96**, 373–382 (1993). doi: 10.1007/BF00317508
86. Y. Huang, P. Gilna, W. Li, Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* **25**, 1338–1340 (2009). doi: 10.1093/bioinformatics/btp161; pmid: 19346323
87. M. J. Sullivan, N. K. Petty, S. A. Beatson, Easyfig: A genome comparison visualizer. *Bioinformatics* **27**, 1009–1010 (2011). doi: 10.1093/bioinformatics/btr039; pmid: 21278367
88. I. Ruiz-Trillo, M. Riutort, D. T. J. Littlewood, E. A. Herniou, J. Baguña, Acoel flatworms: Earliest extant bilaterian Metazoans, not members of Platyhelminthes. *Science* **283**, 1919–1923 (1999). doi: 10.1126/science.283.5409.1919; pmid: 10082465
89. R. J. Gast, T. A. McDonnell, D. A. Caron, srDna-based taxonomic affinities of algal symbionts from a planktonic foraminifer and a solitary radiolarian. *J. Phycol.* **36**, 172–177 (2000). doi: 10.1046/j.1529-8817.2000.99133.x
90. J. Schindelin *et al*., Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012). doi: 10.1038/nmeth.2019; pmid: 22743772
91. M. E. Newman, Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **74**, 036104 (2006). doi: 10.1103/PhysRevE.74.036104; pmid: 17025705

ACKNOWLEDGMENTS

We thank the commitment of the following people and sponsors: Centre National de la Recherche Scientifique (CNRS) (in particular, Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, Fund for Scientific Research – Flanders (G.L.M., K.F., S.C., and J.R.), Rega Institute (J.R.), KU Leuven (J.R.), The French Ministry of Research, the French Government "Investissements d'Avenir" programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), ANR (projects POSEIDON/ANR-09-BLAN-0348, PHYTBACK/ANR-2010-1709-01, PROMETHEUS/ANR-09-PCS-GENM-217, TARA GIRUS/ANR-09-PCS-GENM-218, European Union FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376, ERC Advanced Grant Awards to CB (Diatomite: 294823), Gordon and Betty Moore Foundation grant (3790) to M.B.S., Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to S.G.A., TANIT (CONES 2010-0036) from the Agència de Gestió d´Ajusts Universitaris i Reserca funded to S.G.A., JSPS KAKENHI grant number 26430184 to H.O., FWO, BIO5, Biosphere 2, Agnès b., the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, Etienne Bourgois, the *Tara* schooner, and its captain and crew. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries that graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge the EMBL Advanced Light Microscopy Facility (ALMF), and in particular R. Pepperkok. The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. Data described herein is available at www.raeslab.org/companion/ocean-interactome.html, at the EBI under the projects PRJEB402 and PRJEB6610, and at Pangaea http://doi.pangaea.de/10.1594/PANGAEA.840721, http://doi.pangaea.de/10.1594/PANGAEA.840718, http://doi.pangaea.de/10.1594/PANGAEA.843018 and http://doi.pangaea.de/10.1594/PANGAEA.843022 and on table S1. The data release policy regarding future public release of *Tara* Oceans data are described in Pesant *et al*. (*67*). All authors approved the final manuscript. This article is contribution number 25 of Tara Oceans. The supplementary materials contain additional data.

TARA OCEANS COORDINATORS

Silvia G. Acinas,[1] Peer Bork,[2] Emmanuel Boss,[3] Chris Bowler,[4] Colomban De Vargas,[5,6] Michael Follows,[7] Gabriel Gorsky,[8,9] Nigel Grimsley,[10,11] Pascal Hingamp,[12] Daniele Iudicone,[13] Olivier Jaillon,[14,15,16] Stefanie Kandels-Lewis,[2] Lee Karp-Boss,[3] Eric Karsenti,[17,18] Uros Krzic,[19] Fabrice Not,[5,6] Hiroyuki Ogata,[20] Stephane Pesant,[21,22] Jeroen Raes,[23,24,25] Emmanuel G. Reynaud,[26] Christian Sardet,[8] Mike Sieracki,[27] Sabrina Speich,[28,29] Lars Stemmann,[8] Matthew B. Sullivan,[30] Shinichi Sunagawa,[2] Didier Velayoudon,[31] Jean Weissenbach,[14,15,16] Patrick Wincker[14,15,16]

[1]Department of Marine Biology and Oceanography, Institute of Marine Sciences ICM-CSIC, Barcelona, Spain. [2]Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany. [3]School of Marine Sciences, University of Maine, Orono, USA. [4]Environmental and Evolutionary Genomics Section, Institut de Biologie de l'Ecole Normale Supérieure, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 8197, Institut National de la Santé et de la Recherche Médicale U1024, Ecole Normale Supérieure, Paris, France. [5]CNRS, UMR 7144, Station Biologique de Roscoff, Roscoff, France. [6]Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Roscoff, France. [7]Dept of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, USA. [8]CNRS, UMR 7093, Laboratoire d'Océanographie de Villefranche (LOV), Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [9]Sorbonne Universités, UPMC Paris 06, UMR 7093, Laboratoire d'Océanographie de Villefranche (LOV), Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [10]CNRS UMR 7232, BIOM, Banyuls-sur-Mer, France. [11]Sorbonne Universités, OOB, UPMC Paris 06, Banyuls-sur-Mer, France. [12]Aix Marseille Université, CNRS, IGS UMR 7256, Marseille, France. [13]Laboratory of Ecology and Evolution of Plankton, Stazione Zoologica Anton Dohrn, Naples, Italy. [14]CEA, Genoscope, Evry France. [15]CNRS, UMR 8030, Evry, France. [16]Université d'Evry, UMR 8030, Evry, France. [17]Environmental and Evolutionary Genomics Section, Institut de Biologie de l'Ecole Normale Supérieure, CNRS, UMR 8197, Institut National de la Santé et de la Recherche Médicale U1024, Ecole Normale Supérieure, Paris, France. [18]Directors' Research, European Molecular Biology Laboratory, Heidelberg, Germany. [19]Cell Biology and Biophysics, European Molecular Biology Laboratory, Heidelberg, Germany. [20]Institute for Chemical Research, Kyoto University, Kyoto, Japan. [21]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [22]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [23]Department of Microbiology and Immunology, Rega Institute KU Leuven, Leuven, Belgium. [24]VIB Center for the Biology of Disease, VIB, Leuven, Belgium. [25]Laboratory of Microbiology, Vrije Universiteit Brussel, Brussels, Belgium. [26]School of Biology and Environmental Science, University College Dublin, Dublin, Ireland. [27]Bigelow Laboratory for Ocean Sciences, East Boothbay, USA. [28]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, Paris, France. [29]Laboratoire de Physique des Océan, UBO-IUEM, Polouzané, France. [30]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, USA. [31]DVIP Consulting, Sèvres, France.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/1262073/suppl/DC1
Table S1

3 October 2014; accepted 18 March 2015
10.1126/science.1262073

RESEARCH

How HIV RNA gets packaged into the virus

Keane et al., *p. 917*

IN *SCIENCE* JOURNALS

Edited by **Stella Hurtley**

Glaciers flowing into the sea off the Southern Antarctic Peninsula

GLACIER MASS LOSS

Increasingly rapid ice sheet melting

Glaciers on the Southern Antarctic Peninsula have begun losing mass at a rapid and accelerating rate. Wouters *et al*. documented the dramatic thinning of the land-based ice, which began in 2009, using satellite altimetry and gravity observations. The melting and weakening of ice shelves reduce their buttressing effect, allowing the glaciers to flow more quickly to the sea. — HJS

Science, this issue p. 899

TUMOR EVOLUTION

Normal skin's curiously abnormal genome

Within every tumor, a battle is being waged. As individual tumor cells acquire new mutations that promote their survival and growth, they clonally expand at the expense of tumor cells that are "less fit." Martincorena *et al*. sequenced 234 biopsies of sun-exposed but physiologically normal skin from four individuals (see the Perspective by Brash). They found a surprisingly high burden of mutations, higher than that of many tumors. Many of the mutations known to drive the growth of cutaneous squamous cell carcinomas were already under strong positive selection. More than a quarter of normal skin cells carried a driver mutation, and every square centimeter of skin contained hundreds of competing mutant clones. — PAK

Science, this issue p. 880; see also p. 867

CYSTIC FIBROSIS

Skirting quality control to treat cystic fibrosis

Patients with cystic fibrosis (CF) have fluid and mucus buildup in their lungs because of mutations that cause misfolding, intracellular retention, and degradation of the cystic fibrosis transmembrane conductance regulator (CFTR). Although drugs can improve the cell surface delivery of mutant CFTR proteins, which are usually partially functional, cells still degrade the mutant CFTR. Loureiro *et al*. found that increasing the interaction between the scaffold protein NHERF1 and mutant CFTR prevented mutant CFTR from being marked for degradation. These manipulations increased the levels of partially functional CFTR on the surface of cultured lung epithelial cells from CF patients. — LKF

Sci. Signal. **8**, ra48 (2015).

ORGANIC CHEMISTRY

Stitching C-N bonds from nitro groups

Numerous compounds in pharmaceutical research have carbon-nitrogen bonds, and chemists are always looking for ways to make them more efficiently. Gui *et al*. present a method that links the carbon in an olefin to the nitrogen in a nitroaromatic compound (see the Perspective by Kürti). Nitroaromatics are readily available, and the method tolerates a wide range of other chemical groups present on either reacting partner. — JSY

Science, this issue p. 886; see also p. 863

SANITATION SUBSIDIES

Helping the poor invest in sanitation

Almost a third of the world's people do not have access to hygienic latrines. Improving access to and increasing the use of latrines would reduce deaths and poor health caused by diarrheal disease. Guiteras *et al.* tested the relative benefits of supplying health information, offering a financial subsidy to purchasers of hygienic latrines, or increasing the availability of latrines for purchase. Providing the subsidy worked best:

ALSO IN *SCIENCE* JOURNALS

Edited by **Stella Hurtley**

INNATE LYMPHOID CELLS

Cells acting at the intersection of immunity

For years, scientists divided the immune system into two arms: innate and adaptive. The cell types involved in the two arms differ in specificity and in how quickly they respond to infections. More recently, immunologists discovered a family of immune cells termed "innate lymphoid cells," which straddle these two arms. Eberl *et al.* review current understanding of innate lymphoid cells. Like innate immune cells, they respond to infection quickly and do not express antigen receptors; however, they secrete a similar suite of inflammatory mediators as T lymphocytes. Better understanding of the processes regulating these cells may allow for their therapeutic manipulation. — KLM

Science, this issue p. 879

CARBON CYCLE

The difference is found at the margins

The terrestrial biosphere absorbs about a quarter of all anthropogenic carbon dioxide emissions, but the amount that they take up varies from year to year. Why? Combining models and observations, Ahlström *et al*. found that marginal ecosystems—semi-arid savannas and low-latitude shrublands—are responsible for most of the variability. Biological productivity in these semi-arid regions is water-limited and strongly associated with variations in precipitation, unlike wetter tropical areas. Understanding carbon uptake by these marginal lands may help to improve predictions of variations in the global carbon cycle. — HJS

Science, this issue p. 895

MICROBIOLOGY

Why methanol-oxidizing bacteria love lanthanides

Although the lanthanide elements are not rare in Earth's crust, they are highly insoluble and difficult to separate. A biological role for these elements has therefore seemed implausible, but recent findings challenge this belief. In a Perspective, Skovran and Martinez-Gomez explain that some methanol-using bacteria contain an enzyme for methanol oxidation that is active only when lanthanide ions are present in the growth medium. Related enzymes have been found in other bacteria, suggesting a wider role of lanthanides in bacterial methanol oxidation. Further insight into the biological role of lanthanides may help toward developing bioremediation for lanthanide mining sites or allow the growth of new species in the lab. — JFU

Science, this issue p. 862

EPIGENETICS

Chromatin state and the single cell

Identifying the chromatin state of any single cell, which may or may not have a different function or represent different stages relative to others collected within any single culture, experiment, or tissue, has been challenging. Cusanovitch *et al*. skirted previously identified technological limitations to identify regions of accessible chromatin at single-cell resolution. Combinatorial cellular indexing, a strategy for multiplex barcoding of thousands of single cells per experiment, was successfully used to investigate the genome-wide chromatin accessibility landscape in each of over 15,000 single cells. — LMZ

Science, this issue p. 910

VIROLOGY

A viral DNA form that survives extremes

The prokaryote *Sulfolobus islandicus* lives at extreme temperatures (~80°C) and acidity (pH 3). It is infected by the rudivirus SIRV2. DiMaio *et al.* determined the structure of the SIRV2 virus using cryo–electron microscopy to understand how the virus survives these brutal conditions. Most DNA in nature assumes a B-form shape. The virion, on the other hand, contains highly unusual A-form DNA that may help it survive adverse conditions. The viral capsid protein forms an extended α-helical structure that wraps around the viral DNA, possibly stabilizing the A-form DNA. — GR

Science, this issue p. 914

RNA STRUCTURE

Structural signals that direct HIV packaging

During the viral replication cycle of HIV, unspliced dimeric RNA genomes are efficiently packaged into new virions at the host cell membrane. Packaging is directed by a region at the start of the genome, the 5′ leader. The architecture of the 5′ leader remains controversial. Keane *et al*. developed nuclear magnetic resonance methods to determine the structure of a 155-nucleotide-long region of the 5′ leader that can direct viral packaging. The structure shows how the 5′ leader binds to the HIV protein that directs packaging, how unspliced dimeric genomes are selected for packaging, and how translation is suppressed when the genome dimerizes. — VV

Science, this issue p. 917

Nonsubsidized households were more likely to purchase latrines when other households in their village were subsidized. — GJC

Science, this issue p. 903

CANCER

Marrow-infiltrating lymphocytes in ACT

Adoptive T cell therapy (ACT) has had success in treating some types of cancer, but widespread use is limited in part by a lack of tumor-specific targets. Tumor-infiltrating T cells may overcome this limitation for solid tumors. Noonan *et al.* performed a phase I clinical trial and showed that bone marrow can be a source of ACT for hematologic malignancies such as multiple myeloma. Marrow-infiltrating lymphocytes provided myeloma-specific immunity in the bone marrow for up to 1 year after ACT, and increased progression-free survival. — ACC

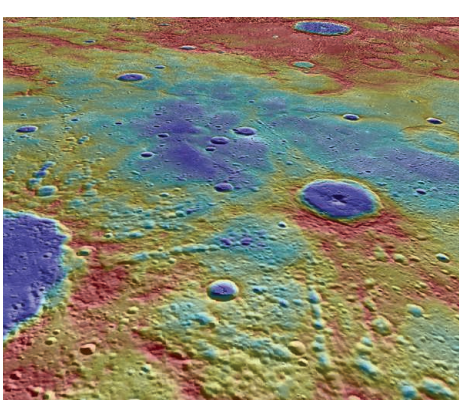
Sci. Transl. Med. **7**, 288ra78 (2015).

PLANETARY SCIENCE

Old minerals expose an ancient field

Mercury is the only terrestrial planet other than Earth with an active, internally generated magnetic field. Results from the MESSENGER spacecraft indicate that the field is almost as old as the planet. Johnson *et al.* took advantage of close flybys to extract evidence of an ancient magnetic field. Certain minerals are able to "lock in" the signature of a field at the time they crystallize. This remnant magnetization was found in a region on Mercury believed to be 3.8 billion years old. — BG

Science, this issue p. 892



MESSENGER's view of Mercury

NEUROPHYSIOLOGY

Brain imagination to control external devices

Studies in monkeys have implicated the brain's posterior parietal cortex in high-level coding of planned and imagined actions. Aflalo *et al.* implanted two microelectrode arrays in the posterior parietal cortex of a tetraplegic patient (see the Perspective by Pruszynski and Diedrichsen). They asked the patient to imagine various types of limb or eye movements. As predicted, motor imagery involved the same types of neural population activity involved in actual movements, which could potentially be exploited in prosthetic limb control. — PRS

Science, this issue p. 906; see also p. 860

EVOLUTION

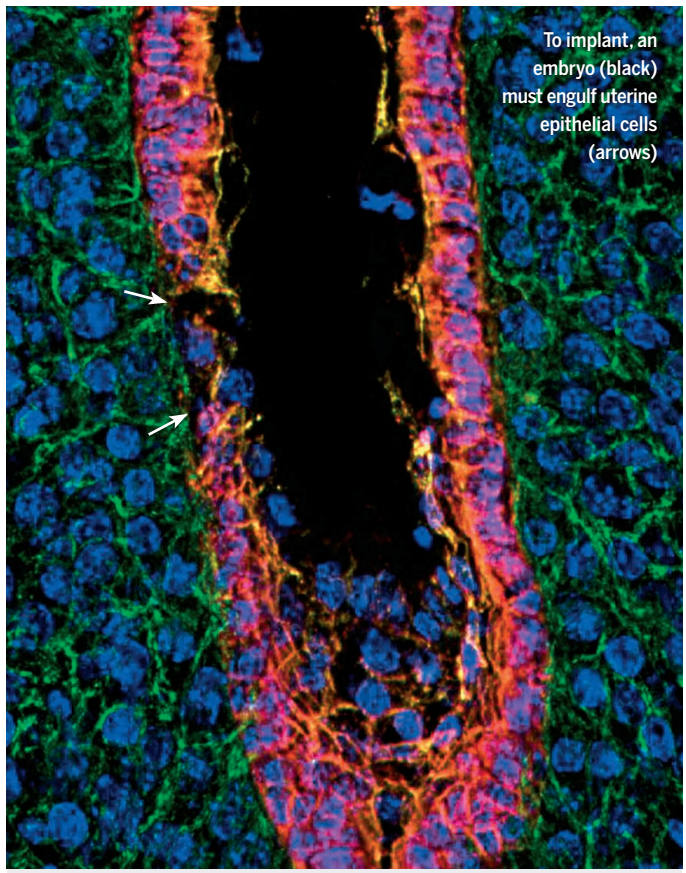
Staying the same across a billion years

How far across evolution do families of genes retain their function? Yeast and humans are separated by roughly a billion years of evolutionary history, and yet genes from one can substitute for orthologous genes in the other. To study this effect systematically, Kachroo *et al.* replaced over 400 essential yeast genes with their human orthologs. Roughly half of the human genes could functionally replace their yeast counterparts. Genes being in the same pathway was as important as sequence or expression similarity in determining replaceability. — GR

Science, this issue p. 921

IN OTHER JOURNALS

Edited by **Kristen Mueller** *and* **Jesse Smith**



To implant, an embryo (black) must engulf uterine epithelial cells (arrows)

IMPLANTATION

Embryos engulf mom to latch on

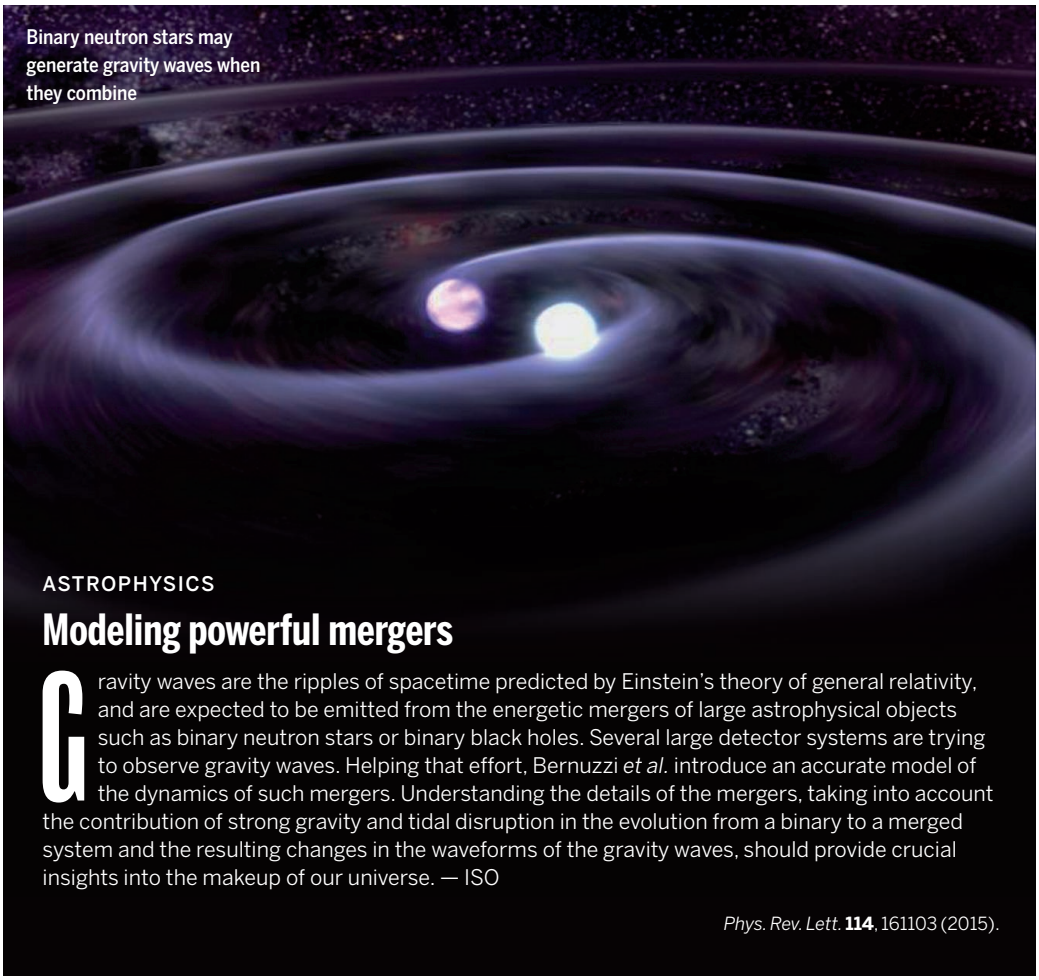
In mammals, to ensure a viable pregnancy, a developing embryo must implant into the wall of the uterus. Previous studies suggested that this depended on maternal uterine epithelial cells dying by apoptosis, a form of programmed cell death. However, Li *et al*. now report in mice that cells from the developing embryo actively engulf live cells of the uterine epithelial barrier, in a process called entosis. This then allows the developing embryo to anchor itself to the uterine stromal bed. Although scientists had previously reported a role for entosis in cancer, these results suggest that this process may be more widespread. — BAP

Cell Rep. **11**, 358 (2015).

BIOGEOGRAPHY

Unevenly blowing in the wind

Scientists, including Charles Darwin, first reported airborne microbes nearly two centuries ago. Many of these organisms cannot be cultured, and only recently have molecular approaches allowed scientists to begin to identify them. To better understand the distribution of airborne fungi, Barberán *et al.* examined dust samples collected from homes across



Binary neutron stars may generate gravity waves when they combine

ASTROPHYSICS

Modeling powerful mergers

Gravity waves are the ripples of spacetime predicted by Einstein's theory of general relativity, and are expected to be emitted from the energetic mergers of large astrophysical objects such as binary neutron stars or binary black holes. Several large detector systems are trying to observe gravity waves. Helping that effort, Bernuzzi *et al*. introduce an accurate model of the dynamics of such mergers. Understanding the details of the mergers, taking into account the contribution of strong gravity and tidal disruption in the evolution from a binary to a merged system and the resulting changes in the waveforms of the gravity waves, should provide crucial insights into the makeup of our universe. — ISO

Phys. Rev. Lett. **114**, 161103 (2015).

the United States. They found impressive microbial diversity in them, with only about a quarter of species being known. Some fungi exhibited strong geographic patterns, such as the allergy-triggering *Alternaria* spp. in the Great Plains and *Cladosporium* in humid regions. Cities showed more homogeneous distributions. — CA

Proc. Natl. Acad. Sci. U.S.A. **112**, 5756 (2015).

SCIENCE AND THE PUBLIC

Citizen scientists fight an oak killer

Sudden oak death (SOD), caused by a fungus-like pathogen, has killed millions of trees in California and Oregon. In a recent example of the value of citizen science for both research and the public good, Meentemeyer *et al.* showed that the involvement of trained volunteers for the past 6 years enabled researchers to learn more about the spread of the disease, build predictive maps of disease risk, and provide decision-makers with information that could help prioritize efforts. High-school students, teachers, and others used a symptom detection guide and a mobile mapping tool and then sampled leaves for analysis. Amateurs equaled professionals in their ability to recognize infected leaves. — BJ

Front. Ecol. Environ. **13**, 189 (2015).

Citizen scientists at work

FLUID DYNAMICS

Uncool heat pipes in microgravity

Heat pipes are efficient heat transfer systems commonly used to cool things such as microprocessors. Heat pipes have a hot end that evaporates liquid, which flows as vapor to a cold end that condenses it. The liquid then normally returns to the hot end through capillary action, completing a circuit with a net cooling effect, although the hot end commonly dries out, lowering the performance of the device—at least on Earth. Kundan *et al.* investigated how heat pipes work in the microgravity of the International Space Station. Surprisingly, on the station, the hot end quickly floods, because of changes in surface tension caused by the lower gravity. This observation suggests that heat pipes will have different performance limitations in space. — BG

Phys. Rev. Lett. 10.1103/PhysRevLett.114.146105 (2015).

PHYSICS

Surprises in spiral domains

Antimony telluride (Sb_2Te_3), a semiconductor with thermoelectric applications, has a layered hexagonal close-packed structure. Hauer *et al*. grew Sb_2Te_3 platelets using a solvothermal technique that developed a spiral growth pattern around a screw dislocation. Scattering-type scanning near-field microscopy of mid-infrared reflectivity surprisingly revealed triangular domains of opposite phase that were not seen with platelets grown by other methods. They attribute the contrast to growth twins that had different levels of antisite defects, which act as electronic dopants and affect its plasma frequency. — PDS

Nano Lett. 10.1021/nl503697c (2015).

PSYCHOLOGY

Judgments that lead to job offers

Job seekers often need to send out hundreds of resumes in order to get a handful of interviews. But do applicants really need to meet their potential employers face-to-face for the best chance of success? Schroeder and Epley investigated this by having business school students or actors apply for jobs by composing elevator pitches for delivery via text and audio or video recordings, and museum visitors or professional recruiters judged the candidates' intellect and their likelihood of being hired. In all of the combinations, the audio pitches outperformed written ones and did just as well as the videos, suggesting that a person's voice is the key. — GJC

Psychol. Sci. **26**, 10.1177/0956797615572906 (2015).

CREDITS: (TOP TO BOTTOM) NASA; DOUGLAS SCHMIDT/UC BERKELEY

REVIEW SUMMARY

INNATE LYMPHOID CELLS

Innate lymphoid cells: A new paradigm in immunology

Gérard Eberl,*† Marco Colonna, James P. Di Santo, Andrew N. J. McKenzie

BACKGROUND: Innate lymphoid cells (ILCs) are a growing family of immune cells that mirror the phenotypes and functions of T cells. Natural killer (NK) cells can be considered the innate counterparts of cytotoxic $CD8^+$ T cells, whereas ILC1s, ILC2s, and ILC3s may represent the innate counterparts of $CD4^+$ T helper 1 (T_H1), T_H2, and T_H17 cells. However, in contrast to T cells, ILCs do not express antigen receptors or undergo clonal selection and expansion when stimulated. Instead, ILCs react promptly to signals from infected or injured tissues and produce an array of secreted proteins, termed cytokines, that direct the developing immune response into one that is adapted to the original insult. Thus, the power of ILCs may be controlled or unleashed to regulate or enhance immune responses in disease prevention and therapy.

ADVANCES: As with B cells and T cells, ILCs develop from the common lymphoid progenitor, but dedicated transcription factors supress the B and T cell fates and direct the generation of the different types of ILCs. ILC precursors may migrate from their primary site of production into infected and injured tissues, where they complete their maturation, similar to the differentiation of naïve T cells into T_H effectors. Cytokines produced by local cells as well as stress ligands and bacterial and dietary compounds regulate the maturation and activation of ILCs into effectors that play a major role in early immune responses to pathogens and symbionts, helminths, and allergen. The cytokines they produce induce innate responses in stromal, epithelial, and myeloid cells and regulate the activity of dendritic cells (DCs), which play a central role in the cross-talk between ILCs and T cells. In particular, ILCs activate tissue-resident DCs to migrate to lymph nodes, where they elicit specific T cell responses, which in turn regulate ILCs. ILCs also directly regulate T cells through the presentation of peptide antigens on major histocompatibility complex II. However, ILCs are also involved in immunopathology, during which their production of cytokines exacerbates the inflammatory process.

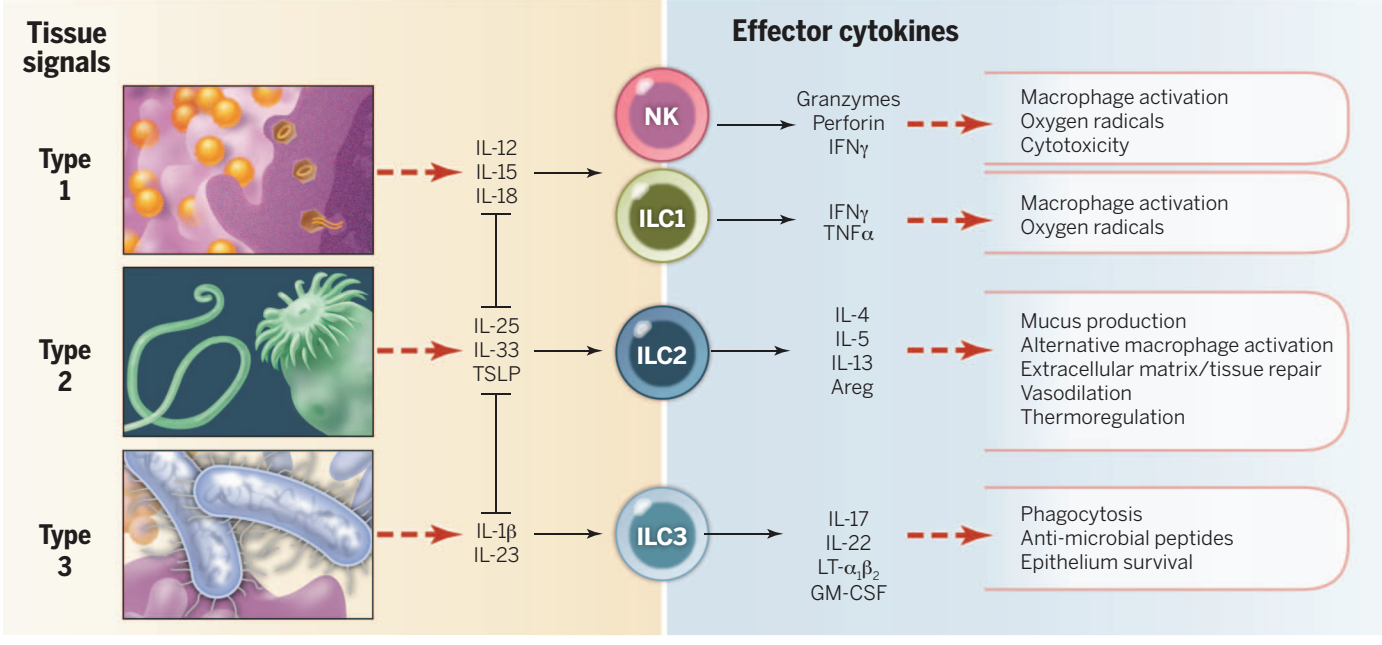
ILCs also play an intriguing role beyond immunity. In adipose tissues, they regulate thermogenesis and prevent local inflammation that may lead to metabolic syndrome, insulin resistance, and obesity-associated asthma. The functions of ILCs in host metabolism are a new area of research that will lead to insights into how the immune system is implicated in host functions not directly related to defense. Furthermore, ILCs are involved in repair responses upon infection and injury of epithelial cells, stromal cells, and stem cells.

ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.aaa6566

OUTLOOK: A logical next step will be the identification of molecules that allow manipulation of ILCs and the orchestration of the optimal immune response after vaccination and immunotherapy—or in contrast, to block detrimental responses. The combination of a prompt activation of ILCs with both effector and regulatory functions, with the expansion of antigen-specific B and T cells, should lead to new and powerful avenues in clinical immunology.

*The list of author affiliations is available in the full article online.
†Corresponding author. E-mail: gerard.eberl@pasteur.fr
Cite this article as G. Eberl *et al.*, *Science* **348**, aaa6566 (2015). DOI:10.1126/science.aaa6566



Signals from injured or infected tissues expand and activate NK cells, ILC1s, ILC2s, and ILC3s. The effector functions of ILCs mirror the functions of $CD8^+$ and $CD4^+$ T cells, with the major difference being the prompt activation of ILCs and their lack of (relatively slow) antigen-dependent clonal selection and expansion.

REVIEW

INNATE LYMPHOID CELLS

Innate lymphoid cells: A new paradigm in immunology

Gérard Eberl,[1]* Marco Colonna,[2] James P. Di Santo,[3] Andrew N. J. McKenzie[4]

Innate lymphoid cells (ILCs) are a growing family of immune cells that mirror the phenotypes and functions of T cells. However, in contrast to T cells, ILCs do not express acquired antigen receptors or undergo clonal selection and expansion when stimulated. Instead, ILCs react promptly to signals from infected or injured tissues and produce an array of secreted proteins termed cytokines that direct the developing immune response into one that is adapted to the original insult. The complex cross-talk between microenvironment, ILCs, and adaptive immunity remains to be fully deciphered. Only by understanding these complex regulatory networks can the power of ILCs be controlled or unleashed in order to regulate or enhance immune responses in disease prevention and therapy.

During hematopoiesis, the common lymphoid progenitor (CLP) gives rise to antigen receptor–bearing T and B lymphocytes. Until quite recently, only two types of lymphoid cells had been recognized as deriving from CLPs but devoid of any antigen receptors. The first of these cells were the natural killer (NK) cells, which complement the cytotoxic $CD8^+$ T cells in killing infected, stressed, or transformed cells (*1*). The second were lymphoid tissue inducer (LTi) cells, which induce the development of lymph nodes and Peyer's patches (*2*, *3*). However, since 2008 the world of lymphoid cells has expanded dramatically. LTi-like cells were found that also express markers associated with NK cells and were termed NK22 cells, or natural cytotoxicity receptor 22 (NCR22) cells, for their concomitant expression of the cytokine interleukin-22 (IL-22) (*4*–*7*). Natural helper cells and nuocytes were described that expand in response to helminth infection and promote anti-worm and pro-allergic type 2 immune responses (*8*, *9*). Last, noncytotoxic NK-like cells were isolated from the intestinal epithelium (*10*, *11*). To avoid chaos in diversity, it was decided to reunite all these cells into one family of "innate lymphoid cells," or ILCs, and to create three categories—ILC1s, ILC2s, and ILC3s—that reflect the cytokine expression profiles of the classical $CD4^+$ T helper (T_H) cell subsets T_H1, T_H2, and T_H17 cells (Box 1) (*12*).

ILCs share the developmental origin and many of the phenotypes and functions of T cells. However, ILCs are activated by stress signals, microbial compounds, and the cytokine milieu of the surrounding tissue, rather than by antigen, in ways similar to the activation of memory or "innate" T cells, such as invariant NKT cells and subsets of γδ T cells. This mode of activation makes ILCs highly reactive and early effectors during the immune response. Furthermore, ILCs express the effector cytokines normally associated with T helper cells, and therefore, ILCs are expected to play a central role in the regulation of type 1, type 2, and type 3 (or T_H17 cell) responses, which control intracellular pathogens, large parasites, and extracellular microbes, respectively. The activity of ILCs may thus be harnessed to enhance responses against pathogens and tumors, during vaccination and immunotherapy, or inhibited to prevent autoimmune or allergic inflammation. Recent data also show that the role of ILCs extends beyond immunity into physiology through the regulation of fat metabolism and body temperature (*13*–*15*). In this Review, we discuss these intriguing issues in the light of the most recent developments.

Development and evolution of ILCs

Developing away from adaptive lymphocyte fate

ILCs develop from CLPs that give rise to B cell and T cell precursors, NK cell precursors (NKPs), and the recently described common helper ILC precursors (ChILPs) that express Id2 and variable levels of promyelocytic leukemia zinc finger (PLZF) (Fig. 1) (*16*–*18*). ChILPs generate all ILC groups but not NK cells, whereas $PLZF^+$ ILC precursors generate all ILC groups but not NK

Box 1. Warning: the limits of nomenclature.

The classification of ILCs into ILC1s, ILC2s, and ILC3s reflects both the phenotypical and the functional characteristics of T_H cells and serves to structure research into their phylogeny and functions. However, this classification also generates some debates because ILCs and T_H cells can coexpress cytokines of more than one type. For example, ILC3s and T_H17 cells are found to coexpress IFN-γ and IL-17—which are characteristic of type 1 and type 3 responses, respectively—during pathological inflammation (*56*, *103*, *128*). How should these cells be referred to, ILC3/1 cells or IFN-γ–expressing ILC3s? Furthermore, ILC3s can evolve into ILC1s by down-regulating the transcription factor RORγt and up-regulating the transcription factor T-bet (*103*, *129*). Therefore, it is possible that IFN-γ–expressing ILC3s are in fact cells that transit from an ILC3 phenotype to an ILC1 phenotype—"so-called ex-ILC3s." To further complicate an already opaque ILC world, a potential ILC2 precursor that is induced by IL-25 has been reported to have the capability to give rise to ILC3-like IL-17 producers, although in naïve mice or upon helminth infection, they appear to default to a more conventional and less plastic ILC2 phenotype (*43*). Last, fate mapping of $PLZF^+$ ILC precursors shows that LTi cells develop along a pathway distinct from that of the other types of ILCs (*17*). In addition, LTi cells and $NKp46^+$ ILC3s can be distinguished on the basis of their gene expression (*106*). This difference may have an evolutionary basis: because the programmed development of lymph nodes and Peyer's patches is induced by LTi cells only in mammals (*130*), LTi cells may be a recent acquisition, whereas ILCs may have appeared with the advent of vertebrates or even before (*49*).

NK cells present another difficulty for classification. NK cells express T-bet and produce IFN-γ and thus are type 1 cells such as T_H1 cells. However, they also express Eomesodermin-dependent perforin and granzymes, as do cytotoxic $CD8^+$ T cells. It is therefore suggested that NK cells mirror $CD8^+$ T cells, whereas ILC1s mirror $CD4^+$ T_H1 cells (*16*, *131*). Thus, NK cells may be termed "cytotoxic ILCs." Distinguishing NK cells from ILC1s can be achieved by fate-mapping of $Id2^+$ or $PLZF^+$ precursor cells (*16*, *17*) or by using Eomesodermin reporter mice. However, it is more difficult to discriminate these two ILC subsets by using surface markers because they vary from tissue to tissue. For example, discriminating the two cell types is relatively straightforward in the liver but more difficult in the spleen and small intestine (*106*). In the liver, ILC1s selectively express TRAIL and VLA1. In the spleen and small intestine, there are no distinctive surface markers identified, although the expression of CXCR6 on ILC1s and of the MHC class I receptors Ly49 and KIRs on NK cells can be partially informative. Last, surface markers used to discriminate these cell types may vary depending on cellular activation.

[1]Institut Pasteur, Microenvironment and Immunity Unit, 75724 Paris, France. [2]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA. [3]Institut Pasteur, Innate Immunity Unit, INSERM U668, 75724 Paris, France. [4]Medical Research Council (MRC) Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge CB2 0QH, UK.
*Corresponding author. E-mail: gerard.eberl@pasteur.fr

cells or LTi cells. ILC development from CLP (via NKP or ChILP) therefore involves a stage of lineage restriction, in which B and T cell potentials are lost and ILC potential is reinforced. This is achieved through the coordinated expression of specific transcription factors that activate or repress target genes that are critical for subset-specific lymphocyte differentiation. For ILC development, several transcription factors have been shown to be critical at the ILC precursor stage, including Id2, Nfil3, and Gata3 (*19*–*24*). Our understanding of how these transcription factors promote ILC fate is incomplete, but one emerging concept involves obligate suppression of alternative lymphoid cell fates, on the basis of reciprocal repression as a means to control binary cell fate decisions. Id2 is a transcriptional repressor that acts to reduce the activity of E-box transcription factors (E2A, E2-2, and HEB), which are critical in early B and T cell development. Thus, increasing expression of Id2 in CLP promotes ILC development at the expense of the B and T cell fates (*20*, *25*). Accordingly, NKP and ChILP express variable levels of Id2, whereas CLPs do not express Id2 (*16*, *26*). In a similar fashion, Gata3 represses B cell fate by blocking EBF1 and thereby facilitates T and ILC differentiation from CLPs (*23*, *24*, *27*).

How Id2 or Gata3 expression is controlled as CLPs differentiate into NKP or ChILP is not fully understood. Signals produced by the microenvironment—for example, bone morphogenic proteins (BMP) and Notch ligands (*28*, *29*)—regulate Id2 expression, a mechanism that could apply to CLPs. Furthermore, the transcription factor Nfil3 links the peripheral circadian clocks involving the nuclear receptor Rev-ERBα to gene regulation (*30*), and its deletion affects multiple developmental processes within the hematopoietic system. In particular, Nfil3 controls differentiation of ILC via Id2 and the transcription factors RAR-related orphan receptor–γt (RORγt), Eomesodermin, and Tox (*21*, *22*, *31*). In addition, soluble factors, including cytokines, regulate Nfil3 expression (*32*), providing a link between signals from the tissue and fate decisions into the ILC lineages.

Do ILCs complete development in response to local cues?

Conventional wisdom suggests that the primary site of ILC development is the liver in the fetus, and the bone marrow after birth, because these primary lymphoid organs harbour CLP, NKP, and ChILP (*16*, *33*, *34*). Once generated, mature ILCs exit these sites, circulate in the blood, and enter tissues following codes based on adhesion molecules and chemokines, similar to the ones used by T cells. This model is supported by the dearth of tissue-resident ILCs under steady-state conditions, with the exception of mucosal sites, and the rapid recruitment of ILCs after infection or injury. However, ILC precursors—the NKP and the ChILP—may leave the fetal liver or the bone marrow and complete their maturation in response to local signals, much in the same way as naïve T cells differentiate into the different effector subsets during inflammation. In this view, ILC precursors would be the innate homologs of naïve T cells.

In support of this hypothesis, NKP and ILC3 precursors are found in human tonsils (*35*). In mouse, ILC3 precursors are found in the fetal gut (*19*), where their mature progeny induce the development of Peyer's patches, as well as after birth in the lamina propria of the small intestine (*36*). Fetal ILC precursors with the capacity to give rise to ILC1s, ILC2s, and ILC3s are present in the mouse intestine and accumulate in the developing Peyer's patches (*37*). The vitamin A metabolite retinoic acid (RA), produced by many types of cells outside lymphoid organs—including nerve cells (*38*), dendritic cells (DCs) (*39*), and stromal cells (*40*)—favors the maturation of ILC3s at the expense of ILC2s (*41*) and is required for the full maturation of ILC3s in the fetus and the adult (*42*). Furthermore, although IL-25 and IL-33 produced by epithelial cells both promote ILC2 differentiation, it has been proposed that IL-25 may act to expand precursors that retain ILC3 potential (*43*). Last, the aryl hydrocarbon receptor Ahr, which is triggered by ligands from diet, is also required for the maintenance and expansion of intestinal ILC3s after birth (*44*–*46*).

ILCs as evolutionary precursors to T cells

Even though the adaptive lymphocyte fate has to be blocked in CLPs to generate ILCs, striking similarities exist between ILC and T cell differentiation. Gata3, Nfil3, and Tcf1 (*21*–*24*, *47*, *48*) are shared by the precursor common to T cells and ILCs, and the signature transcription factors T-bet, Gata3, and RORγt, which determine the development of type 1, 2, or 3 cells, are highly conserved in both innate and adaptive lymphoid cells in mice and men. It is therefore tempting to propose that ILCs are the evolutionary precursors of T cells, even though definitive evidence has yet to be found that ILCs exist in invertebrates or early vertebrates that lack T or B lymphocytes (*49*). The emergence of ILCs, and thus of the lymphoid lineage, must also have provided a fitness advantage. As we now understand the function of ILCs and T_H cells, this advantage would build on the ability to rapidly direct immunity into type 1, 2, or 3 responses that are adapted to counter specific types of threats. Myeloid cells, as well as nonhematopoietic cells such as epithelial cells and stromal cells, produce cytokines in reaction to infection and injury, which activate a particular ILC subset and the production of effector cytokines. The reason why phagocytic myeloid cells, presumably the first type of immune cells to appear during evolution, would not perform this function is unclear, but may be related to the superior capacity of lymphoid cells to expand rapidly.

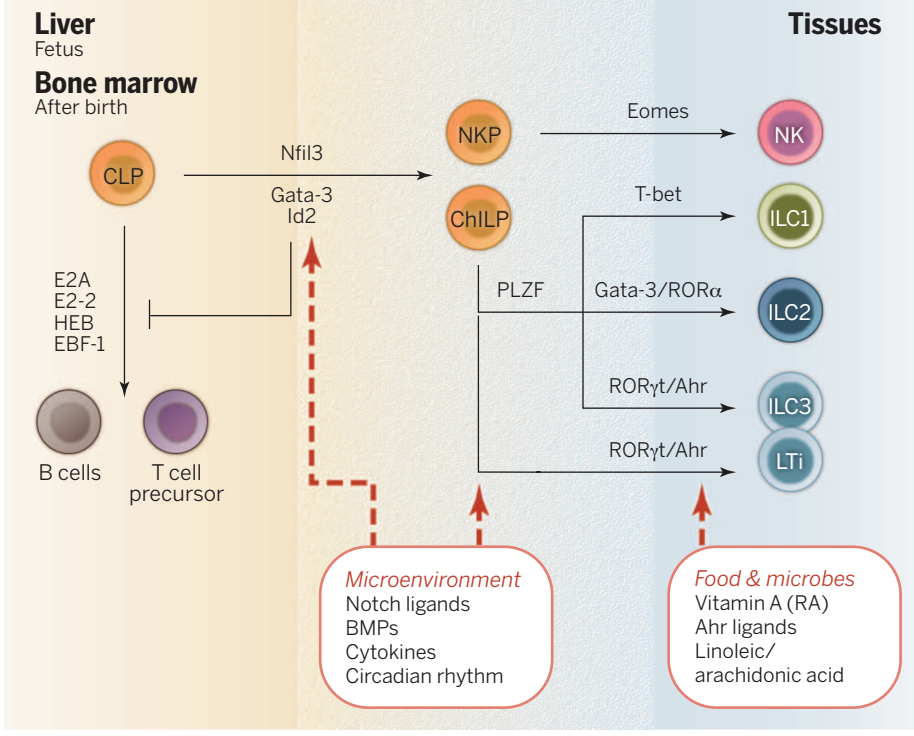


Fig. 1. The development of ILCs. The development of ILCs from common lymphoid progenitors (CLPs) requires Id2-mediated suppression of alternative lymphoid cell fates that generate B and T cells. Factors present in the microenvironment, such as Notch ligands, bone morphogenic proteins (BMPs), and cytokines, as well as the circadian rhythm, control expression of Nfil3, Gata3, and Id2, which determine the progression toward the ILC fate. Distinct precursors give rise to NK cells and ILCs (which, unlike NK cells, are noncytotoxic), while the transcription factor PLZF further divides the progeny of ChILPs into the PLZF-dependent ILC1s, ILC2s, and ILC3s and PLZF-independent LTi cells (although LTi cells tend to be grouped as ILC3s) required for the development of lymph nodes, Peyer's patches, and ILFs. The maturation of ILC precursors into mature ILCs may occur outside of primary lymphoid tissues, in ways similar to the maturation of naïve T_H cells into T_H1, T_H2, T_H17, and regulatory T cells (T_{reg} cells) and in response to a variety of signals produced by the tissue microenvironment.

Once established as a diverse family of innate effector cells, the program of ILC development, differentiation, and function would serve as a "blueprint" for T cells. Emergence of the adaptive arm of the immune system, based on major histocompatibility complex (MHC) restriction and somatic rearrangements of antigen receptor genes, would be layered onto the ILC program, providing an exhaustive range of antigen specificity to the already existing effector cell diversity. Because clonal selection via the T cell receptor results in substantial cellular expansion, T cells may also be freed from the microenvironmental constraints that limit ILC expansion, providing more amplitude to immune effector and regulatory functions, as well as antigen-specific immunological memory.

Activation of ILCs

ILCs translate signal cytokines into effector cytokines

In the absence of adaptive antigen receptors, ILCs react to the microenvironment through cytokine receptors. NK cells and ILC1s expand and secrete interferon-γ (IFN-γ) in response to IL-12, IL-15, and IL-18 produced by myeloid cells as well as by nonhematopoietic cells in response typically to intracellular pathogens (Fig. 2) (*10*, *11*, *16*, *50*). ILC2s, on the other hand, respond to the epithelium-derived cytokines IL-25, IL-33, TSLP (thymic stroma lymphopoietin), basophil-derived IL-4, and products of the arachidonic acid pathway, in response to parasite infection, allergens, and epithelial injury (*8*, *9*, *51*–*53*). Activation of ILC2s leads to the production of high amounts of IL-4, IL-5, and IL-13. Last, ILC3s respond mainly to IL-1β and IL-23 produced by myeloid cells in response to bacterial and fungal infection (*54*–*56*). ILC3s produce lymphotoxins, GM-CSF (granulocyte-macrophage colony-stimulating factor), and IL-22, as well as IL-17 in the fetus, early after birth and during inflammation (*57*, *58*).

Thus, ILCs translate signal cytokines produced by myeloid and nonhematopoietic cells in tissues into effector cytokines that activate local innate and adaptive effector functions. For example, IFN-γ activates the production of microbicidal reactive oxygen species in myeloid cells, induces the production of antibodies for antibody-mediated cytotoxicity, and increases antigen presentation by MHC molecules (*59*). On the other hand, IL-5 induces the recruitment of eosinophils, and IL-13 stimulates the production of mucus by goblet cells [the secretion of which can also be induced by IFN-γ (*60*)] (*61*), whereas IL-17 and IL-22 induce the production of antimicrobial peptides by epithelial cells (*62*) and the recruitment of neutrophils through the expression of CXC chemokines by stromal cells (*63*).

NK cells also express an array of receptors that recognize MHC I, the constant domains of antibodies, and cell-surface molecules associated with cellular transformation, stress, and infection, the activation of which leads to cytotoxicity and the production of IFN-γ (*64*). These NK receptors are not antigen receptors but nevertheless confer some degree of specificity to the reactivity of NK cells. Because individual NK cells express different combinations and levels of NK receptors, triggering of one receptor may lead to the expansion of a subset of NK cells and thus to an increased response, or memory, upon reencounter of the trigger (*65*). Furthermore, a subset of ILC3s expresses the pan-NK marker NKp46 in mouse and NKp44 in human (*4*–*7*). NKp46 appears redundant for ILC3 responses against bacterial infection (*66*), but NKp44 can activate human ILC3s (*67*). Last, ILCs isolated from human tonsils were found to produce IL-5 and IL-13, as well as IL-22, in response to ligands that bind the pattern recognition receptor Toll-like receptor 2 (TLR2) (*68*), indicating that ILCs may also react to microbial compounds. Thus, it is possible that ILCs express different arrays of innate receptors that enable them to react to sets of molecules or proxies for type 1–, 2-, or 3-inducing cellular stresses, injuries or infections. However, although such receptors are well studied for NK cells, they remain to be described for the other types of ILCs.

How diet and the microbiota influence ILC development and activity

As mentioned earlier, the vitamin A metabolite RA is required for full maturation of ILC3s at the expense of ILC2s (*41*, *42*), and food-derived Ahr ligands are required for the maintenance of ILC3s after birth (*44*–*46*). Furthermore, TLR2 ligands can activate human ILC2s and ILC3s in vitro (*68*). That is, however, the state of our knowledge of the direct effects of diet and microbiota on ILCs. In contrast, much more is known on indirect effects of diet and microbes on the activation of ILCs.

In the absence of microbiota in germ-free mice, the activity of ILC3s in the intestine is substantially perturbed. Although the development of lymph nodes and Peyer's patches, induced by LTi cells, is programmed in the fetus, the formation of isolated lymphoid follicles (ILFs) in the intestinal lamina propria after birth is not (*69*). Bacteria are required to trigger the production of β-defensins and the chemokine CCL20 by epithelial cells, which induce the morphogenesis of ILFs through activation of $CCR6^+$ LTi cells

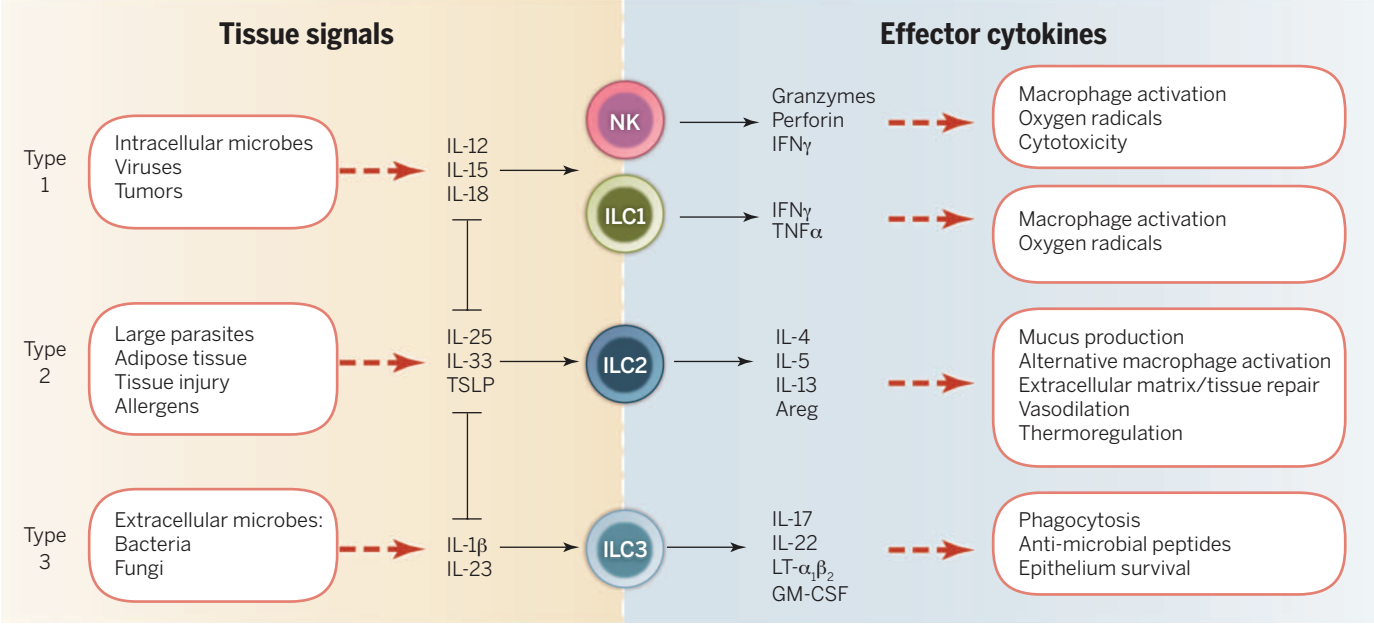


Fig. 2. Activation and functions of ILCs. The tissue signals that expand and activate ILC1s, ILC2s, and ILC3s, and the effector functions of ILCs, mirror the activation and functions of T cells. In this figure, NK cells, ILC1s, ILC2s, and ILC3s could be replaced by $CD8^+$ T cells, T_H1, T_H2, and T_H17 cells, respectively. However, whereas ILCs are activated promptly by tissue signals and therefore act upstream in the immune response, T cells are first selected and expanded on the basis of T cell receptor specificity, a process that typically requires several days.

clustered in so-called cryptopatches (70) and the recruitment of CCR6$^+$ B cells to nascent ILFs (71). The B cell chemoattractant CXCL13, produced by dedicated stromal cells termed "lymphoid stromal cells" (LSCs), is also required for the development of lymphoid tissues through the recruitment of LTi cells from the bloodstream (72) and is induced by RA (38). Furthermore, microbiota induce the expression of CXCL16 by dendritic cells (DCs), which recruits ILC3s to the lamina propria and villi of the small intestine (73). Microbiota also negatively regulate the activity of ILC3s. The expression of IL-17 and IL-22 by ILC3s is highest in the fetus and gradually declines after birth as the intestinal tract is colonized. Microbiota induce the expression of the type 2 cytokine IL-25 by epithelial cells, which activates IL25R$^+$ DCs and the regulation of ILC3s through mechanisms that remain to be elucidated (57).

High-fat diet leads to the build-up of visceral adipose tissue (VAT). Intriguingly, ILC2s are associated with VAT (74) and were originally described as residents of "fat-associated lymphoid clusters" (FALC) on the mesentery (8). The production of IL-5 and IL-13 by ILC2s leads to the recruitment of eosinophils and the generation of alternatively activated macrophages (AAMs) that protect the organism from fat-induced ILC3-mediated inflammatory pathology (74, 75). It is unclear how fat tissue regulates the activation of ILC2s or ILC3s, but this possibly involves metabolites of arachidonic acid, such as prostaglandins and lipoxins, which are respectively activators and inhibitors of ILC2s (76).

Roles of ILCs in immunity

Do ILCs have specific effector functions?

Each cell type in an organism is expected to have a specific function that justifies its evolutionary conservation. However, NK cells, ILC1s, ILC2s, and ILC3s mirror the cytokine production and effector functions of CD8$^+$ T cells, T_H1, T_H2, and T_H17 cells (Fig. 2). Nevertheless, in contrast to T cells, ILCs do not undergo antigen-driven clonal selection and expansion, and therefore, ILCs act promptly like a population of memory T cells. As a consequence, within hours after infection or injury, the effector cytokines IFNγ, IL-5, and IL-13, or IL-17 and IL-22, which can be produced by both ILCs and T cells, are produced mostly by ILCs. In certain tissues, the prompt production of effector cytokines is shared with "innate" T cells, such as mucosa-associated invariant T (MAIT) cells that produce IFN-γ, IL-17, and IL-22 (77); invariant NKT (iNKT) cells that produce IFN-γ or IL-4 (78); and subsets of γδ T cells that produce IFN-γ and IL-17 within different epithelial and mucosal compartments (79–81). Nevertheless, each of these cell types reacts to distinct stimuli. For example, MAIT cells recognize microbial metabolites bound to the MHC-like molecule MR1, and iNKT cells respond to glycolipid moieties bound to the MHC-like molecule CD1d.

Regulation of adaptive immunity by ILCs

Because ILCs are activated early in the immune response to infection and injury, and produce type 1, type 2, and type 3 cytokines, it is expected that they regulate the developing adaptive immune response (82). ILCs have been found to do that in two ways: directly through the expression of MHC class II molecules (MHC II), and indirectly through the regulation of DCs (Fig. 3).

ILC3s were shown nearly two decades ago to express MHC II on their surface (2, 83), but the importance of this expression became clear only recently. ILC3s not only express MHC II but also transcripts for molecules associated with antigen processing and presentation, such as the invariant chain CD74 and the catalyzer of peptide exchange H2-DM, and can process exogenous antigen for presentation to CD4$^+$ T cells (84). In the intestine, ILC3s regulate the activity of T cells specific for microbiota-derived antigens, and as a consequence, the absence of MHC II on ILC3s leads to intestinal inflammation. In contrast, ILC3s activate CD4$^+$ T cells in the spleen upon antigen processing and presentation on MHC II (85). ILC2s also present antigen on MHC II and induce the production of IL-2 and IL-4 by CD4$^+$ T cells, which drive a positive feedback on growth and cytokine production by ILC2s expressing the receptors for IL-2 and IL-4 (86, 87). This dialogue is functionally important as MHC II–deficient ILC2s fail to cause efficient expulsion of parasitic helminths, even in the presence of MHC II$^+$ DCs (86).

ILCs also regulate DCs. The production of IFN-γ by NK cells increases the production of IL-12, IL-15, and IL-18 by DCs, driving a positive feedback loop between NK cells and DCs that promotes the differentiation of T_H1 cells (88). Likewise, the production of IL-13 by ILC2s leads to the activation of DCs, their migration into the draining lymph nodes and the differentiation of T_H2 cells (89). In the absence of ILC2s, the levels of IL-13 are insufficient to instigate the migration of DCs to the lymph nodes in response to lung injury, and T_H2 responses are impaired (89). Last, ILC3s activate DCs through membrane-bound lymphotoxin (LT) $\alpha_1\beta_2$, which in turn produce elevated levels of IL-23, which promotes the activity of ILC3s and the differentiation of T_H17 cells (90), as well as nitric oxide, which activates B cells (91).

Because ILCs promote T cell activation through DCs, it is likely that T cells promote ILC activation through similar mechanisms, establishing positive feedback loops between ILC, T cells, and DCs. However, this cross-talk also provides controls on the activity of ILCs because a decrease in the source of T cell antigen and of signals from the affected tissue should exhaust the positive feedback. In addition, competition between ILC and T cells for common activating cytokines from DCs and the affected tissue may also regulate ILC activity. In agreement with this hypothesis, the activity of ILC3s is increased in the absence of T

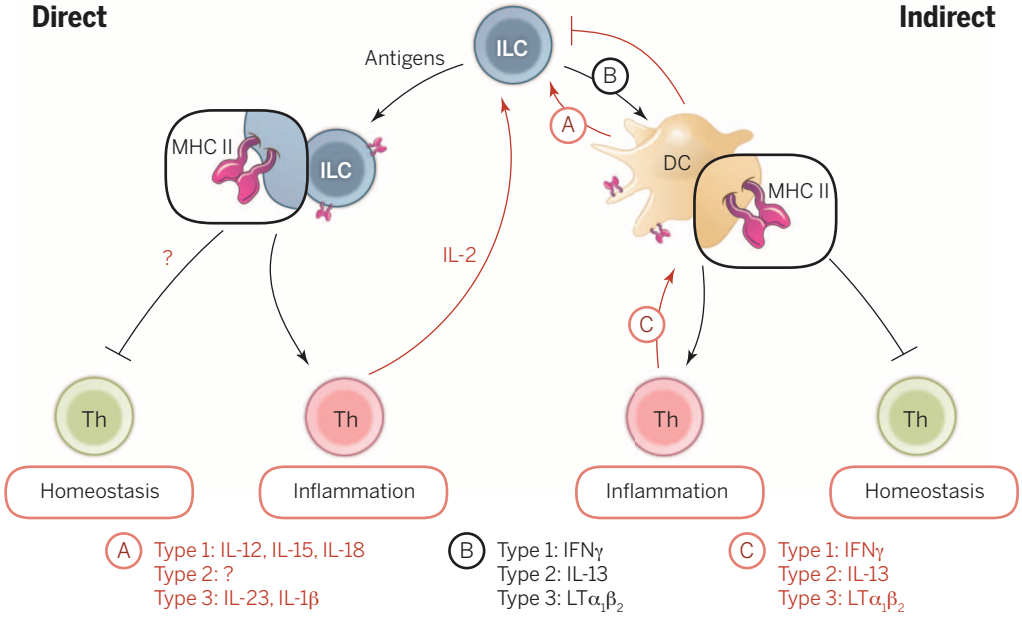


Fig. 3. Regulation of adaptive immunity by ILCs. ILCs regulate T cells both directly through antigen presentation on MHC II, and indirectly through the regulation of DCs. The cross-talk between ILCs, DCs, and T cells establishes a complex regulatory network involving positive and negative feedbacks, the dynamics of which remain to be elucidated. The mechanisms by which ILCs repress CD4$^+$ T_H cell activation remain unclear but may involve the lack of costimulatory molecules in the context of steady state (84). It also remains unclear how DCs negatively regulate the activity of ILCs (57). Red lines depict feedback loops, and "A," "B," and "C" list the type 1, type 2, or type 3 cytokines involved in a specific cross-talk. ILC3s also activate B cells in the intestine through lymphotoxin-mediated recruitment of T_H cells and activation of dendritic cells (91), as well as marginal-zone B cells in the spleen (132).

cells (57). Furthermore, the dependence of ILC2s on IL-2 raises the possibility that both ILC2s and T cells are regulated by regulatory T cells through the removal of IL-2 from the microenvironment.

ILCs in tissue protective and repair responses

ILC2s are involved in tissue-repair responses through the production of amphiregulin (a ligand of the epidermal growth factor receptor) and IL-13. Upon infection of mouse lungs with the H1N1 influenza virus, ILC2s contribute to tissue repair through the expression of amphiregulin (92). Furthermore, injury to the bile duct, which can lead to severe liver disease, leads to the IL-33–mediated activation of ILC2s that promote cholangiocyte proliferation and epithelial restoration through the release of IL-13 (93). In VAT, IL-13 production by ILC2s protects from fat-induced inflammation promoted by ILC3s, which leads to metabolic syndrome, insulin resistance, and diabetes (74). More generally, IL-13 leads to the recruitment of eosinophils and the generation of AAMs (75) and promotes the production of extracellular matrix by stroma cells and mucus by epithelial cells, mechanisms involved both in repair responses and in defense against large parasites (94).

ILC3s promote tissue protective and repair responses through the production of $LT\alpha_1\beta_2$ and IL-22. Infection of lymph nodes with lymphocytic choriomeningitis virus leads to the destruction of lymphoid stromal cells (LSCs). ILC3s restore LSCs through $LT\alpha_1\beta_2$ and activation of LTβ receptor on LSCs (95). IL-22 has a general role in protecting epithelial cells, mostly through the activation of antiapoptotic pathways. In a model of graft-versus-host disease (GvHD), ILC3s protect intestinal epithelial stem cells from GvHD-induced cell death (96). In that context, a subset of ILC3s resists full-body irradiation and provides IL-22 to the stem cells. A similar ILC3-mediated mechanism was found to protect the thymus from the consequences of full-body irradiation (97). IL-22 also protects hepatocytes from acute liver inflammation, but the source of IL-22 was, at the time, attributed to T_H17 cells (98). The source of IL-22 was later recognized to include ILC3s in the $CD45RA^+$ cell transfer model of colitis (99).

ILCs and fat: Roles beyond immunity?

Adipose tissue is associated with the immune system at several levels. Lymph nodes and lymphoid clusters on the mesentery are embedded in adipose tissue for reasons that remain unclear (8). Type 2 responses, including ILC2s, are required to avoid the induction of type 3 responses that lead to metabolic syndrome, insulin resistance, diabetes, as well as obesity-associated asthma (100). In contrast, high-fat diet increases gut permeability and leads to the accumulation of bacteria in VAT, the recruitment and activation of type 1 macrophages, and a shift of the immune response associated with VAT from a protective type 2 to a pathogenic type 3 response (101, 102). Furthermore, ILC2s have recently been shown to regulate thermogenesis from beige fat in a process that appears to involve immune cells beyond immunity (13–15). The sensing of cold by nerves triggers their release of catecholamines that activate the biogenesis and activation of brown adipose tissue (BAT) for thermogenesis. Subcutaneous white adipose tissue (scWAT) can also undergo browning under these circumstances, but its low innervation cannot provide the levels of catecholamines required for the conversion of scWAT into beige fat. Macrophages, however, are recruited to cold-stressed scWAT and produce catecholamines, amplifying the signals released by nerves. This activity of macrophages is dependent on IL-4 produced by eosinophils, as well as on IL-5 and IL-13 produced by ILC2s, replicating the recruitment and activation process induced by ILC2s in VAT. ILC2s also produce methionine-enkephalin peptides, which induce beiging of VAT (15). Last, IL-4 and IL-13 induce the differentiation of adipocyte precursors directly into beige fat (14).

ILCs in pathology

High frequencies of ILC1s are found in Crohn's disease patients and in mouse models of colitis, contributing to the pathology through the production of IFN-γ (10, 11). ILC3s are also associated with inflammatory pathology when producing both IL-17 and IFN-γ during colitis and infection with *Salmonella enterica* (56, 103), as well as with obesity-induced airway hyperreactivity through the production of IL-17 (Fig. 4) (100). The pathogenicity of ILC3s was demonstrated when comparing mice deficient in T and B cells only with those lacking T cells, B cells, and ILCs (56). These studies show that ILC3s can be pathogenic (or sufficient to induce pathology) but nevertheless fail to show that ILC3s are necessary for the development of pathology in the presence of adaptive immunity. The difficulty stems from the lack of mutant mice that lack ILC1s or ILC3s while developing a normal set of T_H1 or T_H17 cells. A chimera system has been established to partially alleviate this difficulty (104). In this system, mature T and B cells are adoptively transferred into Rag-deficient mice, which lack these cell types but develop ILCs. Antibody depletion against a congenic marker depletes ILCs but leaves the T cell compartment intact.

In contrast, the ILC2s field has benefited from RORα-deficient mice that lack ILC2s but not other types of lymphocytes—in particular, T_H2 cells (18, 105). RORα message is also expressed in ILC1s and ILC3s (106) but does not appear to be required for ILC3 development (105). RORα-deficient mice, termed staggerer mice, also develop an undersized cerebellum that translates into behavioral defects (107). Chimeric mice that lack RORα only in the hematopoietic compartment fail to develop acute lung pathology in response to papain, a protease allergen, demonstrating the role of ILC2s in priming the allergic response involving T_H2 cells (89, 105). RORα-deficient mice were further used to show that ILC2s are required to

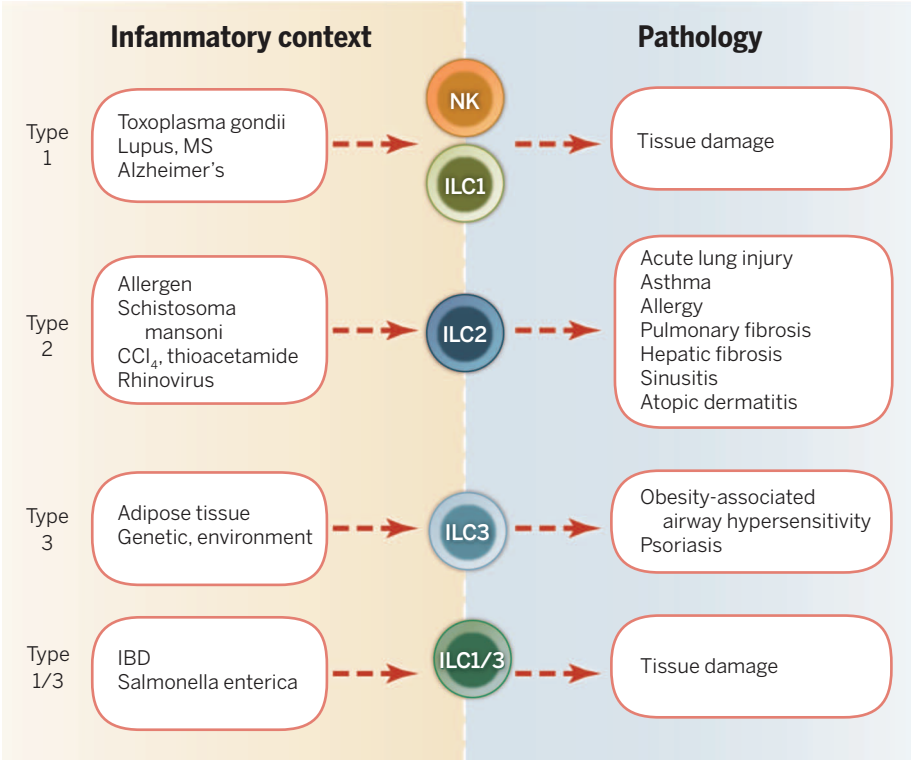


Fig. 4. ILCs in pathology. Pathogens, allergens, chemicals, diet, metabolic states, and genetic factors can induce type 1, type 2, or type 3 inflammatory conditions that lead to pathology involving ILCs. Listed are examples of pathologies shown to involve ILCs, even though in most cases the causative role of ILCs, or their requirement in the pathology, remains to be established. Strong intestinal inflammatory pathology induced during inflammatory bowel disease (IBD) or by *Salmonella enterica* generates ILCs that produce both type 1 (IFN-γ) and type 3 (IL-17) effector cytokines.

expel the helminth *Nyppostrongylus brasiliensis* from the intestine (*18*) and to induce pulmonary fibrosis upon infection with *Schistosoma mansoni* through the production of IL-13 (*108*). The tools available to specifically ablate ILC2s recently expanded after the generation of mice that express the diphtheria toxin receptor (DTR) on ILC2s but not on T cells, allowing for time-controlled ablation of ILC2s (*86*).

ILC2s and IL-13 are also associated with hepatic fibrosis induced in mice by thioacetamide, carbontetrachloride, and *Schistosoma mansoni* (*109*), and with pulmonary fibrosis (*108*), chronic rhinosinusitis (*110*), and atopic dermatitis (*111*, *112*), as well as allergen- (*112*, *113*) and rhinovirus-induced asthma exacerbation in patients (*114*, *115*). Last, ILC2s are proposed to play a central role in asthma-induced obesity. ILC2s in VAT protect from obesity through the release of IL-5 and IL-13 and the recruitment of eosinophils (*74*). However, the accumulation of eosinophils into the asthmatic lungs may prevent their recruitment to VAT and thereby type 2 immunity from protecting the organism from high-fat diet–induced obesity (*116*).

Targeting ILCs for prevention and therapy

Because ILCs act promptly in response to infection and injury, and regulate type 1, type 2, and type 3 responses, they may be targeted to critically enhance or block immune responses early during vaccination, immunotherapy, and inflammatory pathology. Toward this goal, it is imperative that the fundamental molecular signals that regulate ILC diversity and commitment are defined comprehensively. Although ILC-specific targets have not yet been identified, the activation pathways and effector molecules they share with T cells can be targeted early in the immune response. For example, inhibitors of RORγt have been identified primarily to block T_H17-mediated inflammatory pathology, but these inhibitors obviously can be used to block ILC3s as well (*117*, *118*). Similarly, RORα, a nuclear hormone receptor similar to RORγt, may be targeted to modulate ILC2s. Agonists for RORγt and RORα may also be developed to enhance the generation and activity of ILC3s and ILC2s in order to enhance defense against mucosal pathogens or to modulate fat-induced metabolic diseases and allergy. A similar strategy may be followed to modulate the activity of NK cells and ILC1s by targeting T-bet.

The activity of ILC2s is promoted by the arachidonic acid metabolites leukotriene D_4 (LTD_4) and prostaglandin D2 (PGD_2) through the cysteinyl leukotriene receptor 1 (CysLT1R) and the "chemoattractant receptor-homologous molecule expressed on T_H2 cells" CRTH2 (*76*), respectively, but is impaired by the arachidonic metabolites lipoxin A_4 (LXA_4) and maresin-1 (*119*). Thus, an arsenal of lipid mediators, or inhibitors of these mediators (Montelukast, a leukotriene receptor antagonist), may be developed to control the activity of ILCs. The cytokines inducing the development and activity of specific subsets of ILCs—such as IL-12, IL-25 and IL-33, or IL-1β and IL-23 for ILC1s, ILC2s, or ILC3s, respectively, as well as IL-2—may also be targeted, although the precise involvement of ILCs in specific diseases have not been determined within the multifarious effects that arise from blocking these pathways. For example, treatment with Daclizumab, an antibody targeting the IL-2Rα (CD25), of multiple sclerosis patients resulted in a decrease in the frequency of RORγt$^+$ ILCs and an increase in the numbers of NK cells that correlated with drug efficacy (*120*). In addition, Ustekinumab, an antibody directed against the p40 subunit common to IL-12 and IL-23, shows high clinical efficacy against psoriasis (*121*). Furthermore, antibodies against IL-25 and IL-33 have shown efficacy in mouse models of allergic lung inflammation (*122*, *123*), and intravenous antibody to TSLP given before allergen challenge in mild asthmatic patients improves asthma symptoms (*124*). These cytokines can also be blocked by microbial compounds. For example, the excretory/secretory products of the helminth *Heligmosomoides polygyrus* impair the activity of ILC2s in response to airways challenges with extracts of the fungal allergen *Alternaria alternata*, presumably through suppression of the initial *A. alternata*–induced IL-33 production (*125*). Alternatively, microbial compounds may be used to boost one type of ILC in order to block the other types of ILCs. Last, the effector cytokines produced by ILCs may be targeted with antibodies against IFN-γ, IL-5, and IL13, or IL-17. For example, Mepolizumab (antibody to IL-5) and Lebrikizumab (antibody to IL-13) have shown encouraging results in clinical trials against asthma (*126*, *127*).

Concluding remarks

The multiple facets of ILC development, activation, and function need to be further explored before efficient manipulation of ILCs can be achieved in the clinic. The developmental pathways leading to the different types of ILCs appear to be relatively complex, and modulation of these pathways by the microenvironment remains poorly understood, with questions remaining about ILC subset plasticity and stability. It will also be insightful to explore the development of ILCs not only during ontogeny, but also during evolution, in order to assess whether "cytotoxic" ILCs (NK cells) and "helper" ILCs (ILC1s, ILC2s, and ILC3s) served as a blueprint for the appearance of CD8$^+$ cytotoxic and CD4$^+$ T_H cells.

Much remains to be uncovered on the activation and function of ILCs. We propose that ILCs promptly translate signals produced by infected or injured tissues into effector cytokines that activate and regulate local innate and adaptive effector functions. Signals produced by the tissues activating ILCs include cytokines, and possibly also stress ligands and microbial compounds. In terms of function, ILCs and T cells produce similar sets of effector cytokines; however, the hallmark of ILCs is prompt and antigen-independent activation, placing them upstream as probable orchestrators of adaptive responses. Therefore, the cross-regulation of ILCs and T cells, involving DCs as a central platform of information exchange, needs to be deciphered by using new mouse models that allow targeting each cell type individually. Furthermore, a role for ILCs beyond immunity, such as in the regulation of fat metabolism, needs to be unravelled in order to understand the integration of the immune system in host physiology.

Such accumulated knowledge should lead to a new type of immunotherapies based on the manipulation of ILCs. Because ILCs appear to play a major role in adjusting the developing immune response to the original insult, the manipulation of ILCs should allow the optimal shaping of immune responses in prevention and therapy. In the context of immunopathology, the manipulation of ILCs may allow blocking the development of detrimental types of immune responses.

REFERENCES AND NOTES

1. R. Kiessling *et al.*, Killer cells: A functional comparison between natural, immune T-cell and antibody-dependent in vitro systems. *J. Exp. Med.* **143**, 772–780 (1976). doi: 10.1084/jem.143.4.772; pmid: 1082915
2. R. E. Mebius, P. Rennert, I. L. Weissman, Developing lymph nodes collect CD4$^+$CD3$^-$LTβ$^+$ cells that can differentiate to APC, NK cells, and follicular cells but not T or B cells. *Immunity* **7**, 493–504 (1997). doi: 10.1016/S1074-7613(00)80371-4; pmid: 9354470
3. S. Adachi, H. Yoshida, H. Kataoka, S. Nishikawa, Three distinctive steps in Peyer's patch formation of murine embryo. *Int. Immunol.* **9**, 507–514 (1997). doi: 10.1093/intimm/9.4.507; pmid: 9138010
4. M. Cella *et al.*, A human natural killer cell subset provides an innate source of IL-22 for mucosal immunity. *Nature* **457**, 722–725 (2009). doi: 10.1038/nature07537; pmid: 18978771
5. N. Satoh-Takayama *et al.*, Microbial flora drives interleukin 22 production in intestinal NKp46$^+$ cells that provide innate mucosal immune defense. *Immunity* **29**, 958–970 (2008). doi: 10.1016/j.immuni.2008.11.001; pmid: 19084435
6. C. Luci *et al.*, Influence of the transcription factor RORgammat on the development of NKp46$^+$ cell populations in gut and skin. *Nat. Immunol.* **10**, 75–82 (2009). doi: 10.1038/ni.1681; pmid: 19029904
7. S. L. Sanos *et al.*, RORgammat and commensal microflora are required for the differentiation of mucosal interleukin 22-producing NKp46$^+$ cells. *Nat. Immunol.* **10**, 83–91 (2009). doi: 10.1038/ni.1684; pmid: 19029903
8. K. Moro *et al.*, Innate production of T_H2 cytokines by adipose tissue-associated c-Kit$^+$Sca-1$^+$ lymphoid cells. *Nature* **463**, 540–544 (2010). doi: 10.1038/nature08636; pmid: 20023630
9. D. R. Neill *et al.*, Nuocytes represent a new innate effector leukocyte that mediates type-2 immunity. *Nature* **464**, 1367–1370 (2010). doi: 10.1038/nature08900; pmid: 20200518
10. J. H. Bernink *et al.*, Human type 1 innate lymphoid cells accumulate in inflamed mucosal tissues. *Nat. Immunol.* **14**, 221–229 (2013). doi: 10.1038/ni.2534; pmid: 23334791
11. A. Fuchs *et al.*, Intraepithelial type 1 innate lymphoid cells are a unique subset of IL-12- and IL-15-responsive IFN-γ-producing cells. *Immunity* **38**, 769–781 (2013). doi: 10.1016/j.immuni.2013.02.010; pmid: 23453631
12. H. Spits, J. P. Di Santo, The expanding family of innate lymphoid cells: Regulators and effectors of immunity and tissue remodeling. *Nat. Immunol.* **12**, 21–27 (2011). doi: 10.1038/ni.1962; pmid: 21113163
13. Y. Qiu *et al.*, Eosinophils and type 2 cytokine signaling in macrophages orchestrate development of functional beige fat. *Cell* **157**, 1292–1308 (2014). doi: 10.1016/j.cell.2014.03.066; pmid: 24906148
14. M. W. Lee *et al.*, Activated type 2 innate lymphoid cells regulate beige fat biogenesis. *Cell* **160**, 74–87 (2015). doi: 10.1016/j.cell.2014.12.011; pmid: 25543153
15. J. R. Brestoff *et al.*, Group 2 innate lymphoid cells promote beiging of white adipose tissue and limit obesity. *Nature* **519**, 242–246 (2015). doi: 10.1038/nature14115; pmid: 25533952
16. C. S. Klose *et al.*, Differentiation of type 1 ILCs from a common progenitor to all helper-like innate lymphoid cell lineages. *Cell* **157**, 340–356 (2014). doi: 10.1016/j.cell.2014.03.030; pmid: 24725403

17. M. G. Constantinides, B. D. McDonald, P. A. Verhoef, A. Bendelac, A committed precursor to innate lymphoid cells. *Nature* **508**, 397–401 (2014). doi: 10.1038/nature13047; pmid: 24509713
18. S. H. Wong *et al.*, Transcription factor RORα is critical for nuocyte development. *Nat. Immunol.* **13**, 229–236 (2012). doi: 10.1038/ni.2208; pmid: 22267218
19. M. Cherrier, S. Sawa, G. Eberl, Notch, Id2, and RORγt sequentially orchestrate the fetal development of lymphoid tissue inducer cells. *J. Exp. Med.* **209**, 729–740 (2012). doi: 10.1084/jem.20111594; pmid: 22430492
20. Y. Yokota *et al.*, Development of peripheral lymphoid organs and natural killer cells depends on the helix-loop-helix inhibitor Id2. *Nature* **397**, 702–706 (1999). doi: 10.1038/17812; pmid: 10067894
21. C. Seillet *et al.*, Nfil3 is required for the development of all innate lymphoid cell subsets. *J. Exp. Med.* **211**, 1733–1740 (2014). doi: 10.1084/jem.20140145; pmid: 25092873
22. T. L. Geiger *et al.*, Nfil3 is crucial for development of innate lymphoid cells and host protection against intestinal pathogens. *J. Exp. Med.* **211**, 1723–1731 (2014). doi: 10.1084/jem.20140212; pmid: 25113970
23. R. Yagi *et al.*, The transcription factor GATA3 is critical for the development of all IL-7Rα-expressing innate lymphoid cells. *Immunity* **40**, 378–388 (2014). doi: 10.1016/j.immuni.2014.01.012; pmid: 24631153
24. N. Serafini *et al.*, Gata3 drives development of RORγt$^+$ group 3 innate lymphoid cells. *J. Exp. Med.* **211**, 199–208 (2014). doi: 10.1084/jem.20131038; pmid: 24419270
25. M. D. Boos, Y. Yokota, G. Eberl, B. L. Kee, Mature natural killer cell and lymphoid tissue-inducing cell development requires Id2-mediated suppression of E protein activity. *J. Exp. Med.* **204**, 1119–1130 (2007). doi: 10.1084/jem.20061959; pmid: 17452521
26. V. Male *et al.*, The transcription factor E4bp4/Nfil3 controls commitment to the NK lineage and directly regulates Eomes and Id2 expression. *J. Exp. Med.* **211**, 635–642 (2014). doi: 10.1084/jem.20132398; pmid: 24663216
27. M. E. García-Ojeda *et al.*, GATA-3 promotes T-cell specification by repressing B-cell potential in pro-T cells in mice. *Blood* **121**, 1749–1759 (2013). doi: 10.1182/blood-2012-06-440065; pmid: 23287858
28. T. Nakahiro, H. Kurooka, K. Mori, K. Sano, Y. Yokota, Identification of BMP-responsive elements in the mouse Id2 gene. *Biochem. Biophys. Res. Commun.* **399**, 416–421 (2010). doi: 10.1016/j.bbrc.2010.07.090; pmid: 20674548
29. P. Zhang, Y. Zhao, X. H. Sun, Notch-regulated periphery B cell differentiation involves suppression of E protein function. *J. Immunol.* **191**, 726–736 (2013). doi: 10.4049/jimmunol.1202134; pmid: 23752615
30. X. Yu *et al.*, T_H17 cell differentiation is regulated by the circadian clock. *Science* **342**, 727–730 (2013). doi: 10.1126/science.1243884; pmid: 24202171
31. D. M. Gascoyne *et al.*, The basic leucine zipper transcription factor E4BP4 is essential for natural killer cell development. *Nat. Immunol.* **10**, 1118–1124 (2009). doi: 10.1038/ni.1787; pmid: 19749763
32. W. Xu *et al.*, Nfil3 orchestrates the emergence of common helper innate lymphoid cell precursors. *Cell Reports* **10**, 2043–2054 (2015). doi: 10.1016/j.celrep.2015.02.057; pmid: 25801035
33. C. A. Vosshenrich *et al.*, Roles for common cytokine receptor gamma-chain-dependent cytokines in the generation, differentiation, and maturation of NK cell precursors and peripheral NK cells in vivo. *J. Immunol.* **174**, 1213–1221 (2005). doi: 10.4049/jimmunol.174.3.1213; pmid: 15661875
34. S. Sawa *et al.*, Lineage relationship analysis of RORγ+ innate lymphoid cells. *Science* **330**, 665–669 (2010). pmid: 20929731
35. N. K. Crellin, S. Trifari, C. D. Kaplan, T. Cupedo, H. Spits, Human NKp44$^+$IL-22$^+$ cells and LTi-like cells constitute a stable RORC$^+$ lineage distinct from conventional natural killer cells. *J. Exp. Med.* **207**, 281–290 (2010). doi: 10.1084/jem.20091509; pmid: 20142432
36. C. Possot *et al.*, Notch signaling is necessary for adult, but not fetal, development of RORγt$^+$ innate lymphoid cells. *Nat. Immunol.* **12**, 949–958 (2011). doi: 10.1038/ni.2105; pmid: 21909092
37. J. K. Bando, H. E. Liang, R. M. Locksley, Identification and distribution of developing innate lymphoid cells in the fetal mouse intestine. *Nat. Immunol.* **16**, 153–160 (2015). doi: 10.1038/ni.3057; pmid: 25501629
38. S. A. van de Pavert *et al.*, Chemokine CXCL13 is essential for lymph node initiation and is induced by retinoic acid and neuronal stimulation. *Nat. Immunol.* **10**, 1193–1199 (2009). doi: 10.1038/ni.1789; pmid: 19783990
39. D. Mucida *et al.*, Reciprocal T_H17 and regulatory T cell differentiation mediated by retinoic acid. *Science* **317**, 256–260 (2007). pmid: 17569825
40. I. Vicente-Suarez *et al.*, Unique lamina propria stromal cells imprint the functional phenotype of mucosal dendritic cells. *Mucosal Immunol.* **8**, 141–151 (2015). doi: 10.1038/mi.2014.51; pmid: 24938743
41. S. P. Spencer *et al.*, Adaptation of innate lymphoid cells to a micronutrient deficiency promotes type 2 barrier immunity. *Science* **343**, 432–437 (2014). doi: 10.1126/science.1247606; pmid: 24458645
42. S. A. van de Pavert *et al.*, Maternal retinoids control type 3 innate lymphoid cells and set the offspring immunity. *Nature* **508**, 123–127 (2014). doi: 10.1038/nature13158; pmid: 24670648
43. Y. Huang *et al.*, IL-25-responsive, lineage-negative KLRG1(hi) cells are multipotential 'inflammatory' type 2 innate lymphoid cells. *Nat. Immunol.* **16**, 161–169 (2015). doi: 10.1038/ni.3078; pmid: 25531830
44. J. S. Lee *et al.*, AHR drives the development of gut ILC22 cells and postnatal lymphoid tissues via pathways dependent on and independent of Notch. *Nat. Immunol.* **13**, 144–151 (2012). doi: 10.1038/ni.2187; pmid: 22101730
45. J. Qiu *et al.*, The aryl hydrocarbon receptor regulates gut immunity through modulation of innate lymphoid cells. *Immunity* **36**, 92–104 (2012). doi: 10.1016/j.immuni.2011.11.011; pmid: 22177117
46. E. A. Kiss *et al.*, Natural aryl hydrocarbon receptor ligands control organogenesis of intestinal lymphoid follicles. *Science* **334**, 1561–1565 (2011). doi: 10.1126/science.1214914; pmid: 22033518
47. L. A. Mielke *et al.*, TCF-1 controls ILC2 and NKp46$^+$RORγt$^+$ innate lymphocyte differentiation and protection in intestinal inflammation. *J. Immunol.* **191**, 4383–4391 (2013). doi: 10.4049/jimmunol.1301228; pmid: 24038093
48. Q. Yang *et al.*, T cell factor 1 is required for group 2 innate lymphoid cell generation. *Immunity* **38**, 694–704 (2013). doi: 10.1016/j.immuni.2012.12.003; pmid: 23601684
49. G. Eberl, J. P. Di Santo, E. Vivier, The brave new world of innate lymphoid cells. *Nat. Immunol.* **16**, 1–5 (2015). doi: 10.1038/ni.3059; pmid: 25521670
50. E. R. Unanue, Macrophages, NK cells and neutrophils in the cytokine loop of Listeria resistance. *Res. Immunol.* **147**, 499–505 (1996). doi: 10.1016/S0923-2494(97)85214-5; pmid: 9127880
51. S. J. Van Dyken *et al.*, Chitin activates parallel immune modules that direct distinct inflammatory responses via innate lymphoid type 2 and γδ T cells. *Immunity* **40**, 414–424 (2014). doi: 10.1016/j.immuni.2014.02.003; pmid: 24631157
52. Y. Motomura *et al.*, Basophil-derived interleukin-4 controls the function of natural helper cells, a member of ILC2s, in lung inflammation. *Immunity* **40**, 758–771 (2014). doi: 10.1016/j.immuni.2014.04.013; pmid: 24837103
53. B. S. Kim *et al.*, Basophils promote innate lymphoid cell responses in inflamed skin. *J. Immunol.* **193**, 3717–3725 (2014). doi: 10.4049/jimmunol.1401307; pmid: 25156365
54. T. Hughes *et al.*, Interleukin-1β selectively expands and sustains interleukin-22$^+$ immature human natural killer cells in secondary lymphoid tissue. *Immunity* **32**, 803–814 (2010). doi: 10.1016/j.immuni.2010.06.007; pmid: 20620944
55. M. Coccia *et al.*, IL-1β mediates chronic intestinal inflammation by promoting the accumulation of IL-17A secreting innate lymphoid cells and CD4$^+$ Th17 cells. *J. Exp. Med.* **209**, 1595–1609 (2012). doi: 10.1084/jem.20111453; pmid: 22891275
56. S. Buonocore *et al.*, Innate lymphoid cells drive interleukin-23-dependent innate intestinal pathology. *Nature* **464**, 1371–1375 (2010). doi: 10.1038/nature08949; pmid: 20393462
57. S. Sawa *et al.*, RORγt$^+$ innate lymphoid cells regulate intestinal homeostasis by integrating negative signals from the symbiotic microbiota. *Nat. Immunol.* **12**, 320–326 (2011). doi: 10.1038/ni.2002; pmid: 21336274
58. A. Mortha *et al.*, Microbiota-dependent crosstalk between macrophages and ILC3 promotes intestinal homeostasis. *Science* **343**, 1249288 (2014). doi: 10.1126/science.1249288; pmid: 24625929
59. U. Boehm, T. Klamp, M. Groot, J. C. Howard, Cellular responses to interferon-gamma. *Annu. Rev. Immunol.* **15**, 749–795 (1997). doi: 10.1146/annurev.immunol.15.1.749; pmid: 9143706
60. P. Songhet *et al.*, Stromal IFN-γR-signaling modulates goblet cell function during *Salmonella Typhimurium* infection. *PLOS ONE* **6**, e22459 (2011). doi: 10.1371/journal.pone.0022459; pmid: 21829463
61. P. G. Fallon *et al.*, IL-4 induces characteristic Th2 responses even in the combined absence of IL-5, IL-9, and IL-13. *Immunity* **17**, 7–17 (2002). doi: 10.1016/S1074-7613(02)00332-1; pmid: 12150887
62. S. C. Liang *et al.*, Interleukin (IL)-22 and IL-17 are coexpressed by Th17 cells and cooperatively enhance expression of antimicrobial peptides. *J. Exp. Med.* **203**, 2271–2279 (2006). doi: 10.1084/jem.20061308; pmid: 16982811
63. P. Ye *et al.*, Requirement of interleukin 17 receptor signaling for lung CXC chemokine and granulocyte colony-stimulating factor expression, neutrophil recruitment, and host defense. *J. Exp. Med.* **194**, 519–528 (2001). doi: 10.1084/jem.194.4.519; pmid: 11514607
64. L. L. Lanier, NK cell recognition. *Annu. Rev. Immunol.* **23**, 225–274 (2005). doi: 10.1146/annurev.immunol.23.021704.115526; pmid: 15771571
65. J. C. Sun, J. N. Beilke, L. L. Lanier, Adaptive immune features of natural killer cells. *Nature* **457**, 557–561 (2009). doi: 10.1038/nature07665; pmid: 19136945
66. N. Satoh-Takayama *et al.*, The natural cytotoxicity receptor NKp46 is dispensable for IL-22-mediated innate intestinal immune defense against Citrobacter rodentium. *J. Immunol.* **183**, 6579–6587 (2009). doi: 10.4049/jimmunol.0901935; pmid: 19846871
67. T. Glatzer *et al.*, RORγt$^+$ innate lymphoid cells acquire a proinflammatory program upon engagement of the activating receptor NKp44. *Immunity* **38**, 1223–1235 (2013). doi: 10.1016/j.immuni.2013.05.013; pmid: 23791642
68. N. K. Crellin *et al.*, Regulation of cytokine secretion in human CD127$^+$ LTi-like innate lymphoid cells by Toll-like receptor 2. *Immunity* **33**, 752–764 (2010). doi: 10.1016/j.immuni.2010.10.012; pmid: 21055975
69. D. Bouskra *et al.*, Lymphoid tissue genesis induced by commensals through NOD1 regulates intestinal homeostasis. *Nature* **456**, 507–510 (2008). doi: 10.1038/nature07450; pmid: 18987631
70. Y. Kanamori *et al.*, Identification of novel lymphoid tissues in murine intestinal mucosa where clusters of c-kit$^+$ IL-7R$^+$ Thy1$^+$ lympho-hemopoietic progenitors develop. *J. Exp. Med.* **184**, 1449–1459 (1996). doi: 10.1084/jem.184.4.1449; pmid: 8879216
71. K. G. McDonald *et al.*, CC chemokine receptor 6 expression by B lymphocytes is essential for the development of isolated lymphoid follicles. *Am. J. Pathol.* **170**, 1229–1240 (2007). doi: 10.2353/ajpath.2007.060817; pmid: 17392163
72. K. M. Ansel *et al.*, A chemokine-driven positive feedback loop organizes lymphoid follicles. *Nature* **406**, 309–314 (2000). doi: 10.1038/35018581; pmid: 10917533
73. N. Satoh-Takayama *et al.*, The chemokine receptor CXCR6 controls the functional topography of interleukin-22 producing intestinal innate lymphoid cells. *Immunity* **41**, 776–788 (2014). doi: 10.1016/j.immuni.2014.10.007; pmid: 25456160
74. A. B. Molofsky *et al.*, Innate lymphoid type 2 cells sustain visceral adipose tissue eosinophils and alternatively activated macrophages. *J. Exp. Med.* **210**, 535–549 (2013). doi: 10.1084/jem.20121964; pmid: 23420878
75. J. C. Nussbaum *et al.*, Type 2 innate lymphoid cells control eosinophil homeostasis. *Nature* **502**, 245–248 (2013). doi: 10.1038/nature12526; pmid: 24037376
76. L. Xue *et al.*, Prostaglandin D2 activates group 2 innate lymphoid cells through chemoattractant receptor-homologous molecule expressed on T_H2 cells. *J. Allergy Clin. Immunol.* **133**, 1184–1194.e7 (2014). doi: 10.1016/j.jaci.2013.10.056; pmid: 24388011
77. E. Treiner *et al.*, Selection of evolutionarily conserved mucosal-associated invariant T cells by MR1. *Nature* **422**, 164–169 (2003). doi: 10.1038/nature01433; pmid: 12634786
78. A. Bendelac, M. N. Rivera, S. H. Park, J. H. Roark, Mouse CD1-specific NK1 T cells: Development, specificity, and function. *Annu. Rev. Immunol.* **15**, 535–562 (1997). doi: 10.1146/annurev.immunol.15.1.535; pmid: 9143699
79. K. D. Jensen *et al.*, Thymic selection determines gammadelta T cell effector fate: Antigen-naive cells make interleukin-17 and antigen-experienced cells make interferon gamma. *Immunity* **29**, 90–100 (2008). doi: 10.1016/j.immuni.2008.04.022; pmid: 18585064
80. B. Martin, K. Hirota, D. J. Cua, B. Stockinger, M. Veldhoen, Interleukin-17-producing gammadelta T cells selectively

expand in response to pathogen products and environmental signals. *Immunity* **31**, 321–330 (2009). doi: 10.1016/j.immuni.2009.06.020; pmid: 19682928
81. C. E. Sutton *et al*., Interleukin-1 and IL-23 induce innate IL-17 production from γδ T cells, amplifying Th17 responses and autoimmunity. *Immunity* **31**, 331–341 (2009). doi: 10.1016/j.immuni.2009.08.001; pmid: 19682929
82. G. Gasteiger, A. Y. Rudensky, Interactions between innate and adaptive lymphocytes. *Nat. Rev. Immunol.* **14**, 631–639 (2014). doi: 10.1038/nri3726; pmid: 25132095
83. G. Eberl *et al*., An essential function for the nuclear receptor RORγt in the generation of fetal lymphoid tissue inducer cells. *Nat. Immunol.* **5**, 64–73 (2004). doi: 10.1038/ni1022; pmid: 14691482
84. M. R. Hepworth *et al*., Innate lymphoid cells regulate $CD4^+$ T-cell responses to intestinal commensal bacteria. *Nature* **498**, 113–117 (2013). doi: 10.1038/nature12240; pmid: 23698371
85. N. von Burg *et al*., Activated group 3 innate lymphoid cells promote T-cell-mediated immune responses. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 12835–12840 (2014). doi: 10.1073/pnas.1406908111; pmid: 25136120
86. C. J. Oliphant *et al*., MHCII-mediated dialog between group 2 innate lymphoid cells and $CD4^+$ T cells potentiates type 2 immunity and promotes parasitic helminth expulsion. *Immunity* **41**, 283–295 (2014). doi: 10.1016/j.immuni.2014.06.016; pmid: 25088770
87. A. S. Mirchandani *et al*., Type 2 innate lymphoid cells drive $CD4^+$ Th2 cell responses. *J. Immunol.* **192**, 2442–2448 (2014). doi: 10.4049/jimmunol.1300974; pmid: 24470502
88. L. Jiao *et al*., NK cells promote type 1 T cell immunity through modulating the function of dendritic cells during intracellular bacterial infection. *J. Immunol.* **187**, 401–411 (2011). doi: 10.4049/jimmunol.1002519; pmid: 21642541
89. T. Y. Halim *et al*., Group 2 innate lymphoid cells are critical for the initiation of adaptive T helper 2 cell-mediated allergic lung inflammation. *Immunity* **40**, 425–435 (2014). doi: 10.1016/j.immuni.2014.01.011; pmid: 24613091
90. A. V. Tumanov *et al*., Lymphotoxin controls the IL-22 protection pathway in gut innate lymphoid cells during mucosal pathogen challenge. *Cell Host Microbe* **10**, 44–53 (2011). doi: 10.1016/j.chom.2011.06.002; pmid: 21767811
91. A. A. Kruglov *et al*., Nonredundant function of soluble LTα3 produced by innate lymphoid cells in intestinal homeostasis. *Science* **342**, 1243–1246 (2013). doi: 10.1126/science.1243364; pmid: 24311691
92. L. A. Monticelli *et al*., Innate lymphoid cells promote lung-tissue homeostasis after infection with influenza virus. *Nat. Immunol.* **12**, 1045–1054 (2011). doi: 10.1038/ni.2131; pmid: 21946417
93. J. Li *et al*., Biliary repair and carcinogenesis are mediated by IL-33-dependent cholangiocyte proliferation. *J. Clin. Invest.* **124**, 3241–3251 (2014). doi: 10.1172/JCI73742; pmid: 24892809
94. J. E. Allen, T. E. Sutherland, Host protective roles of type 2 immunity: Parasite killing and tissue repair, flip sides of the same coin. *Semin. Immunol.* **26**, 329–340 (2014). doi: 10.1016/j.smim.2014.06.003; pmid: 25028340
95. E. Scandella *et al*., Restoration of lymphoid organ integrity through the interaction of lymphoid tissue-inducer cells with stroma of the T cell zone. *Nat. Immunol.* **9**, 667–675 (2008). doi: 10.1038/ni.1605; pmid: 18425132
96. A. M. Hanash *et al*., Interleukin-22 protects intestinal stem cells from immune-mediated tissue damage and regulates sensitivity to graft versus host disease. *Immunity* **37**, 339–350 (2012). doi: 10.1016/j.immuni.2012.05.028; pmid: 22921121
97. J. A. Dudakov *et al*., Interleukin-22 drives endogenous thymic regeneration in mice. *Science* **336**, 91–95 (2012). doi: 10.1126/science.1218004; pmid: 22383805
98. L. A. Zenewicz *et al*., Interleukin-22 but not interleukin-17 provides protection to hepatocytes during acute liver inflammation. *Immunity* **27**, 647–659 (2007). doi: 10.1016/j.immuni.2007.07.023; pmid: 17919941
99. L. A. Zenewicz *et al*., Innate and adaptive interleukin-22 protects mice from inflammatory bowel disease. *Immunity* **29**, 947–957 (2008). doi: 10.1016/j.immuni.2008.11.003; pmid: 19100701
100. H. Y. Kim *et al*., Interleukin-17-producing innate lymphoid cells and the NLRP3 inflammasome facilitate obesity-associated airway hyperreactivity. *Nat. Med.* **20**, 54–61 (2014). doi: 10.1038/nm.3423; pmid: 24336249
101. P. D. Cani *et al*., Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet-induced obesity and diabetes in mice. *Diabetes* **57**, 1470–1481 (2008). doi: 10.2337/db07-1403; pmid: 18305141
102. R. Burcelin, L. Garidou, C. Pomié, Immuno-microbiota cross and talk: The new paradigm of metabolic diseases. *Semin. Immunol.* **24**, 67–74 (2012). doi: 10.1016/j.smim.2011.11.011; pmid: 22265028
103. C. S. Klose *et al*., A T-bet gradient controls the fate and function of CCR6-$RORγt^+$ innate lymphoid cells. *Nature* **494**, 261–265 (2013). doi: 10.1038/nature11813; pmid: 23334414
104. G. F. Sonnenberg, L. A. Monticelli, M. M. Elloso, L. A. Fouser, D. Artis, $CD4^+$ lymphoid tissue-inducer cells promote innate immunity in the gut. *Immunity* **34**, 122–134 (2011). doi: 10.1016/j.immuni.2010.12.009; pmid: 21194981
105. T. Y. Halim *et al*., Retinoic-acid-receptor-related orphan nuclear receptor alpha is required for natural helper cell development and allergic inflammation. *Immunity* **37**, 463–474 (2012). doi: 10.1016/j.immuni.2012.06.012; pmid: 22981535
106. M. L. Robinette *et al*., Transcriptional programs define molecular characteristics of innate lymphoid cell classes and subsets. *Nat. Immunol.* **16**, 306–317 (2015). doi: 10.1038/ni.3094; pmid: 25621825
107. B. A. Hamilton *et al*., Disruption of the nuclear hormone receptor RORα in staggerer mice. *Nature* **379**, 736–739 (1996). doi: 10.1038/379736a0; pmid: 8602221
108. E. Hams *et al*., IL-25 and type 2 innate lymphoid cells induce pulmonary fibrosis. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 367–372 (2014). pmid: 24344271
109. T. McHedlidze *et al*., Interleukin-33-dependent innate lymphoid cells mediate hepatic fibrosis. *Immunity* **39**, 357–371 (2013). doi: 10.1016/j.immuni.2013.07.018; pmid: 23954132
110. J. M. Mjösberg *et al*., Human IL-25- and IL-33-responsive type 2 innate lymphoid cells are defined by expression of CRTH2 and CD161. *Nat. Immunol.* **12**, 1055–1062 (2011). doi: 10.1038/ni.2104; pmid: 21909091
111. B. S. Kim *et al*., TSLP elicits IL-33-independent innate lymphoid cell responses to promote skin inflammation. *Sci. Transl. Med.* **5**, 170ra16 (2013). doi: 10.1126/scitranslmed.3005374; pmid: 23363980
112. M. Salimi *et al*., A role for IL-25 and IL-33-driven type-2 innate lymphoid cells in atopic dermatitis. *J. Exp. Med.* **210**, 2939–2950 (2013). doi: 10.1084/jem.20130351; pmid: 24323357
113. T. A. Doherty *et al*., Allergen challenge in allergic rhinitis rapidly induces increased peripheral blood type 2 innate lymphoid cells that express CD84. *J. Allergy Clin. Immunol.* **133**, 1203–1205.e7 (2014). doi: 10.1016/j.jaci.2013.12.1086; pmid: 24582313
114. J. Beale *et al*., Rhinovirus-induced IL-25 in asthma exacerbation drives type 2 immunity and allergic pulmonary inflammation. *Sci. Transl. Med.* **6**, 256ra134 (2014). doi: 10.1126/scitranslmed.3009124; pmid: 25273095
115. J. Y. Hong *et al*., Neonatal rhinovirus induces mucous metaplasia and airways hyperresponsiveness through IL-25 and type 2 innate lymphoid cells. *J. Allergy Clin. Immunol.* **134**, 429–439.e8 (2014). doi: 10.1016/j.jaci.2014.04.020; pmid: 24910174
116. C. M. Lloyd, S. Saglani, Eosinophils in the spotlight: Finding the link between obesity and asthma. *Nat. Med.* **19**, 976–977 (2013). doi: 10.1038/nm.3296; pmid: 23921744
117. J. R. Huh *et al*., Digoxin and its derivatives suppress T_H17 cell differentiation by antagonizing RORγt activity. *Nature* **472**, 486–490 (2011). doi: 10.1038/nature09978; pmid: 21441909
118. L. A. Solt *et al*., Suppression of T_H17 differentiation and autoimmunity by a synthetic ROR ligand. *Nature* **472**, 491–494 (2011). pmid: 21499262
119. C. Barnig *et al*., Lipoxin A4 regulates natural killer cell and type 2 innate lymphoid cell activation in asthma. *Sci. Transl. Med.* **5**, 174ra26 (2013). doi: 10.1126/scitranslmed.3004812; pmid: 23447017
120. J. S. Perry *et al*., Inhibition of LTi cell development by CD25 blockade is associated with decreased intrathecal inflammation in multiple sclerosis. *Sci. Transl. Med.* **4**, 145ra106 (2012). doi: 10.1126/scitranslmed.3004140; pmid: 22855463
121. B. L. Bartlett, S. K. Tyring, Ustekinumab for chronic plaque psoriasis. *Lancet* **371**, 1639–1640 (2008). doi: 10.1016/S0140-6736(08)60702-3; pmid: 18486724
122. S. J. Ballantyne *et al*., Blocking IL-25 prevents airway hyperresponsiveness in allergic asthma. *J. Allergy Clin. Immunol.* **120**, 1324–1331 (2007). doi: 10.1016/j.jaci.2007.07.051; pmid: 17889290
123. A. J. Coyle *et al*., Crucial role of the interleukin 1 receptor family member T1/ST2 in T helper cell type 2-mediated lung mucosal immune responses. *J. Exp. Med.* **190**, 895–902 (1999). doi: 10.1084/jem.190.7.895; pmid: 10510079
124. G. M. Gauvreau *et al*., Effects of an anti-TSLP antibody on allergen-induced asthmatic responses. *N. Engl. J. Med.* **370**, 2102–2110 (2014). doi: 10.1056/NEJMoa1402895; pmid: 24846652
125. H. J. McSorley, N. F. Blair, K. A. Smith, A. N. McKenzie, R. M. Maizels, Blockade of IL-33 release and suppression of type 2 innate lymphoid cell responses by helminth secreted products in airway allergy. *Mucosal Immunol.* **7**, 1068–1078 (2014). doi: 10.1038/mi.2013.123; pmid: 24496315
126. J. Corren *et al*., Lebrikizumab treatment in adults with asthma. *N. Engl. J. Med.* **365**, 1088–1098 (2011). doi: 10.1056/NEJMoa1106469; pmid: 21812663
127. I. D. Pavord *et al*., Mepolizumab for severe eosinophilic asthma (DREAM): A multicentre, double-blind, placebo-controlled trial. *Lancet* **380**, 651–659 (2012). doi: 10.1016/S0140-6736(12)60988-X; pmid: 22901886
128. H. Spits *et al*., Innate lymphoid cells—A proposal for uniform nomenclature. *Nat. Rev. Immunol.* **13**, 145–149 (2013). doi: 10.1038/nri3365; pmid: 23348417
129. C. Vonarbourg *et al*., Regulated expression of nuclear receptor RORγt confers distinct functional fates to NK cell receptor-expressing $RORγt^+$ innate lymphocytes. *Immunity* **33**, 736–751 (2010). doi: 10.1016/j.immuni.2010.10.017; pmid: 21093318
130. G. Eberl, Inducible lymphoid tissues in the adult gut: Recapitulation of a fetal developmental pathway? *Nat. Rev. Immunol.* **5**, 413–420 (2005). doi: 10.1038/nri1600; pmid: 15841100
131. A. Diefenbach, M. Colonna, S. Koyasu, Development, differentiation, and diversity of innate lymphoid cells. *Immunity* **41**, 354–365 (2014). doi: 10.1016/j.immuni.2014.09.005; pmid: 25238093
132. G. Magri *et al*., Innate lymphoid cells integrate stromal and immunological signals to enhance antibody production by splenic marginal zone B cells. *Nat. Immunol.* **15**, 354–364 (2014). doi: 10.1038/ni.2830; pmid: 24562309

ACKNOWLEDGMENTS

G.E. is supported by grants from the Institut Pasteur, the Agence Nationale de la Recherche (ANR), and the Pasteur-Weizmann Council. M.C. is supported by grants from the U.S. National Institutes of Health (1U01AI095542, R01DE021255, and R21CA16719). J.P.D. is supported by grants from the Institut Pasteur, Inserm, La Ligue National Contre le Cancer, and the ANR. A.N.J.M. is supported by the UK Medical Research Council (MRC) and the Wellcome Trust.

10.1126/science.aaa6566

RESEARCH ARTICLES

TUMOR EVOLUTION

High burden and pervasive positive selection of somatic mutations in normal human skin

Iñigo Martincorena,[1] Amit Roshan,[2] Moritz Gerstung,[1] Peter Ellis,[1] Peter Van Loo,[1,3,4] Stuart McLaren,[1] David C. Wedge,[1] Anthony Fullam,[1] Ludmil B. Alexandrov,[1] Jose M. Tubio,[1] Lucy Stebbings,[1] Andrew Menzies,[1] Sara Widaa,[1] Michael R. Stratton,[1] Philip H. Jones,[2]* Peter J. Campbell[1,5]*

How somatic mutations accumulate in normal cells is central to understanding cancer development but is poorly understood. We performed ultradeep sequencing of 74 cancer genes in small (0.8 to 4.7 square millimeters) biopsies of normal skin. Across 234 biopsies of sun-exposed eyelid epidermis from four individuals, the burden of somatic mutations averaged two to six mutations per megabase per cell, similar to that seen in many cancers, and exhibited characteristic signatures of exposure to ultraviolet light. Remarkably, multiple cancer genes are under strong positive selection even in physiologically normal skin, including most of the key drivers of cutaneous squamous cell carcinomas. Positively selected mutations were found in 18 to 32% of normal skin cells at a density of ~140 driver mutations per square centimeter. We observed variability in the driver landscape among individuals and variability in the sizes of clonal expansions across genes. Thus, aged sun-exposed skin is a patchwork of thousands of evolving clones with over a quarter of cells carrying cancer-causing mutations while maintaining the physiological functions of epidermis.

The standard narrative of tumor evolution depicts the accumulation of driver mutations in cancer genes, causing waves of expansion of progressively more disordered clones (*1*, *2*). Central to this model is the presumption that randomly distributed somatic mutations must accumulate in normal cells before transformation (*3*), but directly observing them has proved challenging due to the polyclonal composition of normal tissue. Retrospective reconstructions of clonal evolution from sequencing of tumors give only partial insights, leaving us with fundamental gaps in our understanding of the earliest stages of cancer development. Critical but largely unanswered questions include the burden of somatic mutations in normal cells, which mutational processes are operative in normal tissues, the extent of positive selection among competing clones within a organ, and the patterns of clonal expansion induced by the very first driver mutations (*4*, *5*). These questions have been partially addressed in blood cells, where somatic mutations, including some driver mutations, have been found to accumulate at a low rate with increasing age (*6*–*10*).

[1]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, Cambridgeshire, UK. [2]MRC Cancer Unit, Hutchison-MRC Research Centre, University of Cambridge, Cambridge, UK. [3]Francis Crick Institute, London, UK. [4]Department of Human Genetics, University of Leuven, Leuven, Belgium. [5]Department of Haematology, University of Cambridge, Cambridge, UK.
*Corresponding author. E-mail: phj20@mrc-cu.cam.ac.uk (P.H.J.); pc8@sanger.ac.uk (P.J.C.)

To study the burden, mutational processes and clonal architecture of somatic mutations in normal nonhematological tissue, we focused on sun-exposed skin. Previous studies have reported the existence of clonal patches of skin cells carrying *TP53* mutations (*11*–*15*). Motivated by this, we designed a sequencing strategy capable of detecting such clones by performing ultradeep sequencing of small biopsies and adapting algorithms to detect mutations in a small fraction of cells. We used eyelid epidermis because of its relatively high levels of sun exposure and because it is one of the few body sites to have normal skin excised (blepharoplasty). This procedure is performed for age-related loss of elasticity of the underlying dermis, which can cause eyelid drooping sometimes severe enough to occlude vision, although the epidermis remains physiologically and histologically normal. From four individuals undergoing bilateral blepharoplasty, we obtained the resected eyelids, all of which had normal epidermis free of macroscopic lesions. The donors, three female and one male, ranged from 55 to 73 years of age and had variable histories of sun exposure (table S1). Three were of Western European origin and one was of South Asian origin. We separated the underlying dermis and took multiple biopsies of the epidermis from each eyelid (Fig. 1, A and B). In total, 234 biopsies of 0.79 to 4.71 mm^2 in area were analyzed. We sequenced the coding exons of 74 genes implicated in skin and other cancers to an average effective coverage of 500× (supplementary methods S1.2 and fig. S7). We also performed whole-genome sequencing to ~147× depth on one biopsy in which a predominant clone was found by the targeted gene screen.

Mutational signature of ultraviolet light exposure in normal skin

To identify somatic mutations in the skin biopsies, we adapted an algorithm designed to detect subclonal variants in cancers (*16*) (supplementary methods S1.3 and figs. S3 and S6), based on building a per-base model of background sequencing errors and identifying loci that have a statistical excess of mismatched base calls. This allowed us to detect mutations present in as few as 1% of the cells of a biopsy, detecting mutant clones ranging from 0.01 mm^2 to several square millimeters in size. Overall, we identified 3760 somatic mutations across the 234 biopsies (Fig. 1C and data set S1). Several lines of evidence confirm that the overwhelming majority of these variant calls are genuine somatic mutations (supplementary methods S1.3.2 and figs. S1 and S2).

The pattern of mutations we identified closely matched that expected for ultraviolet (UV) light exposure and that seen in skin cancers (Fig. 1, D to F, and fig. S8). There was a predominance of C>T mutations, especially when the mutated cytosine was preceded by another pyrimidine (namely, a TpC or CpC context), and there were high rates of CC>TT dinucleotide substitutions. This signature is consistent with the known chemistry of sunlight-induced damage to DNA, in which UV rays catalyze the formation of cyclobutane dimers from adjacent pyrimidines (*17*–*20*). C>T and CC>TT mutations were significantly more frequent on the nontranscribed strand of genes (Fig. 1, D and E), which is consistent with transcription-coupled repair (*21*).

We also observed enrichment of C>A (G>T) mutations, with no obvious sequence context but a strong bias toward higher rates of C>A mutations on the transcribed strand (Fig. 1D). Assuming that the strand bias results from transcription-coupled repair, this indicates that the damaged base is the guanine in the C:G pairing. This signature is also seen in cutaneous squamous cell carcinoma (cSCC) cancers, particularly in those with a relatively low mutation burden (fig. S8), but less frequently in basal cell carcinomas (BCCs) and melanomas. A significant fraction of mutations seen after in vitro exposure of cells to UV rays are not the canonical transitions at dipyrimidine sites, with C>A transversions being prominent (*20*). One hypothesis for the pathogenesis of this signature is the oxidation of guanine residues (typically 8-oxoguanine) by reactive oxygen species generated by sunlight (*22*). 8-oxoguanine is subject to transcription-coupled repair (*23*), consistent with the strand bias we see.

Pervasive positive selection of somatic mutations in normal skin

In the Darwinian model of cancer evolution, clones with driver mutations in cancer genes have a selective advantage over those without. In genomic data across multiple tumors, this manifests as an enrichment of protein-altering

mutations in cancer genes as compared to that expected for the background mutation rate. To explore whether clonal selection is operative in normal skin cells, we adapted a dN/dS model that accounts for the context-dependent mutation spectrum and that estimates the background mutation rate of each gene separately using synonymous mutations (*24*) (Fig. 2, supplementary methods S1.4, and fig. S6). One major advantage of this approach is that the mutation rate is estimated locally, thus inherently correcting for the variation in mutation rate across the genome, differences in read depth across the genes surveyed, and the mutational spectrum observed in each individual. Genes under positive selection can be identified, and the number of driver mutations can be quantified from the excess of nonsynonymous mutations (*24*).

Remarkably, six genes had a significant excess of protein-altering base substitutions after correcting for multiple-hypotheses testing (Fig. 2), with five of these also showing excess rates of indels and/or dinucleotide subs (Fig. 2D and supplementary methods S1.4). *NOTCH1* was the most frequently mutated gene in the cohort and showed the highest observed-to-expected ratios of missense, nonsense, and essential splice site mutations. *NOTCH2* and *NOTCH3* also carried a significant excess of protein-altering mutations. NOTCH receptors are key regulators of stem cell biology in a number of organs (*25*) and are a frequent target of inactivating mutations in epithelial cancers (*26*–*29*) and activating mutations in lymphoid malignancies (*30*, *31*). The distribution of somatic mutations within the *NOTCH1* and *NOTCH2* genes was not random, with heavy clustering of amino acid substitutions in the extracellular epidermal growth factor–like domains and large numbers of protein-truncating mutations distributed throughout the genes, matching that observed in cutaneous and head and neck SCCs (Fig. 2I). The density of positively selected driver mutations was surprisingly high. From the excess of protein-altering mutations, we estimated the density of cell clones carrying driver mutations to be 57.1 clones/cm^2 [confidence interval ($CI_{95\%}$): 51 to 61/cm^2] for NOTCH1, 24.6 clones/cm^2 ($CI_{95\%}$: 19 to 28/cm^2) for NOTCH2, and 1.3 clones/cm^2 ($CI_{95\%}$: 0.6 to 1.6/cm^2) for NOTCH3. (Fig. 2C and supplementary methods S1.4.2). Thus, on average we found 83 clones carrying positively selected driver mutations in *NOTCH* genes for every square centimeter of aged, sun-exposed skin.

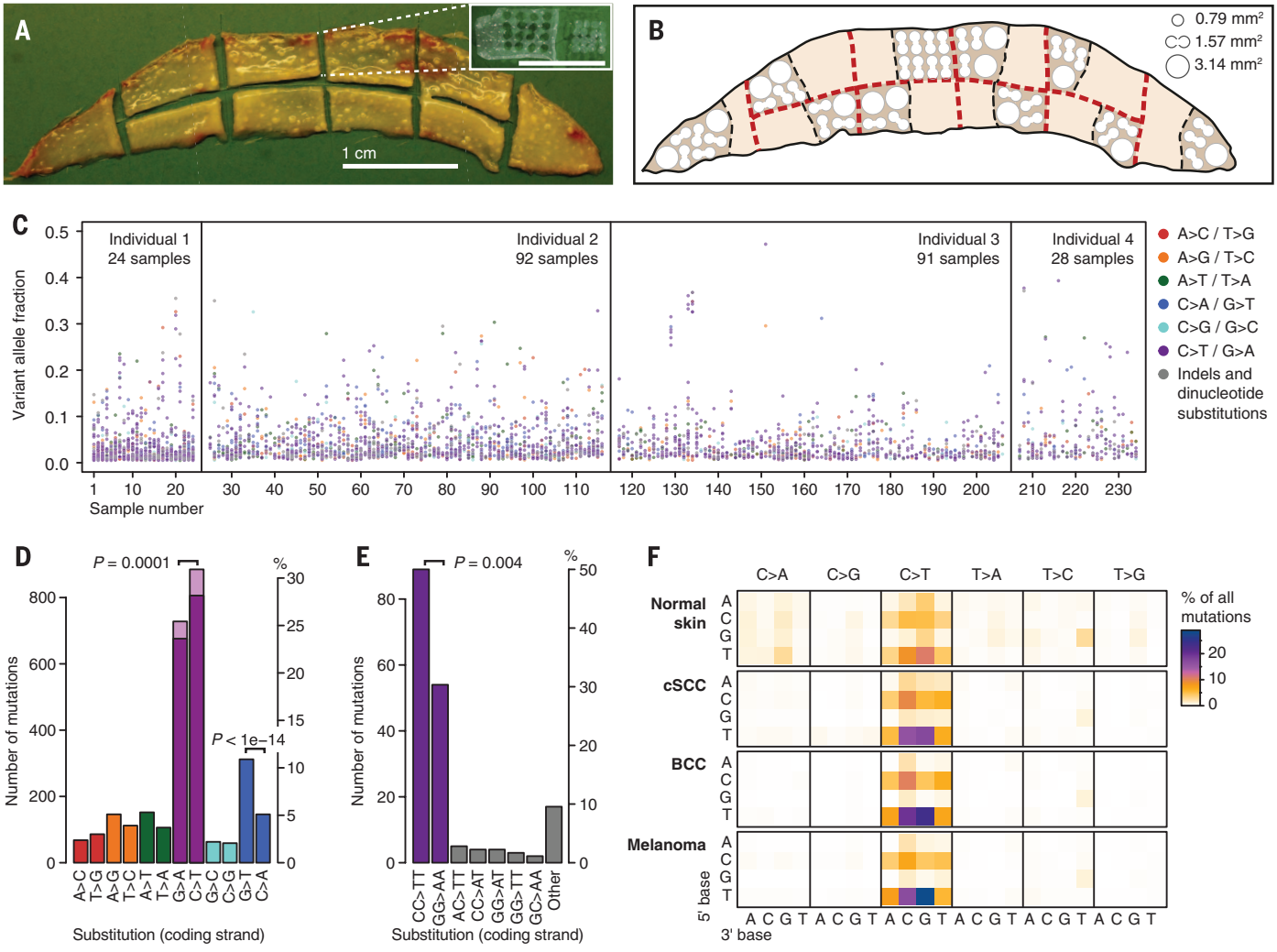


Fig. 1. Burden and spectrum of mutations in normal human skin. (**A**) Excised human eyelid viewed from the dermal surface. The inset shows a sample region of epidermis after the dermis has been removed and biopsies taken. (**B**) Locational map of harvested areas from an eyelid showing locations of 0.79 mm^2, 1.57 mm^2, and 3.14 mm^2 biopsies. (**C**) Distribution of the variant allele fraction (i.e. the fraction of sequencing reads reporting the mutation of all reads across the locus) for the 3760 mutations found across the 234 samples from four individuals, colored by mutation type. (**D** and **E**) Total counts in the coding (untranscribed) versus the noncoding (transcribed) strand for single base substitutions (D) and dinucleotides (E). The counts of C>T (G>A) mutations in a dipyrimidine context are shown in dark purple. *P* values reflect the transcription strand asymmetry (exact Poisson test). (**F**) Heat map of the relative rates of each mutation type, depending on the nucleotides upstream and downstream of the mutated base. Rates are normalized for sequence composition of the targeted genes.

In SCCs of the skin and other organs, both copies of *NOTCH1* are frequently inactivated (*28*, *29*), typically through a point mutation combined with a copy number alteration. We developed an algorithm to identify small populations of cells with copy number alterations across the genes targeted for sequencing by phasing heterozygous single-nucleotide polymorphisms (SNPs) (*32*) (supplementary methods S1.6 and fig. S4). *NOTCH1* was the gene most frequently subject to copy number changes (Fig. 3), with 27 out of 234 biopsies having detectable alterations (Fig. 3B).

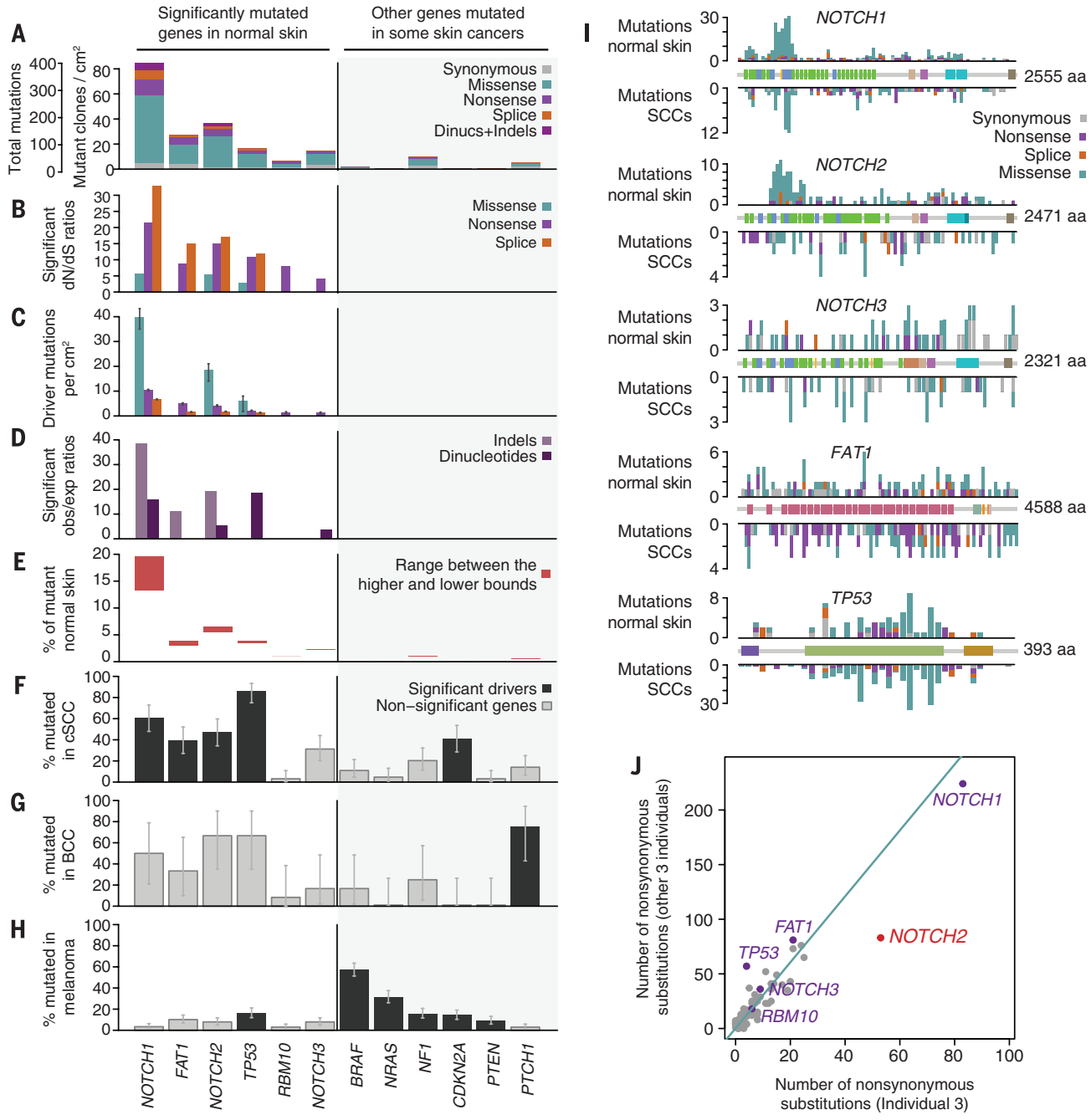


Fig. 2. Pervasive positive selection of oncogenic mutations in normal skin. (A) to (E) Patterns of selection in six genes recurrently mutated in normal skin and in six other genes frequently implicated in skin cancers. (**A**) Number of mutations per gene classified by their functional impact. (**B**) dN/dS ratios for genes under significant positive selection (only statistically significant ratios are shown). (**C**) Estimated number of driver mutations per square centimeter of normal skin. (**D**) Enrichment of indels and dinucleotides in driver genes (bars show significant observed-to-expected ratios only). (**E**) Estimated percentage of cells in normal skin carrying mutations in each gene. Lower-bound estimates were obtained assuming the possibility of up to two driver mutations per cell, whereas higher-bound estimates are obtained by allowing only one driver mutation per gene per cell. (**F** to **H**) Percentage of cSCC, BCC, and melanoma tumors that carry a nonsynonymous point mutation in each gene. Genes found to be significantly recurrently mutated in each cancer type are shown in black (supplementary results S2.2). (**I**) Distribution of mutations across five driver genes in normal skin (above the gene diagrams) and in SCCs (below), including 67 cutaneous SCCs and 319 TCGA head and neck cancer exomes. The gene diagrams show the location of encoded protein domains. (**J**) Differential selection in *NOTCH2* across individuals (supplementary methods S1.5).

Only occasional copy number alterations were detected in other genes, although our power to detect these was variable because of differences in the numbers of heterozygous SNPs sequenced. When we estimate the percentage of cells carrying a *NOTCH1* copy number change in a biopsy, we find that there is often a *NOTCH1* point mutation apparently occurring in the same fraction of cells in the biopsy (Fig. 3C). This overlap, which occurs much more frequently than expected by chance ($P < 10^{-5}$, supplementary methods S1.6.1), demonstrates that biallelic inactivation of *NOTCH1* is already frequent in normal skin cells and not restricted to SCCs.

FAT1 showed a statistically significant excess of inactivating mutations across all classes, including nonsense and essential splice site substitutions and short indels [false discovery rate (FDR)–adjusted *P* value (q) = 8×10^{-11}, 9×10^{-6}, and 2×10^{-4}, respectively; Fig. 2, B to D, and supplementary methods S1.4.5]. *FAT1* is a cadherin-like protein that suppresses tumor growth by blocking β-catenin signaling and is recurrently mutated in a range of cancers (*33*), including cutaneous (table S2) and head and neck SCCs (*34*, *35*). Consistent with previous analyses of mutant clones in normal skin (*11*), we found an estimated 9.5 clones/cm^2 carrying a driver mutation in *TP53* (4.6 to 11.8/cm^2; $q = 4 \times 10^{-6}$). In addition, we saw canonical hotspot mutations in several oncogenes, including *KRAS*, *NRAS*, and *HRAS*.

We found evidence of positive selection in other genes that have not previously been implicated in skin cancer. *RBM10*, which encodes an RNA-binding protein, is subject to recurrent inactivating mutations in lung adenocarcinoma (*36*), and we also see an excess of protein-truncating mutations in normal skin ($q = 0.009$; Fig. 2B). *RBM10* is not a known skin cancer gene, although it may conceivably emerge as a rare driver in cSCCs with further sequencing. Additionally, in an analysis for excess mutations at hotspots, *FGFR3* showed significant recurrence at two canonical residues (supplementary methods S1.4.3). The same hotspot mutations have been found in ~40% of seborrheic keratoses (*37*). These skin growths have an incidence that is 15 times higher than that of skin cancers (*38*), but they never become invasive or malignant. This observation suggests that there may be a class of genes in which somatic mutations give a clonal selective advantage in normal tissue, but do not cause, or could even inhibit, hallmarks of the cancer phenotype such as invasion or dissemination.

We compared the catalog of significantly mutated genes in normal skin to published exome sequencing studies from the three commonest classes of skin cancer, namely cSCCs (*28*, *39*, *40*), BCC (*41*) and melanoma (*42*). When analyzed using the same statistical methodology, there was little overlap in positively selected genes in normal skin compared to either BCC or melanoma (Fig. 2, G and H; supplementary results S2.2; and tables S2 to S4). In contrast, we found that the pattern in normal skin closely matched that of cSCC, with *NOTCH1*, *NOTCH2*, *FAT1*, and *TP53* all being significantly mutated in the latter (Fig. 2F). Point mutations in *CDKN2A* were not found to be under positive selection in normal skin, despite this gene being a frequent driver in cSCC, inactivated by point mutations or homozygous deletions. Although our design does not allow us to reliably detect homozygous deletions, we found only three *CDKN2A* point mutations (two missense and one synonymous) across all 234 samples of normal skin, whereas ~31% ($CI_{95\%}$: 14 to 52%) of cSCCs carry nonsynonymous point mutations in the gene. These data suggest that the selective forces acting on physiologically normal skin resemble those in squamous cell carcinomas, with remarkable similarities between the driver mutations in each. However, *CDKN2A* inactivation

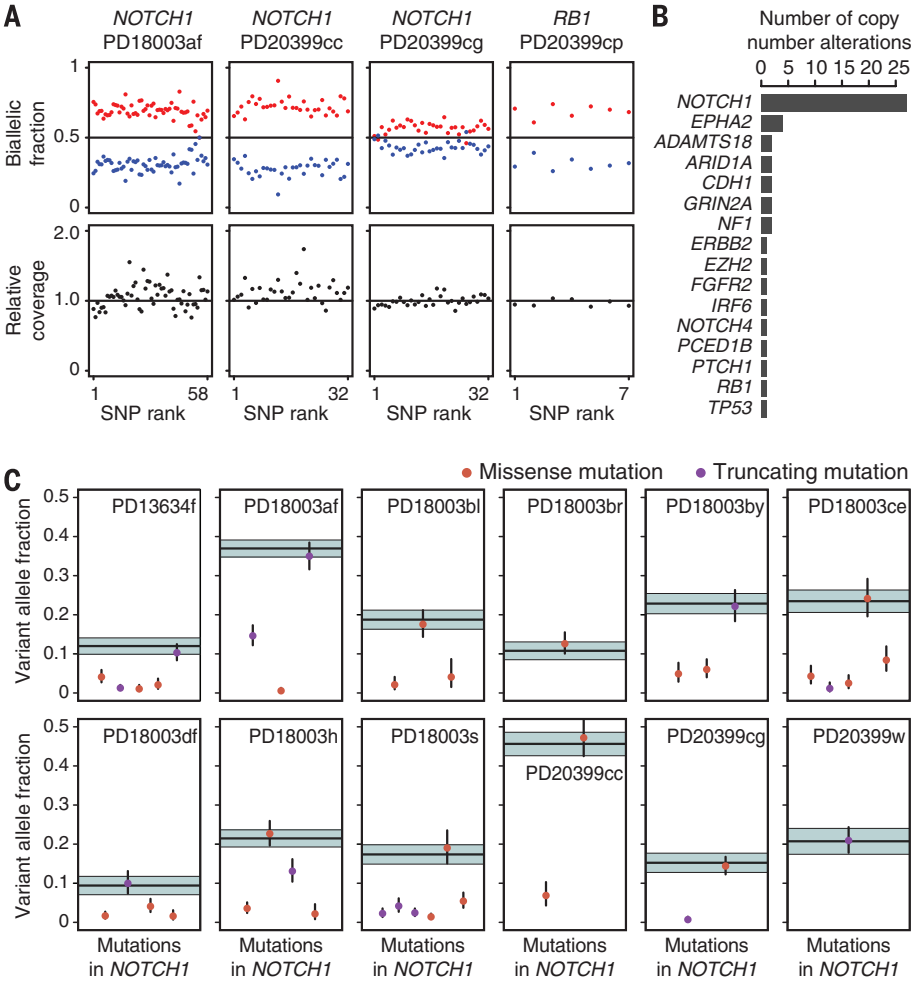


Fig. 3. Frequent copy number aberrations and biallelic loss of *NOTCH1* in normal skin. (**A**) Example of four skin samples with subclonal copy number aberrations in *NOTCH1* and *RB1*. Every point represents a heterozygous SNP within the affected gene, and aberrations manifest as allelic imbalances, with a higher fraction of reads (biallelic fraction) supporting one of the alleles of the gene (in red). The extent of the deviation from 0.5 depends on the number of gene copies gained or lost and on the proportion of the biopsy occupied by the subclone (supplementary methods S1.6). (**B**) Number of copy number aberrations detected per gene. (**C**) In *NOTCH1*, a substitution is often found in the same fraction of cells as a deletion of the other allele (dot colocalizing with a horizontal band), showing that the loss of both copies of *NOTCH1* is frequent in normal skin cells. Horizontal lines represent the expected variant allele fraction for a mutation inactivating the only remaining allele of a gene in the same clone, with colored shadows representing 95% confidence intervals. Orange and purple dots represent the allele fraction of missense and nonsense mutations in the biopsy, with 95% CIs (supplementary methods S1.6.1 and fig. S5).

appears to be specific to cancer clones, suggesting that its loss confers no selective advantage until more advanced stages of cancer evolution. The absence of mutations characteristic of melanomas is consistent with the fact that around 95% of the cells in the epidermis are keratinocytes (*43*), whereas melanomas originate from melanocytes. The absence of the *PTCH1* mutations seen in BCC is notable, especially given that BCC has a three-fold higher incidence than cSCC in populations of European ancestry (*44*). This may be consistent with BCC originating from cells infrequent or absent in the eyelid epidermis, such as from hair follicles (*45*), although our data cannot rule out other explanations.

Surprisingly, one of the four individuals in our series contributed a disproportionate number of mutations in *NOTCH2* (39% of all mutations in *NOTCH2* compared to 24% in other genes). A formal test of heterogeneity confirmed that *NOTCH2* showed a variable rate of driver mutations among individuals (q = 0.0005; Fig. 2J; supplementary methods S1.5; and figs. S9 and S10). Because the dN/dS method used inherently accounts for gene-specific coverage and patient-specific mutation spectrum, this finding is likely to reflect a true biological difference among the four individuals rather than a bias arising from some aspect of the experimental design. One conceivable explanation is that some difference in the local eyelid environment provides a stronger pressure for *NOTCH2* mutations; another, more likely explanation is that the genetic background of each individual could lead to differences in the strength of selective advantage across genes. The patient with different selection strength for *NOTCH2* was of South Asian ancestry, whereas the other three were Western Europeans, although this needs considerably larger sample sizes to address formally. Nonetheless, these data illustrate the exciting potential of such study designs to detect inter-individual differences in the driver landscape that cannot be extracted from sequencing a single established cancer per patient.

Mutant clonal expansions

Together with the mutation rate, the size of the clonal expansions induced by driver mutations in normal tissue is critical to understanding the evolution of cancer, since both factors together determine the size of the pool of cells that can

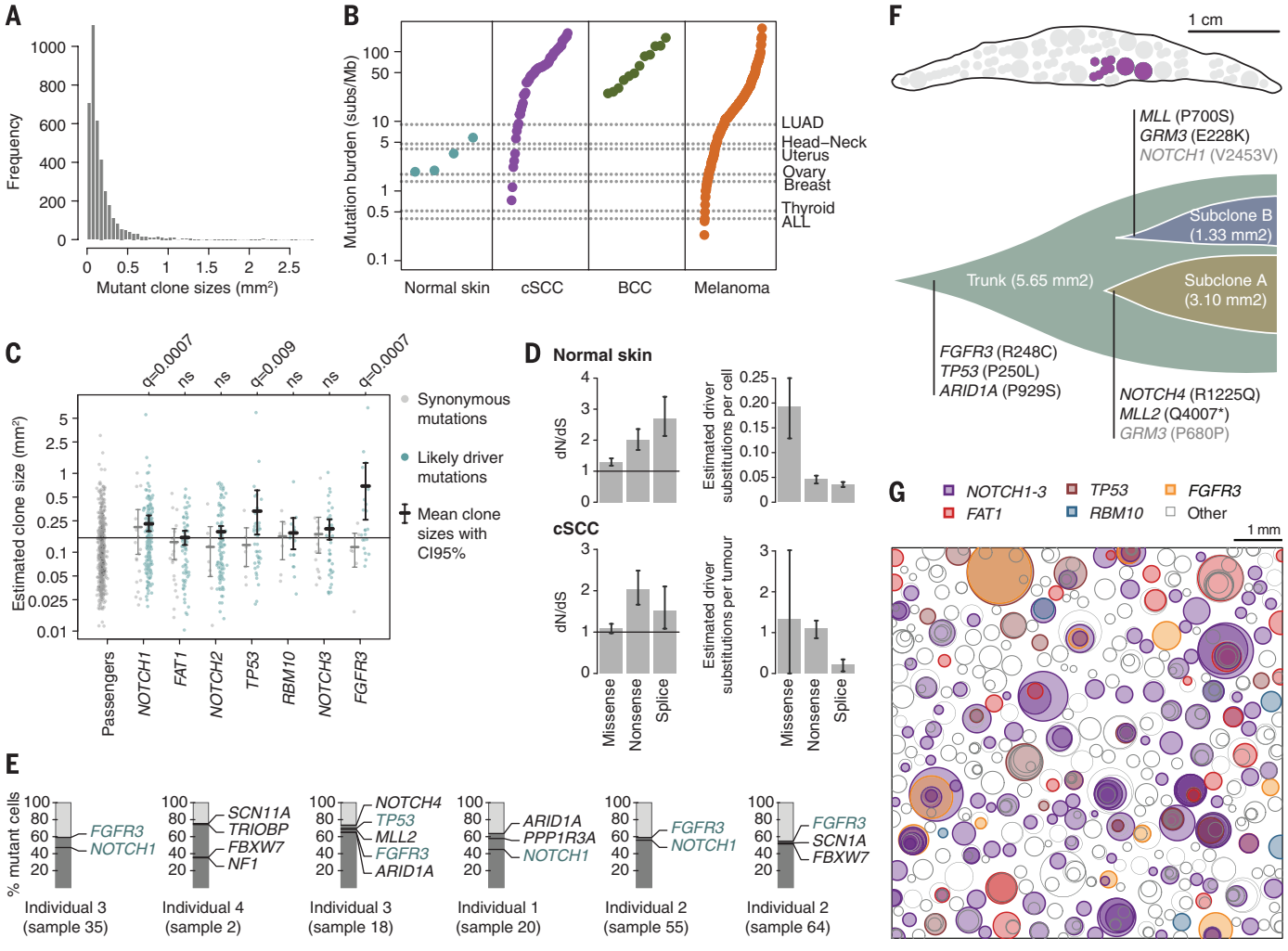


Fig. 4. Mutant clone sizes and clonal dynamics in normal skin. (**A**) Distribution of clone sizes of all mutations. (**B**) Mutation burden per megabase in the normal skin of four individuals and in a range of human cancers (supplementary methods S1.8). (**C**) Clone sizes of likely driver and passenger mutations in normal skin. Driver mutations are defined as those mutation types found to be under significant positive selection in each gene (Fig. 2). CIs and FDR-adjusted P values (q) values were obtained using 10,000 random permutations of the gene labels assigned to each mutation. (**D**) Global dN/dS estimates across all 74 genes analyzed in the study in normal skin and cSCC. This allows us to estimate the number of driver mutations per normal cell or per tumor as the number of mutations fixed by positive selection (supplementary methods S1.4.2). (**E**) Identification of mutations co-occurring in the same subclone, using the pigeonhole principle (*32*). (**F**) Subclonal structure of a large clone found to overlap with six biopsies (shown in purple in the eyelid locational map). (**G**) Schematic representation of the mutant clones in an average 1 cm^2 of normal eyelid skin. To generate the figure, a number of biopsies were randomly selected to amount to 1 cm^2 of sequenced skin, and all clones observed in these biopsies were represented as circles randomly distributed in space. The density, size, and simulated nesting of clones are all based on the sequencing data obtained in this study.

acquire sequential hits (*4*, *5*). With our experimental design, the observed fraction of sequencing reads reporting a mutation correlates accurately with the fraction of cells in a biopsy that carry the mutation, once we correct for the local copy number at that locus, enabling us to estimate clone sizes (*32*) (supplementary sections S1.7 and S2.6.1). For the majority of mutations identified here, the variant allele fraction was <5% (Fig. 1B), indicating that most mutations were seen in only a small proportion of cells in the biopsy, typically <10%, with many mutations seen in only 1 to 2% of cells. There were exceptions, however, and some biopsies carried somatic mutations found in most of the cells. We find that the distribution of mutant clone sizes in aged, sun-exposed skin has a heavy right tail (Fig. 4A and fig. S11A), with some clones as large as several square millimeters in surface area.

To estimate the average burden of somatic substitutions per skin cell, we can integrate the estimated fraction of mutant cells across the biopsies from each of the four donors (supplementary methods S1.8). This reveals that the mutation burden estimated from coding sequence is at least two to six somatic mutations/Mb/cell (Fig. 4B). This estimate is at the lower end of the burden of mutations in cSCCs (1 to 380/Mb) and melanomas (0.5 to 200/Mb), and higher than the average mutation burdens seen in many adult solid tumors (Fig. 4B). Using the variant allele fraction, we estimate that 14 to 21% of skin cells carry *NOTCH1* mutations, with 5 to 7% having *NOTCH2* and 2 to 3% *NOTCH3* mutations (Fig. 2E). *TP53* mutations and *FAT1* mutations are present in 3 to 5% of skin cells, remarkably similar to the estimate of 4% from immunohistochemical studies of *TP53* clones in human skin (*11*). Thus, about a quarter of all skin cells in these biopsies carried NOTCH mutations, the vast majority of which are driver mutations.

In current models of cancer development, driver mutations cause clonal expansions, widening the pool of cells that is susceptible to further driver mutations until enough accumulate to drive transformation and invasion. We compared the clone size of mutations in driver genes against that of synonymous mutations in non-driver genes, which are likely selectively neutral (Fig. 4C). We find that whereas the average clone size for neutral mutations was 0.15 mm^2 ($CI_{95\%}$: 0.13 to 0.17), it was significantly larger for driver mutations in *NOTCH1* (average 0.23 mm^2; q = 0.002), *TP53* (0.33 mm^2; q = 0.009), and *FGFR3* (0.69 mm^2; q = 0.0007; permutation test). Clone sizes for *FAT1*, *NOTCH2*, and *NOTCH3* mutations were not significantly increased. Although some putatively neutral mutations in this data set may be hitchhiking in clones with driver mutations, the difference in clone sizes between driver and neutral mutations is unexpectedly small. The large excess of truncating mutations in these genes demonstrates that clones carrying these mutations must have had a strong selective advantage at some stage. Indeed, lineage tracing in mice has revealed that clones carrying *TP53* mutations grow nearly exponentially in UV-exposed epidermis (*13*). Yet, exponential growth must slow relatively early in the expansion of the clones to explain both the limited range of clone sizes observed here (Fig. 4C) and their similarity across individuals of different ages (Fig. 1C). Such constraints on clonal growth are likely to represent a critical protection against progressive accumulation of driver mutations and cancer. The physiological mechanisms underpinning this are unknown, but "imprisonment" of *TP53*-mutant clones has been observed in murine epidermis (*46*), possibly driven by interactions between the clone and surrounding cells and density-dependent growth constraints.

In contrast to the relatively small clone sizes of canonical cSCC driver genes, clones with activating *FGFR3* mutations were among the largest observed. It is striking that the driver mutations inducing the largest clonal expansions in normal skin were those associated with benign tumors, namely seborrheic keratosis. This shows that the size of clonal expansion induced by a somatic mutation need not correlate with its potential to induce malignant transformation.

Our data reveal notable similarities between normal and cancer cells, with normal cells carrying thousands of mutations, including oncogenic driver mutations subject to strong positive selection. A major difference between the normal cells sequenced here and cancer cells seems to be the number of driver mutations per cell (Fig. 4D). Using dN/dS, we estimate that normal cells in the skin of these four subjects carry an average of 0.27 ($CI_{95\%}$: 0.19 to 0.35) driver point mutations per cell. Using the same method for cSCCs, we estimate an average of 2.7 ($CI_{95\%}$: 0.91 to 4.65) driver point mutations per tumor in the genes sequenced in this study.

At an average of 0.27 driver mutations per cell, there may be many normal cells with several drivers coexisting. When clones represent a large enough fraction of the biopsy, we can apply deductive reasoning to demonstrate co-occurrence of mutations in the same clone of cells (*32*). In our data, there were six large clones for which this was possible (Fig. 4E), with three showing two or more likely driver mutations in the same subclone. In one massive clone that spanned six adjacent biopsies, we found all cells carrying a canonical activating mutation in *FGFR3* together with a known driver mutation in *TP53*, and two separate subclonal expansions (Fig. 4F).

To obtain a more comprehensive picture of the mutational landscape of normal cells, we performed whole genome sequencing to 147x depth on a biopsy containing this clone. This identified 73,904 base substitutions and 2248 small indels, with a mutation signature largely dominated by UV light exposure (fig. S12, B and C). About 14,000 of these were clonal (~4.6/Mb), presumably hitchhiking with the *FGFR3* and *TP53* mutations, but the rest were subclonal, often in <20% cells (fig. S12A). Integrating the allele frequencies, we estimate an average of 21,102 mutations per genome per cell (~7/Mb) in this sample. The mutation rate was found to vary along the genome, with higher rates in lowly expressed genes and in repressed chromatin (fig. S13), as observed in cancer (*47*) and human evolution (*48*).

Discussion

We found the frequency of driver mutations in physiologically normal skin cells surprisingly high. For example, there were more *NOTCH1* mutations in just 5 cm^2 of aged, sun-exposed skin analyzed here than have been identified in more than 5000 cancers sequenced by TCGA (The Cancer Genome Atlas). About 20% of normal skin cells carry driver mutations in *NOTCH1*, with some but not overwhelming enrichment in the matching cancer (60% of cSCCs have *NOTCH1* mutations). Several other cancer genes were under positive selection in normal skin, and we found clones carrying two to three driver mutations that had not acquired malignant potential, raising the question of what combinations of events are sufficient for transformation. These observations may not be entirely unexpected: For cancers to occur with the frequency they do in the general population, there may be a vast underlying reservoir of competing clones part or much of the way to malignant transformation. A rather sobering corollary is that if we had a systemic targeted therapeutic that killed all cells with inactivated *NOTCH1*, we might successfully treat 60% of cSCCs but with considerable collateral damage to physiologically normal skin.

Studying tumor evolution by sequencing established cancers is akin to inferring the rules of a musical talent quest by identifying similarities across the show's annual winners. Successful aspirants undoubtedly have common properties that identify necessary criteria for victory, but there is no substitute for directly observing the competition in its raw, early, local heats. Here, we have found hundreds of evolving clones per square centimeter of skin (Fig. 4G); thousands of mutations per skin cell; variability among individuals in the profile of driver mutations; and variability among cancer genes in clonal dynamics. Scaled up across the range of organ systems, cell types, and mutational exposures—and encompassing aging, predisposing diseases, and genetic backgrounds—such studies promise to reveal fundamental insights into the earliest stages of cancer development.

REFERENCES AND NOTES

1. B. Vogelstein *et al.*, *Science* **339**, 1546–1558 (2013).
2. L. A. Garraway, E. S. Lander, *Cell* **153**, 17–37 (2013).
3. C. Tomasetti, B. Vogelstein, G. Parmigiani, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 1999–2004 (2013).
4. P. C. Nowell, *Science* **194**, 23–28 (1976).
5. A. L. Jackson, L. A. Loeb, *Genetics* **148**, 1483–1490 (1998).
6. J. S. Welch *et al.*, *Cell* **150**, 264–278 (2012).
7. S. Jaiswal *et al.*, *N. Engl. J. Med.* **371**, 2488–2498 (2014).
8. G. Genovese *et al.*, *N. Engl. J. Med.* **371**, 2477–2487 (2014).
9. M. Xie *et al.*, *Nat. Med.* **20**, 1472–1478 (2014).
10. T. McKerrell *et al.*, *Cell Reports* **10**, 1239–1245 (2015).
11. A. S. Jonason *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 14025–14029 (1996).
12. A. Ziegler *et al.*, *Nature* **372**, 773–776 (1994).
13. A. M. Klein, D. E. Brash, P. H. Jones, B. D. Simons, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 270–275 (2010).
14. H. Nakazawa *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 360–364 (1994).
15. G. Ling *et al.*, *Am. J. Pathol.* **159**, 1247–1253 (2001).
16. M. Gerstung, E. Papaemmanuil, P. J. Campbell, *Bioinformatics* **30**, 1198–1204 (2014).

17. L. B. Alexandrov *et al.*, *Nature* **500**, 415–421 (2013).
18. G. P. Pfeifer, Y. H. You, A. Besaratinia, *Mutat. Res.* **571**, 19–31 (2005).
19. G. P. Pfeifer, A. Besaratinia, *Photochem. Photobiol. Sci.* **11**, 90–97 (2012).
20. D. E. Brash, *Photochem. Photobiol.* **91**, 15–26 (2015).
21. S. C. Shuck, E. A. Short, J. J. Turchi, *Cell Res.* **18**, 64–72 (2008).
22. A. Besaratinia, S. I. Kim, S. E. Bates, G. P. Pfeifer, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 5953–5958 (2007).
23. J. Guo, P. C. Hanawalt, G. Spivak, *Nucleic Acids Res.* **41**, 7700–7712 (2013).
24. C. Greenman, R. Wooster, P. A. Futreal, M. R. Stratton, D. F. Easton, *Genetics* **173**, 2187–2198 (2006).
25. J. Liu, C. Sato, M. Cerletti, A. Wagers, *Curr. Top. Dev. Biol.* **92**, 367–409 (2010).
26. N. J. Wang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 17761–17766 (2011).
27. P. S. Hammerman *et al.*, *Nature* **489**, 519–525 (2012).
28. S. Durinck *et al.*, *Cancer Discov.* **1**, 137–143 (2011).
29. N. Agrawal *et al.*, *Science* **333**, 1154–1157 (2011).
30. X. S. Puente *et al.*, *Nature* **475**, 101–105 (2011).
31. A. P. Weng *et al.*, *Science* **306**, 269–271 (2004).
32. S. Nik-Zainal *et al.*, *Cell* **149**, 994–1007 (2012).
33. L. G. Morris *et al.*, *Nat. Genet.* **45**, 253–261 (2013).
34. N. Stransky *et al.*, *Science* **333**, 1157–1160 (2011).
35. India Project Team of the International Cancer Genome Consortium, *Nat. Commun.* **4**, 2873 (2013).
36. M. Imielinski *et al.*, *Cell* **150**, 1107–1120 (2012).
37. A. Logié *et al.*, *Hum. Mol. Genet.* **14**, 1153–1160 (2005).
38. I. Harvey, S. Frankel, R. Marks, D. Shalom, M. Nolan-Farrell, *Br. J. Cancer* **74**, 1302–1307 (1996).
39. A. P. South *et al.*, *J. Invest. Dermatol.* **134**, 2630–2638 (2014).
40. C. R. Pickering *et al.*, *Clin. Cancer Res.* **20**, 6582–6592 (2014).
41. S. S. Jayaraman, D. J. Rayhan, S. Hazany, M. S. Kolodney, *J. Invest. Dermatol.* **134**, 213–220 (2014).
42. E. Hodis *et al.*, *Cell* **150**, 251–263 (2012).
43. J. A. McGrath, R. A. Eady, F. M. Pope, *Rook's Textbook of Dermatology* (Blackwell Science Ltd Oxford, Oxford, ed. 7, 2004).
44. A. Lomas, J. Leonardi-Bee, F. Bath-Hextall, *Br. J. Dermatol.* **166**, 1069–1080 (2012).
45. M. Grachtchouk *et al.*, *J. Clin. Invest.* **121**, 1768–1781 (2011).
46. W. Zhang, E. Remenyik, D. Zelterman, D. E. Brash, N. M. Wikonkal, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 13948–13953 (2001).
47. B. Schuster-Böckler, B. Lehner, *Nature* **488**, 504–507 (2012).
48. I. Martincorena, N. M. Luscombe, *BioEssays* **35**, 123–130 (2013).

ACKNOWLEDGMENTS

We thank M. S. Kolodney, R. J. Cho, and M. Dimon for kindly providing their published exome sequencing data on BCC and cSCC. This work was supported by the Wellcome Trust (grant 077012/Z/05/Z) and by a Medical Research Council (MRC) Centennial award (P.H.J.). P.J.C. has a Wellcome Trust Senior Clinical Research Fellowship (WT088340MA). P.H.J. is supported by an MRC grant in aid and Cancer Research UK (CRUK) (program grant C609/A17257). I.M. is supported by fellowships from EMBO (grant 1287-2012) and Queens' College, Cambridge. A.R. acknowledges support from the CRUK-Cambridge Cancer Centre Clinical Research Fellowship. J.M.C.T. is supported by Marie Curie Fellowship FP7-PEOPLE-2012-IEF (project number 328264), and P.J.C. holds equity in and is a paid consultant for 14M Genomics Ltd. The sequencing data have been deposited in the European Genome-Phenome Archive (accession nos. EGAS00001000860, EGAS00001000515, and EGAS00001000603). The R code used for variant calling (ShearwaterML) is available in Bioconductor (deepSNV package).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/880/suppl/DC1
Methods
Supplementary Text
Figs. S1 to S13
Tables S1 to S4
References (*49*–*68*)
Author Contributions
Data Set S1

13 January 2015; accepted 13 April 2015
10.1126/science.aaa6806

ORGANIC CHEMISTRY

Practical olefin hydroamination with nitroarenes

Jinghan Gui,[1] Chung-Mao Pan,[1]* Ying Jin,[1]* Tian Qin,[1] Julian C. Lo,[1] Bryan J. Lee,[1] Steven H. Spergel,[2] Michael E. Mertzman,[2] William J. Pitts,[2] Thomas E. La Cruz,[3] Michael A. Schmidt,[3] Nitin Darvatkar,[4] Swaminathan R. Natarajan,[4] Phil S. Baran[1]†

The synthesis and functionalization of amines are fundamentally important in a vast range of chemical contexts. We present an amine synthesis that repurposes two simple feedstock building blocks: olefins and nitro(hetero)arenes. Using readily available reactants in an operationally simple procedure, the protocol smoothly yields secondary amines in a formal olefin hydroamination. Because of the presumed radical nature of the process, hindered amines can easily be accessed in a highly chemoselective transformation. A screen of more than 100 substrate combinations showcases tolerance of numerous unprotected functional groups such as alcohols, amines, and even boronic acids. This process is orthogonal to other aryl amine syntheses, such as the Buchwald-Hartwig, Ullmann, and classical amine-carbonyl reductive aminations, as it tolerates aryl halides and carbonyl compounds.

The formation and manipulation of amines represent a large fraction of the daily activity of practicing organic chemists (*1*). The most useful methods for the synthesis and functionalization of amines currently include alkylation (*2*, *3*), amine-carbonyl reductive amination (*4*), and C–N cross-coupling (*5*–*8*). For example, secondary aromatic and heteroaromatic amines are usually accessed by arylation or alkylation of the parent amine. Given the prevalence of amines in medicinal chemistry (*9*) and some of the limitations of current amine syntheses, we pursued a distinct pathway for their construction.

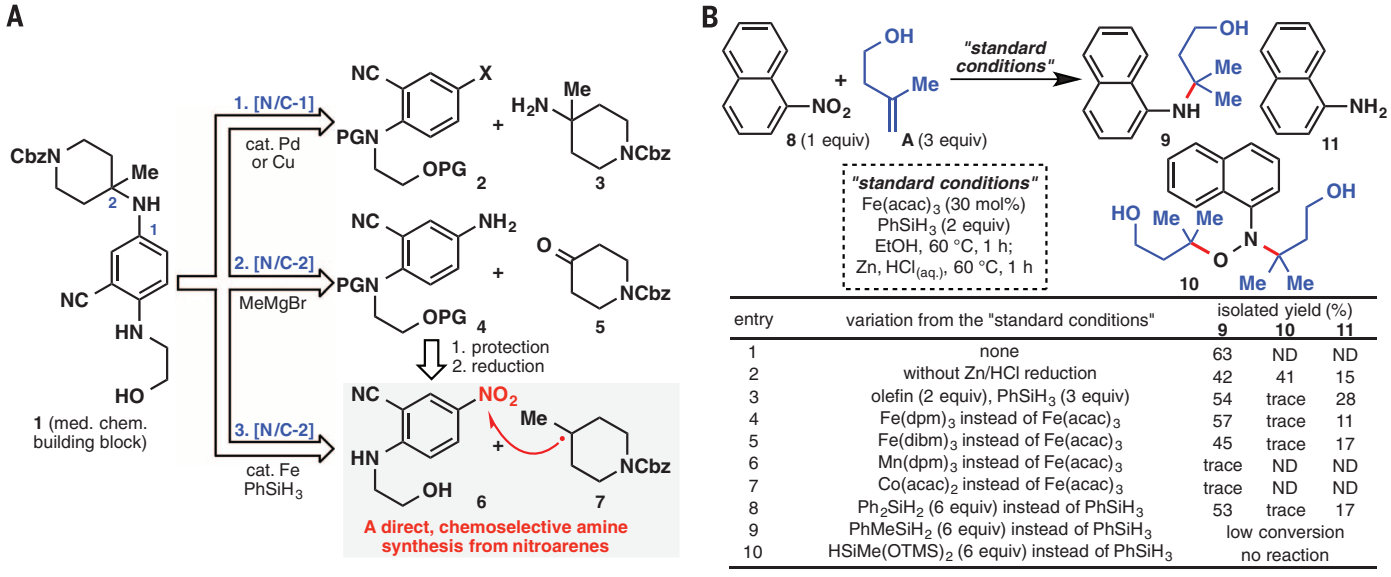
The challenges of amine synthesis can be exemplified with conventional retrosynthetic logic applied to a prototypical medicinal chemistry building block **1** (Fig. 1A). The first disconnection, between nitrogen and C-1, results in a C–N cross-coupling transform and leads to aromatic halide **2** and the hindered primary amine **3** (*10*). Because of functional group incompatibilities, protecting groups on the alcohol and amino groups of **2** might be required. The second approach involves a disconnection between nitrogen and C-2 and proceeds by way of a Grignard addition to an intermediate imine formed between protected amine **4** and ketone **5**. Amine **4** would be derived by protection and reduction of nitroarene **6**, which, unlike **2** and **4**, is commercially available. However, both of these routes contain concession steps arising from the need for protecting groups and external redox manipulations. Thus, a third disconnection was envisaged, in which C–N bond construction using radical **7** is designed to occur concomitantly with the reduction of a nitro group. To our knowledge, there are no practical methods for direct formation of a C–N bond from a nitroarene that liberate a secondary amine. In this work, the invention of such a reaction is reported that uses simple olefins as the radical source, an inexpensive silane and zinc metal as reductants, and an abundant iron salt as a catalyst.

[1]Department of Chemistry, Scripps Research Institute, La Jolla, CA 92037, USA. [2]Discovery Chemistry, Bristol-Myers Squibb, Princeton, NJ 08543, USA. [3]Chemical Development, Bristol-Myers Squibb, New Brunswick, NJ 08903, USA. [4]Kemxtree LLC, 1370 Hamilton Street, Somerset, NJ 08873, USA.
*These authors contributed equally to this work. **†Corresponding author. E-mail: pbaran@scripps.edu**

Reaction development and optimization

There were clues in the literature suggesting the feasibility of this reaction. For example, Russell and Yao demonstrated that *tert*-butyl radicals, derived from the photoinduced decomposition of an organomercury species, could add to both nitroarenes and nitrosoarenes to give mixtures of *N*- and *N*,*O*-alkylated adducts (*11*). Numerous reports have shown that radicals react readily with nitroso compounds (*12*–*15*), as demonstrated by Corey and Gross as a means to generate hindered amines (*16*). Buchwald and colleagues have also demonstrated that hydroxylamine derivatives can be used in olefin hydroamination under copper catalysis (*17*), and Lalic and colleagues have used similar hydroxylamine derivatives to aminate alkyl boranes in a two-step hydroamination process (*18*). Additionally, Yu and colleagues have coupled hydroxylamine derivatives with aromatic C–H bonds to generate anilines under palladium catalysis (*19*). Given the widespread availability of nitro(hetero)arenes and their ease of synthesis, it is surprising that they have not been exploited further beyond their reduction to the corresponding aniline (*20*), the Cadogan carbazole synthesis (*21*, *22*), the Bartoli indole synthesis (*23*), and a few C–C bond-forming reactions (*24*, *25*).

Our recent work on Fe-catalyzed olefin cross-coupling via alkyl radical intermediates (*26*, *27*) led us to attempt the coupling of nitronaphthalene **8** with isoprenyl alcohol **A** (Fig. 1B). To our delight, useful quantities of the desired amine product **9** were isolated upon initial attempts, along



entry	variation from the "standard conditions"	isolated yield (%) 9	10	11
1	none	63	ND	ND
2	without Zn/HCl reduction	42	41	15
3	olefin (2 equiv), $PhSiH_3$ (3 equiv)	54	trace	28
4	$Fe(dpm)_3$ instead of $Fe(acac)_3$	57	trace	11
5	$Fe(dibm)_3$ instead of $Fe(acac)_3$	45	trace	17
6	$Mn(dpm)_3$ instead of $Fe(acac)_3$	trace	ND	ND
7	$Co(acac)_2$ instead of $Fe(acac)_3$	trace	ND	ND
8	Ph_2SiH_2 (6 equiv) instead of $PhSiH_3$	53	trace	17
9	$PhMeSiH_2$ (6 equiv) instead of $PhSiH_3$	low conversion		
10	$HSiMe(OTMS)_2$ (6 equiv) instead of $PhSiH_3$	no reaction		

Fig. 1. Amine synthesis via coupling of nitroarenes and olefins. (**A**) Amine retrosynthesis: a case study from drug discovery. (**B**) Invention and optimization of a nitroarene-based olefin hydroamination. Cbz, benzyloxycarbonyl; PG, protecting group; Me, methyl; Ph, phenyl; Et, ethyl; acac, acetylacetonate; dpm, 2,2,6,6-tetramethylheptane-3,5-dionate; dibm, 2,6-dimethylheptane-3,5-dionate; OTMS, trimethylsilyloxy; ND, not determined.

with substantial amounts of *N*,*O*-alkylated adduct **10** and reduced aminonaphthalene **11**. By modulating the reaction's stoichiometry and introducing a Zn-mediated reduction of the *N*, *O*-alkylated adduct **10** into the same flask, the desired amine **9** was isolated in 63% yield. Fe salts are unique in their ability to catalyze this reaction (Fig. 1B, entries 1 to 5), as Co-based (*28*–*31*) (Fig. 1B, entry 7) and Mn-based (*32*, *33*) (Fig. 1B, entry 6) systems delivered only trace amounts of product. In the absence of olefin, $Co(acac)_2$ and $Mn(dpm)_3$ failed to reduce the nitroarene to the corresponding aniline, a process observed with the use of $Fe(acac)_3$; this indicates that there is some interaction of the Fe-based system with the nitroarene prior to presumed radical addition (see figs. S1 and S2). Among the silanes screened, $PhSiH_3$ proved to be the most effective at facilitating the transformation (Fig. 1B, entries 8 to 10).

Exploring substrate scope

With an optimized set of conditions in hand, the scope of both the olefin and nitroarene partners was extensively evaluated. We subjected 27 different olefin donors (Fig. 2) to hydroamination with an array of nitro(hetero)arenes adorned with a variety of functional groups, for a total of 113 examples (Figs. 3 to 7). In accord with known reactivity trends in Fe-based olefin functionalization (*34*, *35*), adducts were formed at the most substituted carbon of the olefin, with olefins **A** and **B** delivering the same products. Mono-, di-, tri-, and tetrasubstituted olefins serve as viable substrates ranging in complexity. Isobutylene (**E**) can be used to enable facile access to *N*-*tert*-butyl aromatic amines. Several of the olefins, such as **F**, **G**, and **K**, permit access to extremely hindered amines that might be challenging to prepare in other ways (*10*). This is particularly exciting given the documented utility of hindered amines as a method to block metabolism in drug discovery (*36*).

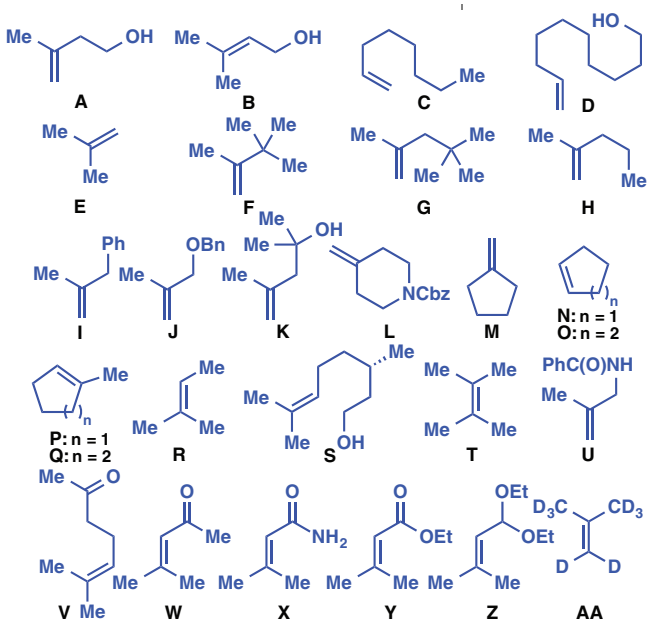
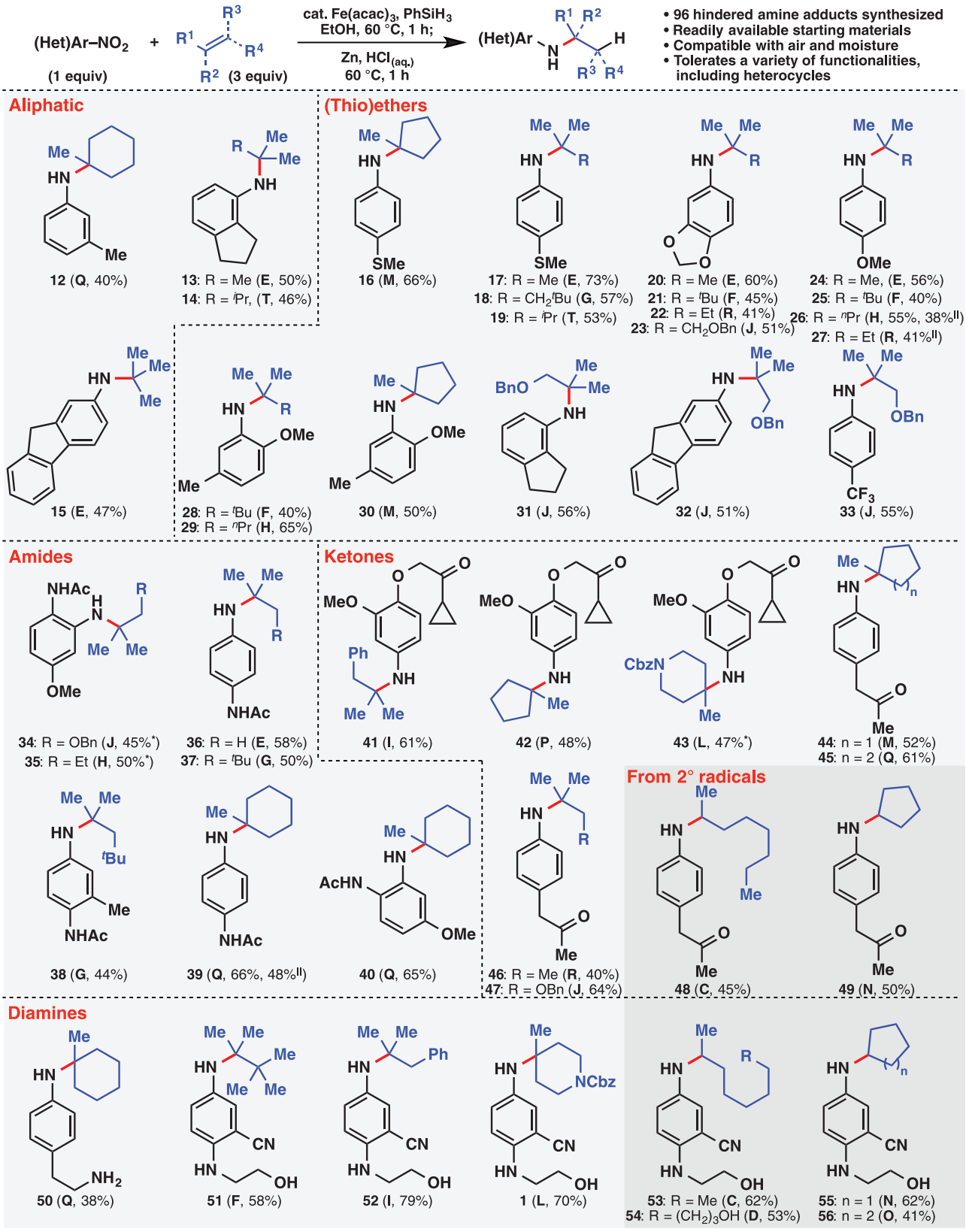


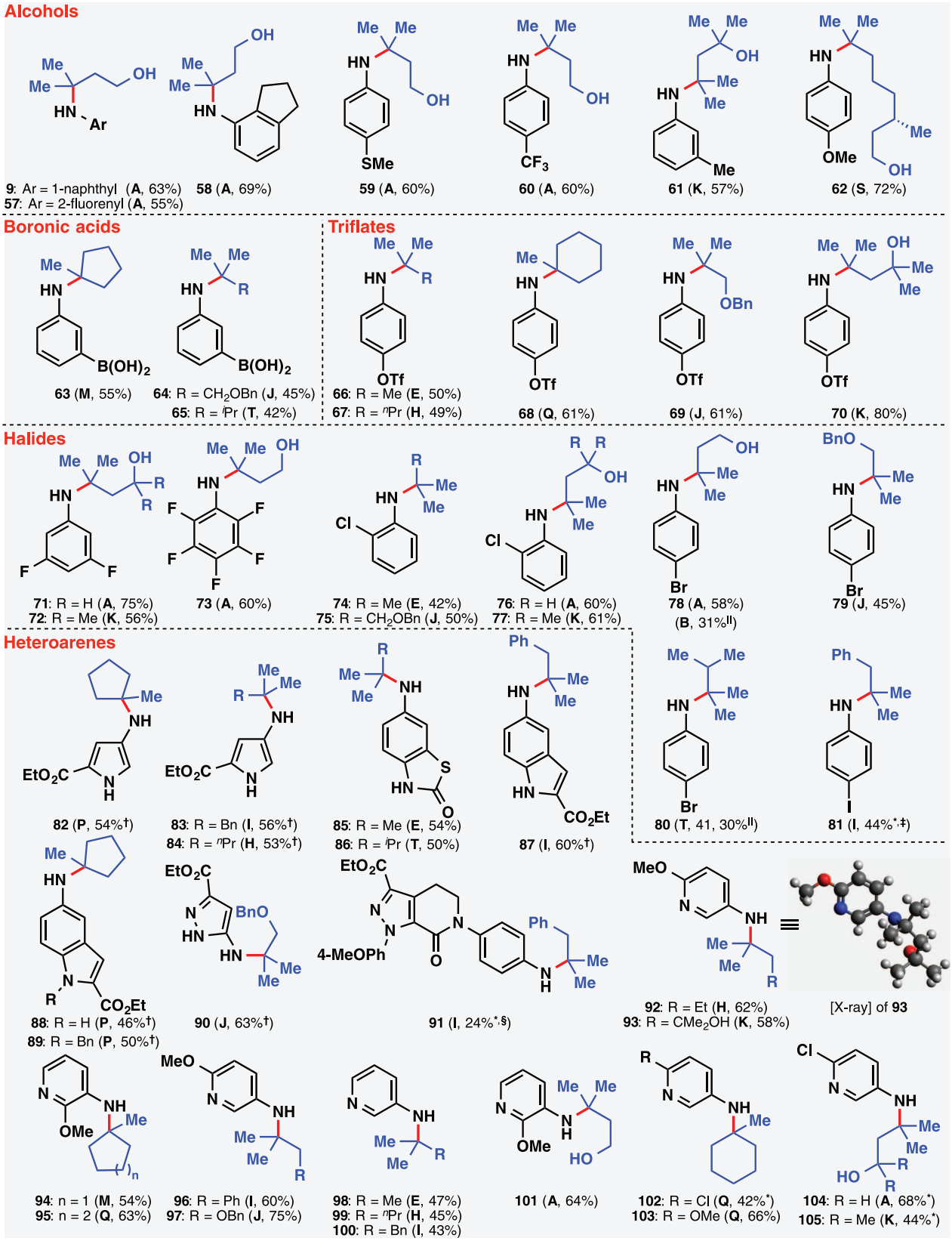
Fig. 2. Olefins explored in the hydroamination process. Bn, benzyl; D, deuterium.

The ability of the reaction to tolerate sensitive functional groups in the nitroarene component is remarkable (Fig. 3). Simple aliphatic functionality (**12**–**15**), thioethers (**16**–**19**), ethers (**20**–**33**), and amides (**34**–**40**) emerge unscathed. Although this is a reductive process, ketones (**41**–**49**) are not reduced, as opposed to the classic carbonyl-amine reductive amination that would require ketones to be protected. Free alcohols and amines are well tolerated (**9**, **50**–**62**), including the showcased example in Fig. 1A, delivering target **1** in 70% isolated yield without the need for protecting group chemistry. Unprotected boronic acids (**63**–**65**), aryl triflates (**66**–**70**), and aryl halides (**71**–**81**: F, Cl, Br, and I) are also tolerated, allowing for downstream C–C, C–O, and C–N cross-coupling chemistry. Most important, nitroheteroarenes can be used to deliver medicinally relevant building blocks containing pyrrole (**82**–**84**), benzothiazolone (**85**, **86**), indole (**87**–**89**), pyrazole (**90**, **91**), indazole (**108**; Fig. 5), triazole (**126**; Fig. 6), and pyridine (**92**–**105**) ring systems. The reaction was also performed on a decagram scale by Kemxtree, a contract research organization responsible for the commercialization of these amine building blocks, to provide dozens of adducts, five of which (**26**, **27**, **39**, **78**, and **80**) are shown in Figs. 3 and 4. Although *tert*-butylated anilines similar to **13** and **24** have been previously made via aminations of



*Without Zn/HCl reduction procedure. †Olefin (5 equiv), $PhSiH_3$ (3 equiv). ‡Olefin (2 equiv), $PhSiH_3$ (3 equiv). §Olefin (2 equiv), $PhSiH_3$ (2 equiv). IIPerformed on decagram scale with no additional optimization.

Fig. 3. Scope of the olefin hydroamination. Isolated yields are shown in parentheses along with the donor olefin used. Standard conditions: nitro(hetero)arene (1 equiv), olefin (3 equiv), $Fe(acac)_3$ (30 mol %), $PhSiH_3$ (2 equiv), EtOH, 60°C, 1 hour; Zn (20 equiv), $HCl_{(aq)}$, 60°C, 1 hour. Bu, butyl; Pr, propyl.



*Without Zn/HCl reduction procedure. †Olefin (5 equiv), $PhSiH_3$ (3 equiv). ‡Olefin (2 equiv), $PhSiH_3$ (3 equiv). §Olefin (2 equiv), $PhSiH_3$ (2 equiv). IIPerformed on decagram scale with no additional optimization.

Fig. 4. Scope of the olefin hydroamination, continued. Isolated yields are shown in parentheses along with the donor olefin used. Standard conditions: nitro(hetero)arene (1 equiv), olefin (3 equiv), $Fe(acac)_3$ (30 mol %), $PhSiH_3$ (2 equiv), EtOH, 60°C, 1 hour; Zn (20 equiv), $HCl_{(aq)}$, 60°C, 1 hour.

carbocations, these processes require either specialized reagents (*37*) or lack the chemoselectivity of the nitroarene-based approach (*38*). Finally, the reaction could also be used to construct adducts that, upon initial inspection, might be accessed using conventional amine-carbonyl reductive amination (**48**, **49**, and **53**–**56**). However, the presence of unprotected carbonyl groups, alcohols, and amines would impede such an approach, thus demonstrating the orthogonality of the hydroamination process to the classical method.

Application to medicinal chemistry targets

Figure 5 depicts three representative examples of how this reaction can simplify the preparation of hindered amine drug candidates: (i) Glucocorticoid receptor modulator intermediate **108** has previously been prepared from nitroindazole **106** in two steps (Fe-mediated reduction followed by ring opening of aziridine **107**) in 24% isolated yield (*39*). Alternatively, direct hydroamination of the same starting material using readily available olefin **U** affords the same target in a single operation (52% isolated yield). (ii) The HIV-1 reverse transcriptase inhibitor intermediate **111** is known to be accessible from nitropyridine **109** using three different transition metals and an expensive, water-sensitive alkylating agent (**110**) in 43% yield over three steps (*40*). The same adduct can be obtained in a single step from the same starting material using a feedstock olefin donor, 2-methyl-2-butene (**R**). (iii) The ORL1 (opioid receptor–like) receptor inhibitor intermediate **113** has previously been prepared by a three-step route involving condensation with ketone **112**, alkyl lithium addition, and deprotection in 37% overall yield (*41*). Alternatively, olefin **114** can react directly with nitrobenzene to deliver the same adduct in similar isolated yield in only 2 hours.

As shown in Fig. 6, there are many opportunities for this reaction to be applied in unusual ways. Cascade amine annulation can be accomplished in the case of olefin **V**, wherein a tandem olefin hydroamination takes place followed by an intramolecular amine-carbonyl reductive amination to deliver the highly substituted *N*-arylpiperidine **116** in 43% isolated yield (Fig. 6A). Electron-deficient olefins can also be used in instances where conjugate addition fails, as exemplified by the synthesis of β-amino derivatives **118** and **120**–**123**, key intermediates in a current medicinal chemistry program at Bristol-Myers Squibb that were otherwise inaccessible via hetero–Michael addition of 2-chloroaniline (**119**) to enone **W** (Fig. 6B). Additionally, this transformation also holds great appeal for isotopic labeling efforts (Fig. 6C). The use of deuterated isobutylene (**AA**) provides deuterated *tert*-butyl intermediate **124**, which can be used as an internal standard for liquid chromatography–mass spectrometry analysis in an ongoing program at Bristol-Myers Squibb. Simple deuterated building blocks of potential interest in medicine (*42*) can also be accessed, such as amine **125** in 77% yield. This single-step procedure obviates the need for costly and time-consuming multistep processes to access these labeled compounds. It could also be applied to the synthesis of radiolabeled products using appropriately labeled ^{3}H or ^{14}C alkenes. Finally, we examined the suitability of this reaction for use in a process setting (Fig. 6D). Because nitroarenes are potentially energetic materials, the temperature profile of the hydroamination to form **126** was studied at 20°C by heat flow calorimetry. No temperature spikes were observed when the catalyst was added; however, there was a ~2°C internal temperature rise upon addition of the $PhSiH_3$ that slowly dissipated over 2 hours as the reaction reached completion. These results indicate the absence of an induction period that could lead to a possible runaway thermal event during large-scale hydroaminations and serve to alleviate some of the concerns when performing the reaction in a process setting.

Substrate limitations

Although this reaction is exceedingly general in its current form, it is not without limitations. For example, nitroalkanes (**127**–**130**; Fig. 7) routinely give low yields of amines; however, it is worth noting that diamines **129** and **130** might not be trivial to directly make in other ways. Products arising from the use of tertiary nitroalkanes were not isolable. During the course of exploring substrate scope (Figs. 3 and 4), we found that esters in the *ortho* position relative to the nitro group were hydrolyzed (**131**), free (thio)phenols (**132**) inhibited the reaction, 2-nitropyridines were not tolerated (**133**), nitroimidazoles (**134**) led to a complex mixture of products, and styrene donors (**135**) gave trace products. Another clear drawback is the need for 2 to 3 equivalents of the olefin, making this chemistry best suited for instances when the nitroarene is more valuable. The main by-products of this reaction include reduction of the nitroarene to the aniline and *N*, *O*-alkylated products (e.g., **10**; Fig. 1) resulting from incomplete reduction, both of which are easily separable by conventional chromatography. Although the average yield of the olefin hydroamination in Fig. 3 is 55%, it is reasonable given the difficulties in purifying free amines and the challenging structures that are accessed. As a comparison, a recent exhaustive review of classical amine-carbonyl reductive amination reported an average yield of ~68% for 564 simple cyclic ketone examples conducted after 1999 (*4*).

Results from preliminary mechanistic experiments suggest that the olefin hydroamination takes place via initial reduction of the nitro(hetero)arene to the corresponding nitroso(hetero)arene, which then forms an adduct with alkyl radicals derived from the donor olefins. Cleavage of the resultant N–O σ bond then liberates the desired hindered secondary amine (see figs. S1 and S2).

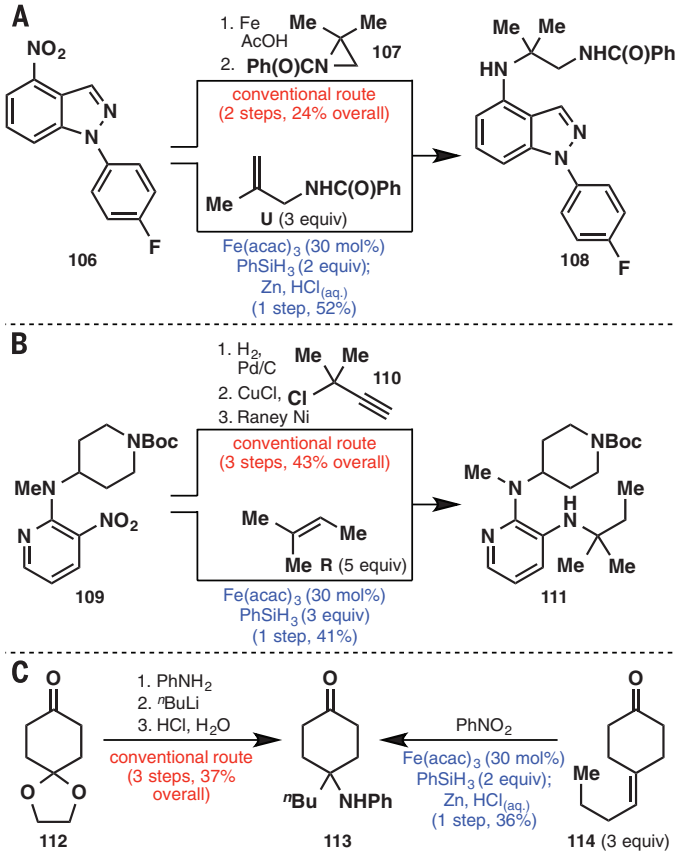


Fig. 5. Olefin hydroamination applied to shorter syntheses of known pharmaceutical targets. (**A**) Glucocorticoid receptor modulator intermediate **108**. (**B**) HIV-1 reverse transcriptase inhibitor intermediate **111**. Boc, *tert*-butyloxycarbonyl. (**C**) ORL1 receptor inhibitor intermediate **113**.

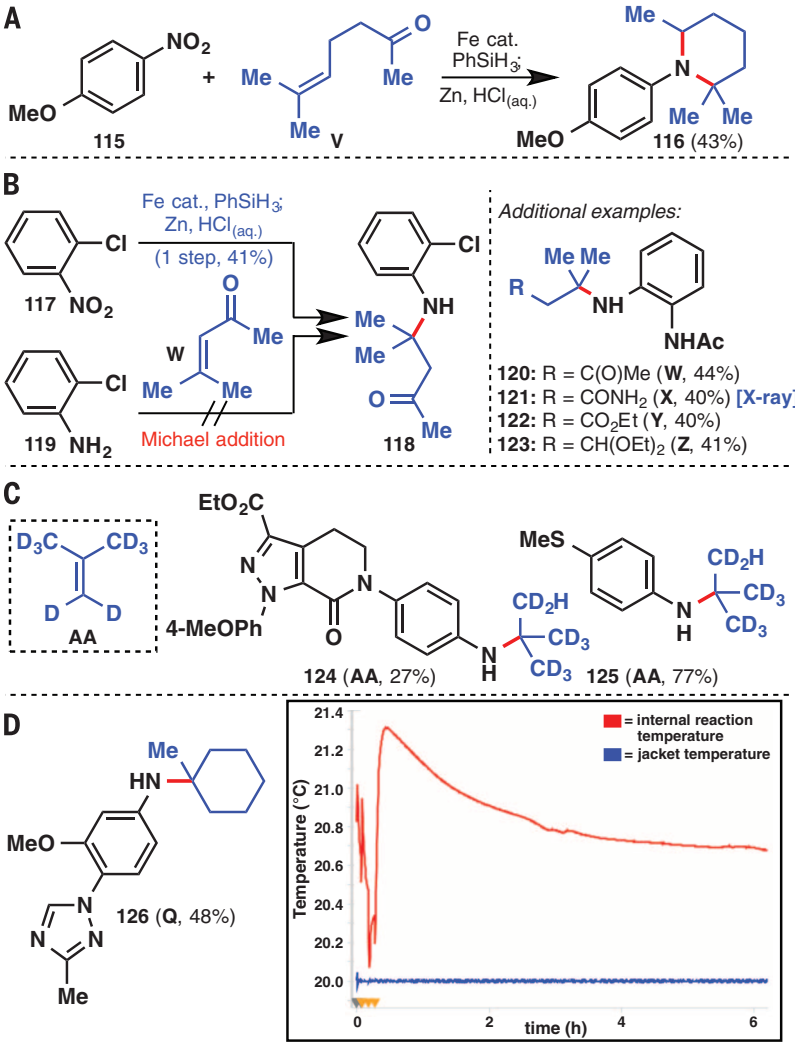
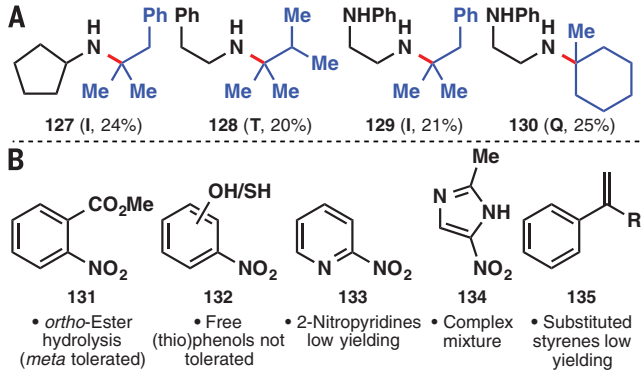


Fig. 6. Creative uses of the olefin hydroamination technique. (**A**) Cascade reductive aminations for amine annulation. (**B**) An efficient method to access hindered Michael adducts. (**C**) Application to deuterium labeling using isobutylene-d_8 (**AA**). (**D**) Temperature profile of the formation of a current intermediate in process chemistry over 6 hours shows the absence of an induction period.

Fig. 7. Limitations of the hydroamination. (**A**) Nitroalkanes give low yields. (**B**) Limitations of functional group tolerance.



REFERENCES AND NOTES

1. S. D. Roughley, A. M. Jordan, *J. Med. Chem.* **54**, 3451–3479 (2011).
2. C. M. Marson, P. Savy, in *Comprehensive Organic Functional Group Transformations II*, C. Ramsden, Ed. (Elsevier, Oxford, 2005), vol. 2, pp. 255–257.
3. G. Guillena, D. J. Ramón, M. Yus, *Chem. Rev.* **110**, 1611–1641 (2010).
4. E. W. Baxter, A. B. Reitz, *Org. React.* **59**, 1–714 (2004).
5. D. L. Boger, J. S. Panek, *Tetrahedron Lett.* **25**, 3175–3178 (1984).
6. D. S. Surry, S. L. Buchwald, *Angew. Chem. Int. Ed.* **47**, 6338–6361 (2008).
7. J. F. Hartwig, *Acc. Chem. Res.* **41**, 1534–1544 (2008).
8. S. V. Ley, A. W. Thomas, *Angew. Chem. Int. Ed.* **42**, 5400–5449 (2003).
9. Y. Ishihara, A. Montero, P. S. Baran, *The Portable Chemist's Consultant* (Apple Publishing Group, Cupertino, CA, 2013).
10. P. Ruiz-Castillo, D. G. Blackmond, S. L. Buchwald, *J. Am. Chem. Soc.* **137**, 3085–3092 (2015).
11. G. A. Russell, C.-F. Yao, *Heteroatom Chem.* **4**, 433–444 (1993).
12. E. G. Janzen, *Acc. Chem. Res.* **4**, 31–40 (1971).
13. N. Inamoto, O. Simamura, *J. Org. Chem.* **23**, 408–410 (1958).
14. T. Hosogai, N. Inamoto, R. Okazaki, *J. Chem. Soc. C* **1971**, 3399–3401 (1971).
15. B. A. Gingras, W. A. Waters, *J. Chem. Soc.* 1920–1924 (1954).
16. E. J. Corey, A. W. Gross, *J. Org. Chem.* **50**, 5391–5393 (1985).
17. S. Zhu, N. Niljianskul, S. L. Buchwald, *J. Am. Chem. Soc.* **135**, 15746–15749 (2013).
18. R. P. Rucker, A. M. Whittaker, H. Dang, G. Lalic, *J. Am. Chem. Soc.* **134**, 6571–6574 (2012).
19. D. Zhu *et al.*, *Angew. Chem. Int. Ed.* **54**, 2497–2500 (2015).
20. N. Ono, *The Nitro Group in Organic Synthesis* (Wiley, New York, 2001).
21. J. I. G. Cadogan, M. Cameron-Wood, R. K. Mackie, R. J. G. Searle, *J. Chem. Soc.* 4831–4837 (1965).
22. H. Gao, Q.-L. Xu, M. Yousufuddin, D. H. Ess, L. Kürti, *Angew. Chem. Int. Ed.* **53**, 2701–2705 (2014).
23. G. Bartoli, R. Dalpozzo, M. Nardi, *Chem. Soc. Rev.* **43**, 4728–4750 (2014).
24. Q.-L. Xu, H. Gao, M. Yousufuddin, D. H. Ess, L. Kürti, *J. Am. Chem. Soc.* **135**, 14048–14051 (2013).
25. H. Gao, D. H. Ess, M. Yousufuddin, L. Kürti, *J. Am. Chem. Soc.* **135**, 7086–7089 (2013).
26. J. C. Lo, Y. Yabe, P. S. Baran, *J. Am. Chem. Soc.* **136**, 1304–1307 (2014).
27. J. C. Lo, J. Gui, Y. Yabe, C.-M. Pan, P. S. Baran, *Nature* **516**, 343–348 (2014).
28. S. Isayama, T. Mukaiyama, *Chem. Lett.* **18**, 1071–1074 (1989).
29. S. W. M. Crossley, F. Barabé, R. A. Shenvi, *J. Am. Chem. Soc.* **136**, 16788–16791 (2014).
30. S. M. King, X. Ma, S. B. Herzon, *J. Am. Chem. Soc.* **136**, 6884–6887 (2014).
31. J. Waser, B. Gaspar, H. Nambu, E. M. Carreira, *J. Am. Chem. Soc.* **128**, 11693–11712 (2006).
32. K. Iwasaki, K. K. Wan, A. Oppedisano, S. W. M. Crossley, R. A. Shenvi, *J. Am. Chem. Soc.* **136**, 1300–1303 (2014).
33. P. Magnus, M. J. Waring, D. A. Scott, *Tetrahedron Lett.* **41**, 9731–9733 (2000).
34. E. K. Leggans, T. J. Barker, K. K. Duncan, D. L. Boger, *Org. Lett.* **14**, 1428–1431 (2012).
35. T. J. Barker, D. L. Boger, *J. Am. Chem. Soc.* **134**, 13588–13591 (2012).
36. D. L. Romero *et al.*, *J. Med. Chem.* **39**, 3769–3789 (1996).
37. J. W. Cran, D. V. Vidhani, M. E. Krafft, *Synlett* **25**, 1550–1554 (2014).
38. J. R. Gage, J. M. Wagner, *J. Org. Chem.* **60**, 2613–2614 (1995).
39. H. Bladh *et al.*, PCT Int. Appl. WO 2007046747 (2007).
40. D. L. Romero *et al.*, *J. Med. Chem.* **39**, 3769–3789 (1996).
41. S. Schunk *et al.*, PCT Int. Appl. WO 2008101659 (2008).
42. T. G. Gant, *J. Med. Chem.* **57**, 3595–3611 (2014).

ACKNOWLEDGMENTS

Supported by a postdoctoral fellowship from the Shanghai Institute of Organic Chemistry, Zhejiang Medicine Co., and Pharmaron (J.G.); a China Scholarship Council and Jilin Medical College visiting scholar fellowship (Y.J.); a NSF graduate fellowship (J.C.L.); and National Institute of General Medical Sciences grant GM-097444, Sigma-Aldrich, and Bristol-Myers Squibb. We thank D.-H. Huang and L. Pasternack for assistance with NMR spectroscopy; A. Rheingold, C. E. Moore, and M. A. Galella for x-ray crystallographic analysis; and M. Eastgate, K. Chen, J. Kadow, and B. Maxwell for helpful discussions. Metrical parameters for the structures of **93** and **121** are available free of charge from the Cambridge Crystallographic Data Center under reference numbers CCDC-1050751 and CCDC-1050750, respectively.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/886/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S20
Tables S1 to S4
References (*43*–*50*)

2 March 2015; accepted 7 April 2015
10.1126/science.aab0245

REPORTS

PLANETARY SCIENCE

Low-altitude magnetic field measurements by MESSENGER reveal Mercury's ancient crustal field

Catherine L. Johnson,[1,2]* Roger J. Phillips,[3] Michael E. Purucker,[4] Brian J. Anderson,[5] Paul K. Byrne,[6,13] Brett W. Denevi,[5] Joshua M. Feinberg,[7] Steven A. Hauck II,[8] James W. Head III,[9] Haje Korth,[5] Peter B. James,[10] Erwan Mazarico,[4] Gregory A. Neumann,[4] Lydia C. Philpott,[1] Matthew A. Siegler,[2,11] Nikolai A. Tsyganenko,[12] Sean C. Solomon[10,13]

Magnetized rocks can record the history of the magnetic field of a planet, a key constraint for understanding its evolution. From orbital vector magnetic field measurements of Mercury taken by the MErcury Surface, Space ENvironment, GEochemistry, and Ranging (MESSENGER) spacecraft at altitudes below 150 kilometers, we have detected remanent magnetization in Mercury's crust. We infer a lower bound on the average age of magnetization of 3.7 to 3.9 billion years. Our findings indicate that a global magnetic field driven by dynamo processes in the fluid outer core operated early in Mercury's history. Ancient field strengths that range from those similar to Mercury's present dipole field to Earth-like values are consistent with the magnetic field observations and with the low iron content of Mercury's crust inferred from MESSENGER elemental composition data.

Mercury is the only inner solar system body other than Earth that currently possesses a global magnetic field generated by a dynamo in a fluid metallic outer core (*1*, *2*). Mercury's field is dipolar, weak (surface field strength ~1% that of Earth's), axially symmetric, and equatorially asymmetric (*3*–*5*). These attributes may indicate an intrinsic north-south asymmetry in the dynamo (*6*). The basic characteristics of the magnetic field have persisted for at least the past ~40 years, the duration of the era of spacecraft exploration of Mercury (*7*), but whether a field was present over longer time scales has been unknown. We show here that Mercury's core dynamo field was also present early in the planet's history, providing critical information on Mercury's interior thermal and dynamic evolution.

[1]Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada. [2]Planetary Science Institute, Tucson, AZ 85719, USA. [3]Planetary Science Directorate, Southwest Research Institute, Boulder, CO 80302, USA. [4]NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA. [5]The Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723, USA. [6]Lunar and Planetary Institute, Houston, TX 77058, USA. [7]Institute for Rock Magnetism, Department of Earth Sciences, University of Minnesota, Minneapolis, MN, 55455, USA. [8]Department of Earth, Environmental, and Planetary Sciences, Case Western Reserve University, Cleveland, OH, 44106, USA. [9]Department of Earth, Environmental and Planetary Sciences, Brown University, Providence, RI 02912, USA. [10]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA. [11]Department of Earth Sciences, Southern Methodist University, Dallas, TX 75205, USA. [12]Institute and Faculty of Physics, Saint Petersburg State University, Saint Petersburg, Russia. [13]Department of Terrestrial Magnetism, Carnegie Institution of Washington, Washington, DC 20015, USA.
*Corresponding author. E-mail: cjohnson@eos.ubc.ca

Magnetized rocks and the fields that result from them are key records of a planet's magnetic field history. Igneous rocks that cool in the presence of an ambient magnetic field can acquire a permanent or remanent magnetization determined by their mineralogy and by the strength and geometry of the magnetizing field. Such magnetization can be altered by subsequent tectonic activity, reheating, burial, shock, or chemical reactions. Lateral variations in the strength or direction of magnetization, or in the thickness or depth of the magnetized layer, give rise to magnetic anomalies that may be detected on or above the surface. Detection of these anomalies depends on the strength and horizontal scale of the magnetization contrasts and on the distance of the observation platform from the magnetized source.

Orbital observations of Mercury's magnetic field by the Magnetometer on the MErcury Surface, Space ENvironment, GEochemistry, and Ranging (MESSENGER) spacecraft have been made from March 2011 to April 2015. MESSENGER's orbit was highly eccentric, and until 2014, minimum (periapsis) altitudes were 200 to 500 km. Fields resulting from remanent crustal magnetization have not been detected in these observations, a result suggesting that remanent magnetization is weak to nonexistent, or coherent only over spatial scales less than a few hundred kilometers.

Magnetic field measurements were obtained by MESSENGER at spacecraft altitudes less than 200 km starting in April 2014. Mercury's offset-axial dipole core field, and fields from the magnetopause and magnetotail current systems and other external sources, dominate the observations (*4*, *5*, *8*, *9*). We estimated these contributions for each orbit, using magnetospheric models developed with MESSENGER data (*5*, *9*), and subtracted them from the vector magnetic field measurements. The remaining signals have magnitudes of a few tens of nT and wavelengths of several hundred to ~1500 km and change substantially from one orbit to the next. They originate mainly from processes operating above the surface of Mercury (*5*, *9*, *10*). These fields mask any smaller-amplitude, shorter-wavelength signals from the planet's interior. We estimated the long-wavelength signals empirically on an orbit-by-orbit basis and removed them by the application of a high-pass filter (*10*) tuned to best separate the short- and long-wavelength signals (Fig. 1).

Typically, the high-pass filtered (HPF) data show either no signals or signals that are correlated with increased variability in the total field at frequencies above 1 Hz. The latter—e.g., those during the time period 1200 to 1260 s in Fig. 1, B and C—are not of internal origin. However, for some orbits, the HPF data show smoothly varying signals that have amplitudes more than three times that of the high-frequency variability. These signals are found close to periapsis (Fig. 1D) and are typically observed on multiple successive orbits (e.g., Fig. 1 and figs. S1 to S4).

We have detected radial (ΔB_r) and colatitudinal (north-south, ΔB_θ) HPF signals with these characteristics over the two regions where MESSENGER periapsis altitudes were lowest (~25 km) in 2014: the Suisei Planitia region (Fig. 2) and a region south of the lobate scarp Carnegie Rupes (*10*). We also detected weaker signals, less than 3 nT in amplitude, over a third region near ~170°E, at times close to periapsis and altitudes of ~95 to ~130 km in 2014. Clear detections have been made on only nightside or dawn-dusk tracks because of lower high-frequency variability in external fields than on the dayside.

Coherent signals across the Suisei Planitia region obtained in orbits from September 2014 display peak amplitudes of ~12 nT at spacecraft altitudes of 27 km, north of Shakespeare basin (Fig. 2A). The dominant wavelength of the signals is ~320 km, but shorter-wavelength signals are also observed. We verified that these results are insensitive to the precise choice of the HPF characteristics, to first order (*10*) (fig. S5), and that the magnetospheric activity index (*11*) was not unusually high during most orbits (*10*) (fig. S6).

The eastern extent of the signals is well constrained by the MESSENGER data, with no signals observed at ~60°N, east of Kosho crater (~220°E), even though periapsis altitudes were below 30 km eastward to 240°E. The westernmost extent (Fig. 2A) corresponds to an orbit-correction maneuver (OCM) that raised periapsis altitude from 25 to 94 km. No signals were detected on orbits immediately following the OCM (fig. S4); this altitude dependence suggests that the source of the fields is internal. Furthermore, the dominant wavelengths of the signals observed at the lowest spacecraft altitudes are consistent with source depth estimates of 7 to 45 km, suggesting magnetized rocks as the source of the observed fields rather than contributions from

the core (*10*). Confirmation of an internal origin is provided by the weaker signals observed at altitudes of 60 to 100 km (fig. S7A) and the absence of signals at ~150 km altitude. These observations are consistent with the upward attenuation of signals from the lowest altitudes predicted for an internal source (*10*) (fig. S7B). Finally, signals very similar in character to those in Fig. 2A were observed over the region in March 2015, at the same local times as September 2014 and at spacecraft altitudes of 14 to 40 km. Larger amplitudes were observed within ~5° latitude of periapsis (~59°N), reflecting the 10-km-lower periapsis altitude. Low-altitude observations from March extend to the western edge of the region and show signals west of Verdi crater, in particular over the adjacent volcanic smooth plains. All data obtained at spacecraft altitudes below 60 km are shown in Fig. 2B.

The HPF signals are seen over, but are not restricted to, regions of lower topography (Fig. 2A and fig S8). In the Suisei Planitia region, signals are seen over regions of both smooth plains and older intercrater plains (Fig. 2B) (*12*, *13*). The largest-amplitude ΔB_r values are spatially associated with smooth plains (Fig. 2B) (*10*). There are no obvious features associated with impact craters in the Suisei region, and no clear signals at the edge of the Borealis basin (fig. S8) have yet been observed, although there are weak signals over the eastern interior of the basin. Contractional structures CS1 and CS2 (Fig. 2A) indicate local association of the signals with tectonic features, but many structures in the region (*14*), such as CS3 and CS4, have no associated ΔB_r signals (Fig. 2A). Similarly, no coherent signals have been seen across Carnegie Rupes (fig. S8). Our observations are consistent with sources at depth that may include a combination of magnetized intrusive material and magnetization contrasts across deep-seated crustal structures (e.g., faults). Features associated with mapped tectonic structures (e.g., the local maximum in ΔB_r over CS1) may reflect sources at shallower depths.

Constraining the time of acquisition of magnetic remanence is difficult because the signals do not correlate with regions of distinctive

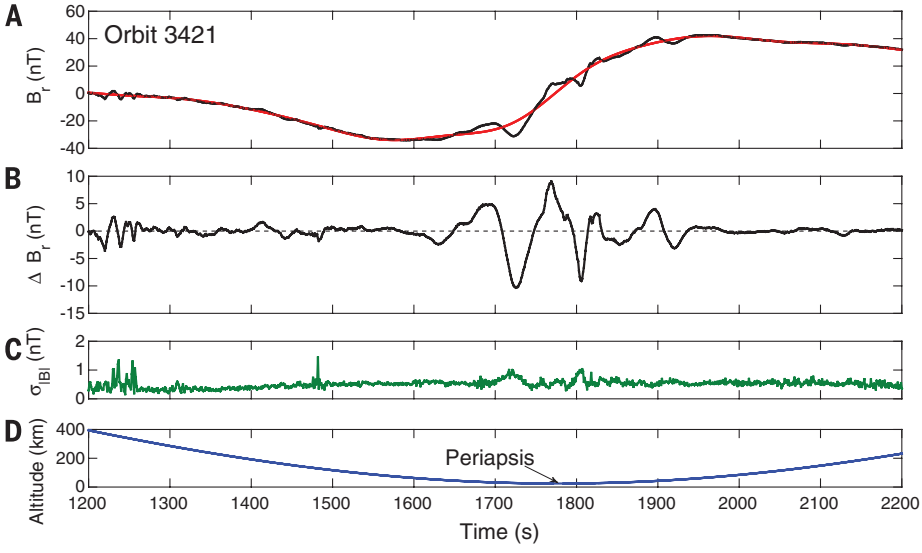


Fig. 1. Magnetic field observations from 8 September 2014 (orbit 3421). (**A**) Radial component of the field, B_r (black), in the Mercury body-fixed frame (*10*) after subtraction of the modeled magnetopause, magnetotail, and offset axial dipole fields, and the low-pass filtered signal (red). (**B**) HPF signal, ΔB_r. (**C**) High-frequency (>1 Hz) variability in the total field, $\sigma_{|B|}$, a measure of the external field noise remaining in the HPF signals. (**D**) Spacecraft altitude. Periapsis altitude was 25 km. 100 s corresponds to a horizontal scale of ~385 km at periapsis. The orbit track is labeled on Fig. 2A.

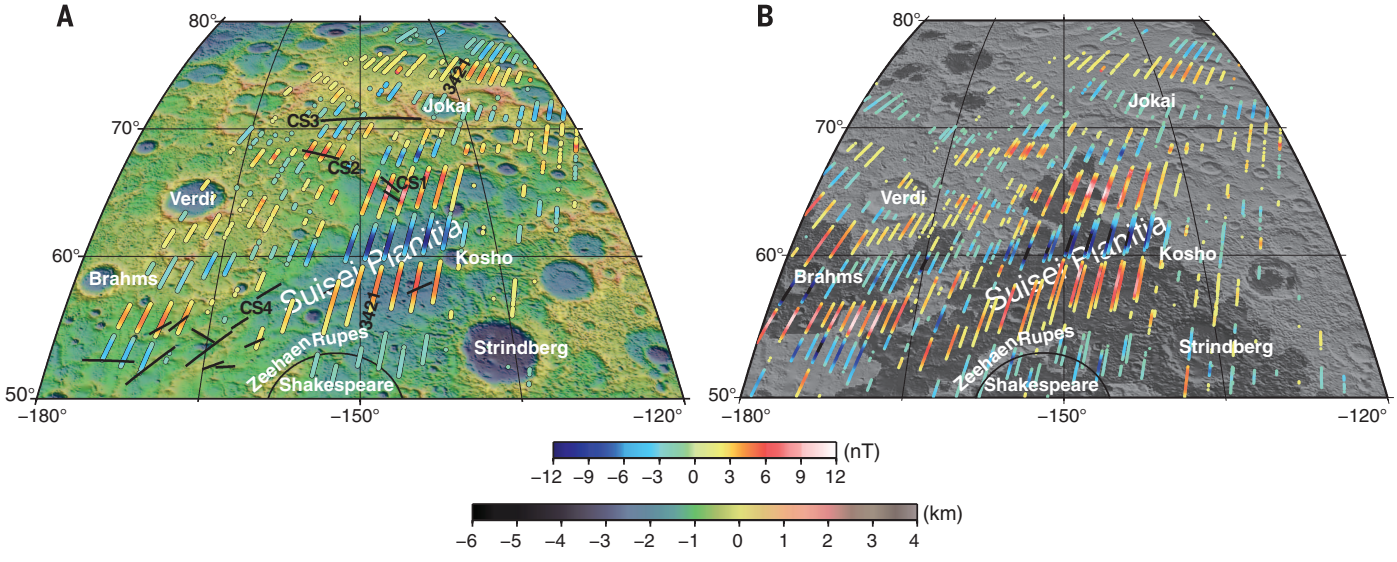


Fig. 2. HPF radial magnetic field, ΔB_r, over Suisei Planitia. The HPF signals shown satisfy $|\Delta B_r| \geq 1$ nT and $|\Delta B_r|/\sigma_{|B|} \geq 3$. Underlying image is of topography derived from Mercury Laser Altimeter measurements (Mollweide projection). Color bars give ΔB_r (nT) and topography (km). 1° of latitude on Mercury corresponds to 43 km. (**A**) Orbits 3411 to 3433 (from September 2014), excluding orbit 3424 (high magnetospheric activity). The time interval between successive orbits is 8 hours. Orbit 3421 (Fig. 1) is labeled. Periapsis local times were 06:00 to 08:30 hours. Spacecraft altitudes were 25 to 60 km. The Shakespeare basin and contractional structures at least 50 km in length, with a strike making an angle greater than 45° to the orbit track, are shown in black. CS1 to CS4 are contractional structures. (**B**) Orbits 3411 to 3433 (as in Fig. 2A) and orbits 3928 to 3940 (from March 2015) at spacecraft altitudes of 14 to 40 km. Periapsis local times were 06:00 to 08:30 hours for all orbits. Underlying image shows the smooth plains in dark gray (*12*) and intercrater plains in light gray. The observations from March 2015 show the repeatability of the signals observed in Fig. 2A and higher amplitudes associated with the lower spacecraft altitudes (peak amplitudes of 20 nT observed at 15 km altitude).

surface ages. However, the presence of signals over relatively large areas (Fig. 2 and fig. S8) that encompass multiple geologic units, together with the observation that the largest-amplitude ΔB_r values occur over the smooth plains, suggests that the smooth plains, the youngest major volcanic deposits on Mercury emplaced 3.7 to 3.9 billion years ago (Ga) (*12*, *15*), provide a lower bound on the average age of magnetization. An average age substantially less than this figure would require processes that operated over large regions after smooth plains emplacement yet left no surficial expression. Such processes could include pervasive intrusions at depth that remained below the Curie temperature of the magnetic carrier mineral(s) after cooling; reheating and subsequent cooling of older intrusive material; subsurface structural deformation of previously magnetized material; or some combination of the two. Although later acquisition of remanence may have occurred locally (e.g., during cooling of impact melt), the association of crustal remanence with diverse terrains suggests that much of the magnetization was acquired in an internal field before 3.7 to 3.9 Ga. The dominance of ΔB_r and ΔB_θ signals across groups of orbits, together with the ΔB_ϕ signals on orbits immediately to the east and west of these groups (*10*) (fig. S9), are consistent with a magnetization that is primarily in the north-south plane and is associated with features that are limited in their east-west extent. The simplest geometry for the field in which such a remanence was acquired is one that, like the current field, was symmetric about the planet's rotation axis (*10*).

The peak strength of the signals over Suisei Planitia provides a lower bound on the magnetization (M) within a layer of a given thickness (*10*). For thicknesses of 4 to 40 km, M values are 0.1 to 0.02 Am^{-1}, respectively, comparable to those inferred for the Moon (*16*). For thermoremanence, M reflects the combined effects of the strength of the magnetizing field ($B_{ancient}$) and the bulk magnetic properties of the crust, given by its thermoremanent susceptibility (χ_{TRM}). Values for $B_{ancient}$ and χ_{TRM} were calculated from the relation $M = \chi_{TRM} B_{ancient} / \mu_0$, where μ_0 is the magnetic permeability of free space, for layer thicknesses ranging from 1 to 100 km (*17*). The thermoremanent susceptibility of Mercury's crust is unknown, because it depends on the magnetic minerals present and on their relative volumetric abundances. The chemically reduced characteristics of Mercury's surface materials (*18*) suggest that iron metal, iron alloys, and iron sulfides are possible magnetic carriers. Given Mercury's low oxygen fugacity (*19*), the paramagnetic iron sulfide troilite is likely to be a more stable mineral than the ferromagnetic pyrrhotite. However, because knowledge of the petrology of Mercury's interior is limited, we evaluated the plausibility of pyrrhotite or a mineral with similar magnetic characteristics (χ_{TRM} and Curie temperature, T_c) as a potential magnetic phase. Susceptibilities for pyrrhotite, iron metal, and high-iron (EH) and low-iron (EL) enstatite chondrites were scaled for volume fractions of the magnetic carrier consistent with the 1.5 to 2 weight percent average iron content inferred from MESSENGER x-ray fluorescence observations (*10*, *20*, *21*). The results (Fig. 3) indicate that for magnetic layer thicknesses of 25 km, consistent with the average source depth and mean crustal thickness (Fig. 4) in the region, EH values for χ_{TRM} require surface field strengths about twice those of the present-day value in the Suisei region (~300 nT). The required field scales inversely with χ_{TRM} and with layer thickness. In particular, EL values for χ_{TRM} require a ~4500-nT field for a 25-km-thick layer.

The field values implied by the magnetic mineralogies for a given layer thickness (Fig. 3) are minima for two reasons. First, they are inferred from M, which is a lower bound on the magnetization, and second, they are derived under the assumption that all the iron is partitioned into magnetic phases. Earth-like fields (~50,000 nT) are permissible if χ_{TRM} is ~6×10^{-4} for a 25-km-thick layer, compatible with 0.1 to 5% of the iron partitioning into magnetic phases. Field strengths weaker than those today are unlikely, on the basis of the values of susceptibility required. Thus, ancient surface field strengths that lie between values comparable to those from Mercury's current dynamo and Earth-like values are most likely given the possible magnetic minerals in Mercury's crust.

We considered two alternative interpretations of the magnetization: first, that it reflects an

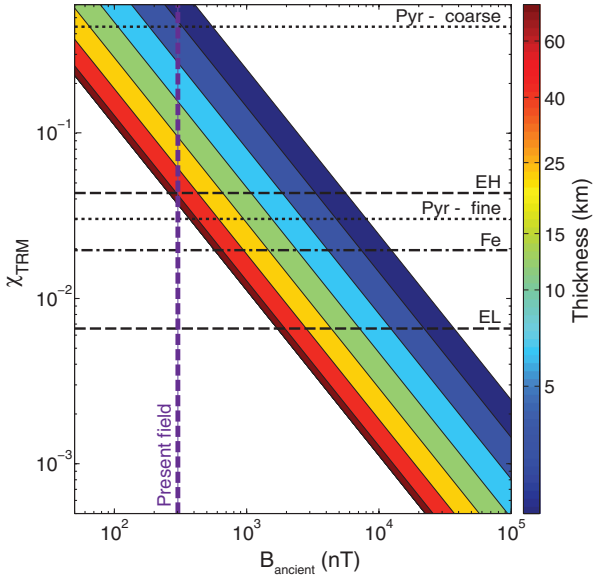


Fig. 3. Mercury's ancient field strength and magnetic mineralogy. Magnetic susceptibility (χ_{TRM}) and magnetizing field strength ($B_{ancient}$) required to produce the observed peak HPF magnetic field strength over Suisei Planitia. Susceptibilities (*10*) for pyrrhotite for grain sizes ranging from 5 to 250 μm (black dotted lines), multidomain iron (black dashed-dot line), and EH and EL enstatite chondrites (black dashed lines) are shown. The values are scaled for volume fractions of the magnetic carrier consistent with the average iron content inferred from MESSENGER observations. The surface field strength for the present internal dipole (vertical dashed line) is that for the average magnetic latitude of the region (46° N).

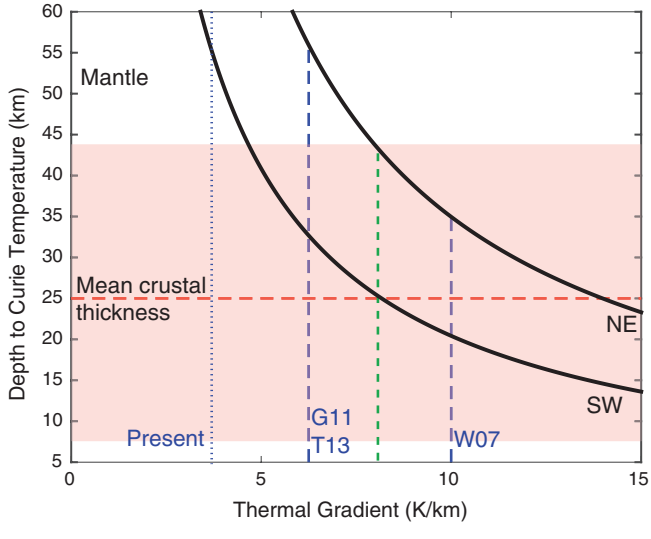


Fig. 4. Depth to the Curie temperature versus thermal gradient on Mercury. Depth to the Curie temperature (T_c = 325°C) for pyrrhotite across the Suisei Planitia region (black lines) as a function of thermal gradient. The two lines reflect the range in depths associated with a decrease in the mean surface temperature from the southwest [closer to the hot pole (*10*) at 0°N, 180°E] to the northeast of the region (fig. S11A). Curves for magnetic minerals with higher T_c lie to right of those shown here (fig. S11B). Blue dashed lines denote the range in thermal gradients at 4 Ga permitted by thermal evolution models W07 (*26*), G11 (*27*), and T13 (*28*); the blue dotted line is the average present value from thermal models (*26*–*28*). The mean crustal thickness (red dashed line) and the range in the mean value (red shaded region) in the region are shown (*22*–*24*). A thermal gradient less than that shown by the green dashed line implies that the uppermost mantle is cooler than the Curie temperature.

induced magnetization in the present field and, second, that it could be a viscous remanent magnetization (VRM) acquired during prolonged exposure of the magnetic minerals to the planetary field and hence reflecting an unknown, but younger, age than that of the smooth plains. Although both of these physical processes are likely to operate, induced magnetizations cannot fully explain the observed HPF field strengths, and the net effect of VRM will be that our estimates of ancient field strength are lower bounds (*10*).

Within the range of uncertainty of crustal thickness (*22*–*24*) and magnetized layer source depths (*10*), most or all of the magnetization could reside within Mercury's crust (Fig. 4). We investigated whether such a scenario is consistent with thermal evolution models, given magnetizations acquired at ~4 Ga. We estimated the depth to T_c for a range of thermal gradients (Fig. 4). The Curie temperature was taken to be 325°C (that of pyrrhotite) as a conservatively low value for our calculations, and we used the maximum average daily surface temperature predicted for a range of Mercury's orbital eccentricities from 0 to 0.4 (*10*, *25*). The results indicate that even for high thermal gradients at 4 Ga (*26*) the depth to T_c in the Suisei Planitia region is at least 20 km. For thermal gradients less than 8 K/km and upper limits on the crustal thickness in the region, the entire crust remains below T_c. These results imply that acquisition and subsequent preservation of an ancient crustal remanence by magnetic carriers with T_c values of at least 325°C are consistent with thermal models (*10*, *26*–*28*), and for carriers with higher T_c some remanence may be carried by upper mantle material. Such a conclusion is predicated on the assumption that the surface temperature pole locations have remained stationary in a body-fixed coordinate system since the time that the remanent magnetization was acquired (*10*). The symmetry of the ancient field with respect to the present rotation axis supports such a presumption by suggesting that, since that epoch, there has been no substantial reorientation of the crust ("true polar wander") with respect to the planet's axis of greatest moment of inertia.

The simplest interpretation of the results presented here is that a core dynamo was present early in Mercury's history. If the dynamo was thermochemically driven [e.g., (*6*, *29*)], this finding provides a strong constraint on models for the thermal evolution of Mercury's interior. In particular, the existence of a core dynamo at the time of smooth plains emplacement presents a new challenge to such models. An early core dynamo can be driven by superadiabatic cooling of the liquid core, but in typical thermal history models this phase has ended by 3.9 Ga. A later dynamo can be driven by the combined effects of cooling and compositional convection associated with formation of a solid inner core (*26*–*28*), but in most thermal history models inner core formation does not start until well after 3.7 Ga. Further progress in understanding the record of Mercury's ancient field can also be made with improved petrological constraints on crustal compositions [e.g., (*30*)], information on the candidate magnetic mineralogies implied, and knowledge of their magnetic properties.

REFERENCES AND NOTES

1. N. F. Ness, K. W. Behannon, R. P. Lepping, Y. C. Whang, K. H. Schatten, *Science* **185**, 151–160 (1974).
2. B. J. Anderson *et al.*, *Science* **321**, 82–85 (2008).
3. B. J. Anderson *et al.*, *Science* **333**, 1859–1862 (2011).
4. B. J. Anderson *et al.*, *J. Geophys. Res.* **117** (E12), E00L12 (2012).
5. C. L. Johnson *et al.*, *J. Geophys. Res.* **117** (E12), E00L14 (2012).
6. H. Cao *et al.*, *Geophys. Res. Lett.* **41**, 4127–4134 (2014).
7. L. C. Philpott *et al.*, *Geophys. Res. Lett.* **41**, 6627–6634 (2014).
8. B. J. Anderson *et al.*, *Geophys. Res. Lett.* **41**, 7444–7452 (2014).
9. H. Korth *et al.*, Modular model for Mercury's magnetospheric magnetic field confined within the average observed magnetopause. JGR Space Physics, doi10.1002/2015JA021022 (2015).
10. Materials and methods are available as supplementary materials on *Science* Online.
11. B. J. Anderson, C. L. Johnson, H. Korth, *Geochem. Geophys. Geosyst.* **14**, 3875–3886 (2013).
12. B. W. Denevi *et al.*, *J. Geophys. Res. Planets* **118**, 891–907 (2013).
13. J. L. Whitten, J. W. Head, B. W. Denevi, S. C. Solomon, *Icarus* **241**, 97–113 (2014).
14. P. K. Byrne *et al.*, *Nat. Geosci.* **7**, 301–307 (2014).
15. S. Marchi *et al.*, *Nature* **499**, 59–61 (2013).
16. M. E. Purucker, J. B. Nicholas, *J. Geophys. Res.* **115** (E12), E12007 (2010).
17. M. A. Wieczorek, B. P. Weiss, S. T. Stewart, *Science* **335**, 1212–1215 (2012).
18. L. R. Nittler *et al.*, *Science* **333**, 1847–1850 (2011).
19. F. M. McCubbin, M. A. Riner, K. E. Vander Kaaden, L. K. Burkemper, *Geophys. Res. Lett.* **39**, L09202 (2012).
20. L. G. Evans *et al.*, *J. Geophys. Res.* **117** (E12), E00L07 (2012).
21. S. Z. Weider, L. R. Nittler, R. D. Starr, T. J. McCoy, S. C. Solomon, *Icarus* **235**, 170–186 (2014).
22. E. Mazarico *et al.*, *J. Geophys. Res. Planets* **119**, 2417–2436 (2014).
23. P. B. James, M. T. Zuber, R. J. Phillips, S. C. Solomon, *J. Geophys. Res. Planets* **120**, 287–310 (2015).
24. S. Padovan, M. A. Wieczorek, J.-L. Margot, N. Tosi, S. C. Solomon, *Geophys. Res. Lett.* **42**, 1029–1038 (2015).
25. A. C. M. Correia, J. Laskar, *Icarus* **201**, 1–11 (2009).
26. J.-P. Williams, O. Aharonson, F. Nimmo, *Geophys. Res. Lett.* **34**, L21201 (2007).
27. M. Grott, D. Breuer, M. Laneuville, *Earth Planet. Sci. Lett.* **307**, 135–146 (2011).
28. N. Tosi, M. Grott, A.-C. Plesa, D. Breuer, *J. Geophys. Res. Planets* **118**, 2474–2487 (2013).
29. S. Stanley, G. A. Glatzmaier, *Space Sci. Rev.* **152**, 617–649 (2010).
30. K. E. Vander Kaaden, F. M. McCubbin, *J. Geophys. Res. Planets* **120**, 195–209 (2015).

ACKNOWLEDGMENTS

We thank the MESSENGER operations and engineering teams for enabling the low-altitude observations reported here. We are also grateful for the contributions of our friend and colleague M. H. Acuña whose expertise was critical to the Magnetometer development. The MESSENGER mission is supported by the NASA Discovery Program and the MESSENGER Participating Scientist Program. C.L.J. and L.C.P. also acknowledge support from the Natural Sciences and Engineering Research Council of Canada. Data from the MESSENGER mission are archived with the NASA Planetary Data System. We thank three reviewers for thoughtful comments that improved the manuscript.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/892/suppl/DC1
Supplementary Text
Figs. S1 to S11
Tables S1 and S2
References (31–53)

7 February 2015; accepted 28 April 2015
Published online 7 May 2015;
10.1126/science.aaa8720

CARBON CYCLE

The dominant role of semi-arid ecosystems in the trend and variability of the land CO_2 sink

Anders Ahlström,[1,2]* Michael R. Raupach,[3]† Guy Schurgers,[4] Benjamin Smith,[1] Almut Arneth,[5] Martin Jung,[6] Markus Reichstein,[6] Josep G. Canadell,[7] Pierre Friedlingstein,[8] Atul K. Jain,[9] Etsushi Kato,[10] Benjamin Poulter,[11] Stephen Sitch,[12] Benjamin D. Stocker,[13,14] Nicolas Viovy,[15] Ying Ping Wang,[16] Andy Wiltshire,[17] Sönke Zaehle,[6] Ning Zeng[18]

The growth rate of atmospheric carbon dioxide (CO_2) concentrations since industrialization is characterized by large interannual variability, mostly resulting from variability in CO_2 uptake by terrestrial ecosystems (typically termed carbon sink). However, the contributions of regional ecosystems to that variability are not well known. Using an ensemble of ecosystem and land-surface models and an empirical observation-based product of global gross primary production, we show that the mean sink, trend, and interannual variability in CO_2 uptake by terrestrial ecosystems are dominated by distinct biogeographic regions. Whereas the mean sink is dominated by highly productive lands (mainly tropical forests), the trend and interannual variability of the sink are dominated by semi-arid ecosystems whose carbon balance is strongly associated with circulation-driven variations in both precipitation and temperature.

Since the 1960s, terrestrial ecosystems have acted as a substantial sink for atmospheric CO_2, sequestering about one-quarter of anthropogenic emissions in an average year (*1*). This ecosystem service, which helps mitigate climate change by reducing the rate of increase of atmospheric greenhouse gases, is due to an imbalance between the uptake of CO_2 through gross primary production (GPP, the aggregate photosynthesis of plants) and the release of carbon to the atmosphere by ecosystem respiration (R_{eco}) and other losses, including wildfires (C_{fire}). The net carbon flux (net biome production, NBP = GPP − R_{eco} − C_{fire}) results from the small

imbalance between the much larger uptake and release fluxes. Consequently, small fractional variations in either of these fluxes can cause substantial absolute variations in net carbon exchange with the atmosphere. These variations account almost entirely for year-to-year variations around the overall trend in atmospheric concentrations of CO_2 (*2*, *3*).

Modeling studies suggest a large uncertainty of the future magnitude and sign of the carbon sink provided by terrestrial ecosystems (*4*–*8*). Robust projections are crucial to assessments of future atmospheric CO_2 burdens and associated climate change, and are therefore central to the effectiveness of future mitigation policies. Reducing the uncertainty of these projections requires better knowledge of the regions and processes governing the present sink and its variations. Inventories suggest that since the beginning of industrialization, the majority of carbon sequestered by the terrestrial biosphere has accumulated in forest ecosystems of the tropics and temperate zones (*9*). However, the relative contributions of ecosystems of different, climatically distinct, regions to variations in the land sink on interannual to multidecadal time scales are not well characterized. Here, we investigated relative regional contributions to the mean sink, to its trend over recent decades, and to the interannual variability (IAV) around the trend.

We used LPJ-GUESS (*10*–*12*), a biogeochemical dynamic global vegetation model, to simulate the geographic pattern and time course of NBP. LPJ-GUESS explicitly accounts for the dependency of plant production and downstream ecosystem processes on the demography (size structure) and composition of simulated vegetation. We forced the model with historical climate (*13*) and CO_2 concentrations, accounting for emissions from land use change and carbon uptake due to regrowth after agricultural abandonment (*14*). We compared the results to an ensemble of nine ecosystem and land surface model simulations from the TRENDY model intercomparison project (*12*, *15*) (hereinafter TRENDY models; table S1). The TRENDY ensemble is similarly based on historical climate and CO_2 but uses a static 1860 land use mask.

Global NBP, as simulated by LPJ-GUESS, shows strong agreement (r^2 = 0.62) with the Global Carbon Project (GCP) estimate of the net land CO_2 flux—an independent, bookkeeping-based estimate derived as the residual of emissions, atmospheric growth, and ocean uptake of CO_2 (*1*) (Fig. 1A). TRENDY models do not account for land use change. Relative to the GCP land flux estimate, they consequently predict a higher average NBP but similar interannual variation. Moreover, the offset between the TRENDY model ensemble mean and the GCP land flux estimate is comparable to the GCP estimate of mean land use change emission flux for the period 1982–2011 (fLUC).

We divided the global land area into six land cover classes, following the MODIS MCD12C1 land cover classification (*12*, *16*): tropical forests (Fig. 1B), extratropical forest, grasslands and croplands (here combined), semi-arid ecosystems (Fig. 1C), tundra and arctic shrub lands, and sparsely vegetated lands (areas classified as barren) (figs. S1 and S2).

When the global terrestrial CO_2 sink (average NBP) and its trend (1982–2011) are partitioned among land cover classes, we find that tropical forests account for the largest fraction (26%, 0.33 Pg C $year^{-1}$) of the average sink over this period (1.23 Pg C $year^{-1}$) (Fig. 1D). In contrast, we find that semi-arid ecosystems dominate the positive global CO_2 sink trend (57%, 0.04 Pg C $year^{-2}$; global, 0.07 Pg C $year^{-2}$) (Fig. 1E). The TRENDY model ensemble shows a consistent pattern, with

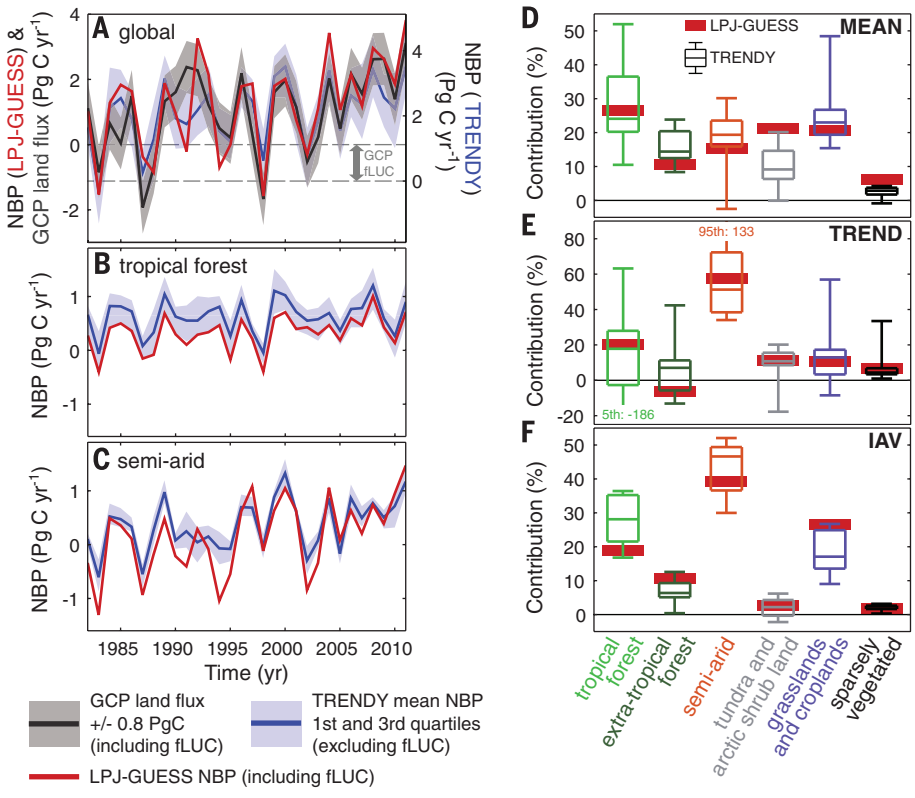


Fig. 1. Global and regional NBP mean, trend, and variations (1982–2011). (**A**) Global NBP from LPJ-GUESS (red line) and GCP land flux time series (black line) with ±0.8 Pg C uncertainty range (shaded gray area). TRENDY models mean (blue line) and first and third quartiles (shaded blue area) are plotted on a separate axis with a time-invariant offset corresponding to the time period average GCP fLUC estimate (1.2 P Pg C $year^{-1}$). (**B**) Tropical forest NBP. LPJ-GUESS (red line) includes emissions from land use change. TRENDY models average (blue line) and first and third quartiles of the ensemble (shaded blue area) do not include emissions from land use change. (**C**) NBP of semi-arid ecosystems from LPJ-GUESS (including land use change emissions) and TRENDY models (excluding land use change emissions); colors and shading as in (B). (**D**) Contribution of land cover classes to global mean NBP (1982–2011) (mean NBP of land cover class as a proportion of mean global NBP). Horizontal lines in box plots show, from top to bottom, 95th, 75th, 50th, 25th, and 5th percentiles. (**E**) Contribution of land cover classes to global NBP trend (land cover class NBP trend as a proportion of global NBP trend). (**F**) Contribution of land cover classes to global NBP IAV (Eq. 1).

[1]Department of Physical Geography and Ecosystem Science, Lund University, 223 62 Lund, Sweden. [2]Department of Earth System Science, School of Earth, Energy and Environmental Sciences, Stanford University, Stanford, CA 94305, USA. [3]Climate Change Institute, Australian National University, Canberra, ACT 0200, Australia. [4]Department of Geosciences and Natural Resource Management, University of Copenhagen, 1350 Copenhagen, Denmark. [5]Institute for Meteorology and Climate Research–Atmospheric Environmental Research, Karlsruhe Institute for Technology, 82476 Garmisch-Partenkirchen, Germany. [6]Biogeochemical Intergration Department, Max Planck Institute for Biogeochemistry, 07745 Jena, Germany. [7]Global Carbon Project, CSIRO Oceans and Atmospheric Flagship, Canberra, ACT, Australia. [8]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, UK. [9]Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [10]Institute of Applied Energy, 105-0003 Tokyo, Japan. [11]Institute on Ecosystems and the Department of Ecology, Montana State University, Bozeman, MT 59717, USA. [12]College of Life and Environmental Sciences, University of Exeter, Exeter EX4 4RJ, UK. [13]Department of Life Sciences, Imperial College, Ascot SL5 7PY, UK. [14]Climate and Environmental Physics, Physics Institute and Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland. [15]Laboratoire des sciences du climat et de l'environnement, CEA Saclay, F-91191 Gif-sur-Yvette Cedex, France. [16]CSIRO Ocean and Atmosphere Flagship, PMB 1, Aspendale, Victoria 3195, Australia. [17]Met Office Hadley Centre, Fitzroy Road, Exeter EX1 3PB, UK. [18]Department of Atmospheric and Oceanic Science and Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD 20742, USA.
***Corresponding author. E-mail: anders.ahlstrom@nateko.lu.se**
†Deceased.

tropical forests dominating the mean sink (median 24%) and semi-arid ecosystems dominating the trend (median 51%). The predominance of semi-arid ecosystems in explaining the global land sink trend is consistent with widespread observations of woody encroachment over semi-arid areas (*17*) and increased vegetation greenness inferred from satellite remote sensing over recent decades (*17*–*19*). Likewise, a recent study attributes the majority of the record land sink anomaly of 2011 to the response of semi-arid ecosystems in the Southern Hemisphere, Australia in particular, to an anomalous wet period; the study further postulates a recent increase in the sensitivity of carbon uptake to precipitation for this region, which is attributed to vegetation expansion (*20*).

We further partitioned IAV in global NBP among land cover classes according to the contribution of individual regions (grid cells or land cover classes) to global NBP IAV (*12*). To this end, we adopted an index that scores individual geographic locations according to the consistency, over time, with which the local NBP flux resembles the sign and magnitude of global NBP (fig. S4):

$$f_j = \frac{\sum_t \frac{x_{jt}|X_t|}{X_t}}{\sum_t |X_t|} \quad (1)$$

where x_{jt} is the flux anomaly (departure from a long-term trend) for region j at time t (in years), and X_t is the global flux anomaly, so that $X_t = \sum_j x_{jt}$. By this definition f_j is the average relative anomaly x_{jt}/X_t for region j, weighted with the absolute global anomaly $|X_t|$. Regions receiving higher and positive average scores are inferred to have a larger contribution in governing global NBP IAV, as opposed to regions characterized by smaller or negative (counteracting) scores (fig. S3). The index we adopt does not characterize the variability of ecosystems of different regions, as, for example, the standard deviation would do (fig. S5); rather, it enables a comparison of their relative importance (contribution) in governing global IAV.

Semi-arid ecosystems were found to account for the largest fraction, 39%, of global NBP IAV, exceeding tropical forest (19%), extratropical forest (11%; all forest, 30%), and grasslands and croplands (27%) (Fig. 1F). The TRENDY model ensemble shows a similar partitioning, with semi-arid ecosystems accounting for 47% (median; tropical forests, 28%; extratropical forest, 6%; all forest, 35%). The overall contributions per land cover class are the sum of both positive and negative contributions that result from differences

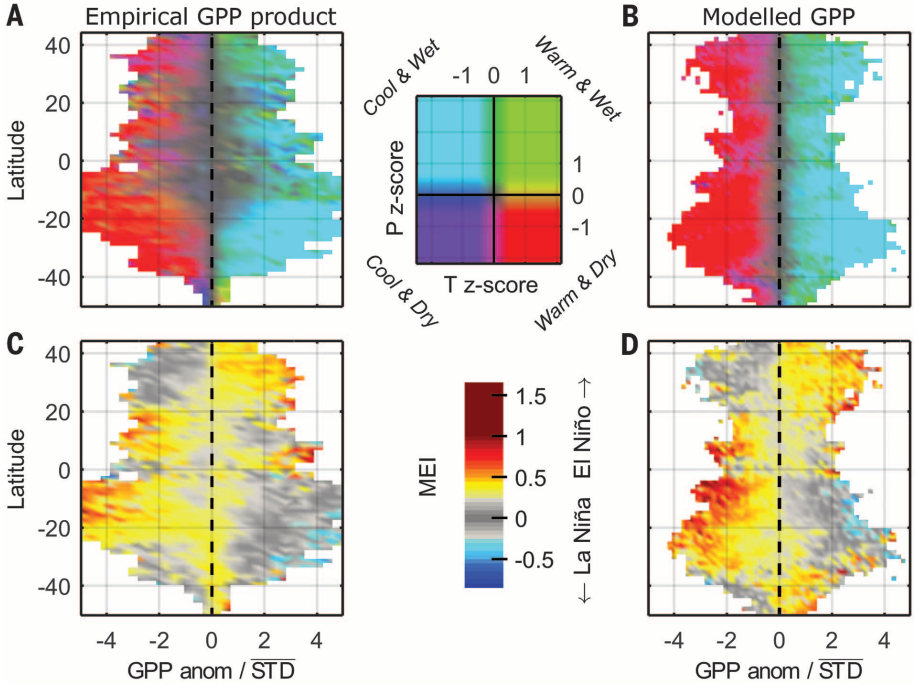


Fig. 2. Climatic covariates of semi-arid ecosystem GPP variations. (**A**) Distribution by latitude of the empirical GPP product anomalies normalized by average standard deviation of GPP in semi-arid lands. The distribution is colored according to average local climatic covariates per latitude zone and distribution bin. (**B**) LPJ-GUESS GPP distribution calculated and colored as in (A). (**C**) Covariation of the multivariate ENSO index [MEI (*31*, *32*)] anomalies with the empirical GPP product. (**D**) Covariation of MEI and modeled GPP anomalies per latitudinal zone. Note that the figure shows the covariates of latitudinal average local GPP anomalies, and not the average covariates based on GPP IAV contribution to NBP IAV.

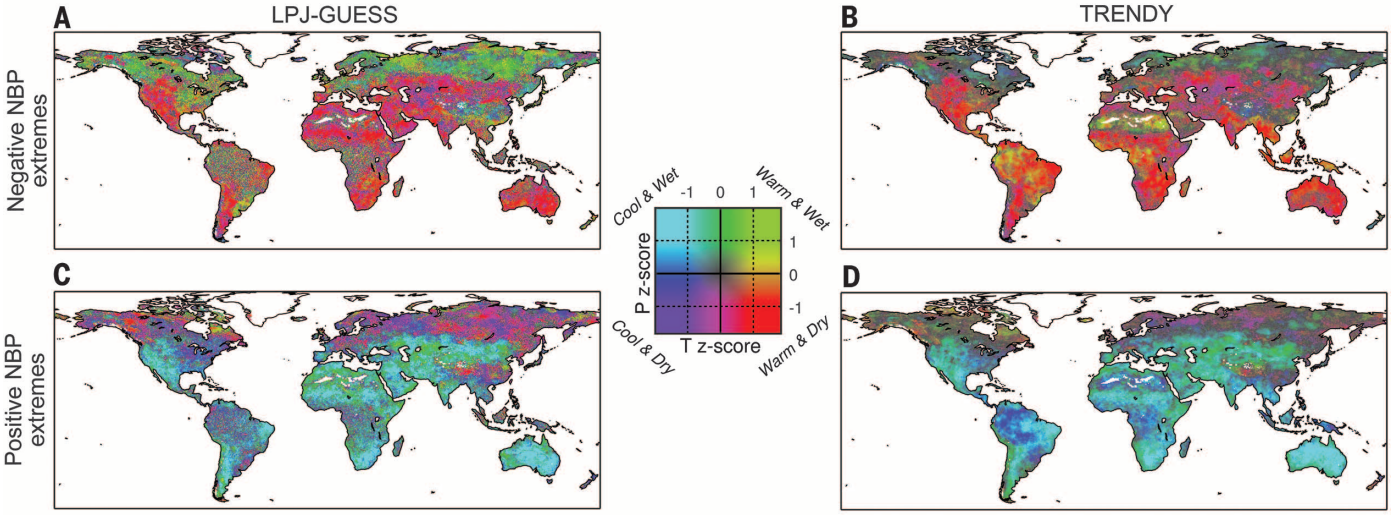


Fig. 3. Climatic covariates of NBP extremes. (**A**) Climatic covariates of LPJ-GUESS negative NBP extremes (1st to 10th percentiles). (**B**) Mean climatic covariates of TRENDY models' negative NBP extremes (1st to 10th percentiles). (**C**) Covariates of LPJ-GUESS positive NBP extremes (90th to 99th percentiles). (**D**) Mean climatic covariates of TRENDY models' positive NBP extremes (90th to 99th percentiles).

in phase between IAV of individual grid cells compared with global IAV (fig. S4). The extent to which negative contributions reduce the overall land cover class contributions is minor for all regions except grasslands and croplands (fig. S6) (LPJ-GUESS, –13%; TRENDY median, –13%) because the latter are distributed widely across climate zones, and because both climate variations and the sensitivity of NBP to climate variations differ among regions.

To partition the global NBP IAV among component fluxes (GPP, R_{eco}, C_{fire}) and among land cover classes, we applied Eq. 1. We found that global NBP IAV is most strongly associated with variation in GPP; interannual GPP anomalies contribute 56% of the global NBP IAV in LPJ-GUESS and a median of 90% in the TRENDY model ensemble. Comparing different land cover classes, the GPP anomalies of semi-arid ecosystems alone contribute 39% in LPJ-GUESS and a median of 65% in the TRENDY model ensemble to global NBP IAV (fig. S7). Semi-arid vegetation productivity thus emerges clearly as the single most important factor governing global NBP IAV.

We used two complementary methods to attribute the variability in GPP—as the inferred primary driver of global NBP IAV—to its environmental drivers. First, we analyzed simulation results from LPJ-GUESS, linking output GPP anomalies to variability in the climatic input data. Second, we used a time-resolved gridded global GPP product derived from upscaled flux tower measurements (*12*, *21*) (hereafter, empirical GPP product). This product uses an empirical upscaling of flux measurements and is thus entirely independent of the modeled GPP in our study.

The three main climatic drivers—temperature (T), precipitation (P), and shortwave radiation (S)—are interdependent and correlated. To account for the combined effects of these drivers, we adopted an analysis of GPP variations from an "impact perspective" (*22*–*24*): We first identified GPP anomalies and then extracted their climatic covariates. The primary challenge of such an analysis on an annual scale is to target climate indices that adequately characterize the "period of climatic influence" (e.g., growing season average, annual averages, minima or maxima of a given climatic forcing). To overcome this challenge, we used semiannual time series of climate drivers constructed via an optimization procedure that weights monthly anomalies of a given climate variable (T, P, or S), accounting for time lags of up to 24 months while making no additional prior assumptions as to the period of influence (*12*). For each GPP event, we extracted climatic covariates as *z* scores of the semiannual climatic drivers.

We evaluated the climatic covariates of GPP anomalies for semi-arid ecosystems from the empirical GPP product and modeled by LPJ-GUESS, focusing on T and P, and found similar responses of GPP to climate with both approaches across all latitude bands (Fig. 2, A and B). Negative GPP anomalies in semi-arid ecosystems are mainly driven by warm and dry (low rainfall) climatic events in most latitudes, suggestive of drought. By contrast, positive GPP anomalies are dominated by cool and wet conditions. Averaging the distributions over latitudes (Fig. 2, A and B) and extracting the climatic covariates per percentile of the GPP distributions shows that GPP varies with climatic conditions on a straight line in T-P space (fig. S8), with a stronger covariation with P than with T. This implies that the full GPP distributions are driven by similar climatic patterns—that is, anomalies that differ in size and sign covary with corresponding differences in size and sign in the drivers. GPP extremes (the tails of the distribution of GPP among years) covary with El Niño–Southern Oscillation (ENSO) across all latitudes (Fig. 2, C and D). In both the model and the empirical GPP product, GPP anomalies are more strongly associated with the positive phase of ENSO (El Niño) than with the negative phase (La Niña); the sign of the relationship varies with latitude. Positive ENSO tends to coincide with negative GPP anomalies in the tropics (30°S to 20°N) and with positive GPP anomalies north of 20°N.

The agreement between climatic covariates of the data-based empirical GPP product and modeled GPP alongside the comparatively robust pattern of the covariation with climate suggests that GPP IAV for semi-arid ecosystems is mediated by climate. Because ENSO covaries with a considerable portion of the GPP distribution, we infer that ENSO is the dominating mode of global circulation variations driving GPP IAV over semi-arid ecosystems. Recent modeling studies have found that extreme El Niño events could become more common under climate change (*25*), which, together with an increased atmospheric demand for water associated with global warming, might exacerbate the impact of El Niño events over semi-arid ecosystems and further increase the role of semi-arid regions in driving global NBP IAV (*26*–*28*).

We repeated the calculation of climatic covariates to simulated NBP for LPJ-GUESS and each of the TRENDY models. The resulting maps of covariates in T-P space are shown as average covariates of negative NBP extremes (Fig. 3, A and B) and positive NBP extremes (Fig. 3, C and D). In general, semi-arid ecosystems stand out as regions in which strong CO_2 uptake events are consistently associated with cool and moist conditions, and strong CO_2 release events with warm and dry conditions. In tropical forests, NBP covaries with both T and P as in semi-arid regions, but also with T alone. In high latitudes, wet or warm and wet conditions lead to negative NBP extremes, whereas dry or warm and dry conditions tend to lead to positive extremes, although the spatial heterogeneity of the covariates is large in this region (Fig. 3).

Our approach offers detailed spatial and temporal disaggregation of drivers and responses, which is important when analyzing drivers or covariates of global NBP IAV because of the high temporal and spatial variability in P (figs. S9 to S11). Using four upscaling levels with increasing spatial and temporal disaggregation [ranging from land surface mean P and T to semiannual P and T, averaged according to the spatial origin of each year's global NBP anomaly (eqs. S5 and S6)], we found that P and NBP IAV become more correlated at higher levels of disaggregation. At the highest disaggregation level, P is almost as strongly correlated with NBP IAV as T, suggesting a strong influence of soil moisture variations on global NBP IAV (*28*). This strong increase in P correlations with disaggregation resolves an apparent conflict between our findings and those of studies using regionally averaged drivers that emphasize the role of T in governing IAV in atmospheric CO_2 (*28*–*30*). For semi-arid ecosystems, T correlations with NBP IAV are slightly stronger than P correlations with NBP IAV (Fig. 4B), partly because of an asymmetric distribution of P and/or an asymmetric response of NBP to P IAV (fig. S12). The correlation of tropical forest P with

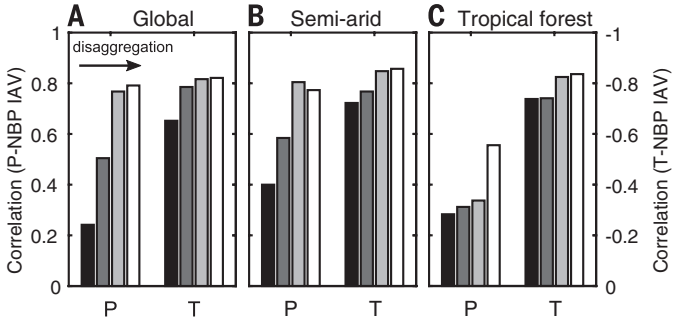


Fig. 4. Correlations between climatic drivers IAV (P and T) and global NBP IAV (mean of all 10 models). (**A**) Global P and T correlations to global NBP IAV. From black to white and left to right, bars represent annual P and T IAV correlations to global NBP IAV with increasing spatial and temporal disaggregation of P and T while averaging to global time series. Black bars represent averaged global land surface P and T weighted by grid cell area; dark gray bars represent P and T weighted by 30-year average contribution to global NBP IAV (Eq. 1 and fig. S4); light gray bars represent averaged P and T weighted by each year's contributions, thus accounting for the difference in the spatial distribution of contributions between years (eqs. S5 and S6); white bars represent semiannual climate drivers averaged to global time series using the annual spatial contributions (as for light gray bars), thereby accounting for the "period of climatic influence" and time lags of up to 24 months. (**B**) Correlations between P and T IAV and global NBP IAV for semi-arid ecosystems. Weights, where applicable, are based on contributions to global NBP IAV as in (A) but with P and T averaged over semi-arid ecosystems only. (**C**) Correlations between P and T IAV and global NBP IAV for tropical forest. Weights, where applicable, are based on contributions to global NBP IAV as in (A) but with P and T averaged over tropical forest only.

NBP IAV increases when we use the semiannual drivers, which suggests the importance of accounting for time lags and the "period of climatic influence" of P variations (*12*), but P correlations with NBP IAV are still weaker than T correlations with NBP IAV (Fig. 4C).

Our analysis provides evidence that semi-arid ecosystems, largely occupying low latitudes, have dominated the IAV and trend of the global land carbon sink over recent decades. Semi-arid regions have been the subject of relatively few targeted studies that place their importance in a global context. Our findings indicate that semi-arid regions and their ecosystems merit increased attention as a key to understanding and predicting interannual to decadal variations in the global carbon cycle.

REFERENCES AND NOTES

1. C. Le Quéré *et al.*, *Earth Syst. Sci. Data* **6**, 235–263 (2014).
2. C. D. Keeling, T. P. Whorf, M. Wahlen, J. van der Plichtt, *Nature* **375**, 666–670 (1995).
3. C. Le Quéré *et al.*, *Nat. Geosci.* **2**, 831–836 (2009).
4. A. Ahlström, G. Schurgers, A. Arneth, B. Smith, *Environ. Res. Lett.* **7**, 044008 (2012).
5. P. Friedlingstein *et al.*, *J. Clim.* **19**, 3337–3353 (2006).
6. A. D. McGuire *et al.*, *Glob. Biogeochem. Cycles* **15**, 183–206 (2001).
7. S. Schaphoff *et al.*, *Clim. Change* **74**, 97–122 (2006).
8. S. Sitch *et al.*, *Glob. Change Biol.* **14**, 2015–2039 (2008).
9. Y. Pan *et al.*, *Science* **333**, 988–993 (2011).
10. A. Ahlström, P. A. Miller, B. Smith, *Geophys. Res. Lett.* **39**, L15403 (2012).
11. B. Smith, I. C. Prentice, M. T. Sykes, *Glob. Ecol. Biogeogr.* **10**, 621–637 (2001).
12. See supplementary materials on *Science* Online.
13. I. Harris, P. D. Jones, T. J. Osborn, D. H. Lister, *Int. J. Climatol.* **34**, 623–642 (2014).
14. G. Hurtt *et al.*, *Clim. Change* **109**, 117–161 (2011).
15. S. Sitch *et al.*, *Biogeosciences* **12**, 653–679 (2015).
16. M. A. Friedl *et al.*, *Remote Sens. Environ.* **114**, 168–182 (2010).
17. N. Andela, Y. Y. Liu, A. I. J. M. van Dijk, R. A. M. de Jeu, T. R. McVicar, *Biogeosciences* **10**, 6657–6676 (2013).
18. R. J. Donohue, T. R. McVicar, M. L. Roderick, *Glob. Change Biol.* **15**, 1025–1039 (2009).
19. R. Fensholt *et al.*, *Remote Sens. Environ.* **121**, 144–158 (2012).
20. B. Poulter *et al.*, *Nature* **509**, 600–603 (2014).
21. M. Jung *et al.*, *J. Geophys. Res.* **16**, G00J07 (2011).
22. J. Zscheischler *et al.*, *Environ. Res. Lett.* **9**, 035001 (2014).
23. M. Reichstein *et al.*, *Nature* **500**, 287–295 (2013).
24. M. D. Smith, *J. Ecol.* **99**, 656–663 (2011).
25. W. Cai *et al.*, *Nat. Clim. Change* **4**, 111–116 (2014).
26. K. E. Trenberth *et al.*, *Nat. Clim. Change* **4**, 17–22 (2014).
27. A. Dai, *Nat. Clim. Change* **3**, 52–58 (2013).
28. X. Wang *et al.*, *Nature* **506**, 212–215 (2014).
29. W. Wang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 13061–13066 (2013).
30. P. M. Cox *et al.*, *Nature* **494**, 341–344 (2013).
31. K. Wolter, M. S. Timlin, in *Proceedings of the 17th Climate Diagnostics Workshop* (University of Oklahoma, Norman, OK, 1993), pp. 52–57; www.esrl.noaa.gov/psd/enso/mei/WT1.pdf.
32. K. Wolter, M. S. Timlin, *Weather* **53**, 315–324 (1998).

ACKNOWLEDGMENTS

This paper is dedicated to the memory of Michael Robin Raupach (1950–2015), whose scientific integrity and novel contributions leave a long-lasting legacy in the field of carbon cycle sciences. The MODIS MOD12C1 land cover product was obtained through the online Data Pool at the NASA Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota (https://lpdaac.usgs.gov/data_access). Supported by the Royal Physiographic Society in Lund (Birgit and Hellmuth Hertz Foundation), Swedish Research Council grant 637-2014-6895, and the Mistra-SWECIA program (A. Ahlström); EC FP7 grant LUC4C (603542) (A. Arneth); OCE Distinguished Visiting Scientist to the CSIRO Ocean and Atmosphere Flagship, Canberra (B.S.); EC FP7 grant EMBRACE (282672) (A. Arneth, M.R., and B.D.S.); the Australian Climate Change Science Program (J.G.C.); NSF grant AGS 12-43071, U.S. Department of Energy grant DE-SC0006706, and NASA LCLUC program grant NNX14AD94G (A.K.J.); the Environmental Research and Technology Development Fund (S-10) of the Ministry of Environment of Japan (E.K.); CSIRO strategic research funds (Y.P.W.); and NOAA grants NA10OAR4310248 and NA09NES4400006 and NSF grant AGS-1129088 (N.Z.). This study is a contribution to the Lund Centre for Studies of Carbon Cycle and Climate Interactions (LUCCI) and the strategic research areas MERGE and BECC.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/895/suppl/DC1
Materials and Methods
Figs. S1 to S12
References (*33*–*56*)

26 October 2014; accepted 24 April 2015
10.1126/science.aaa1668

GLACIER MASS LOSS

Dynamic thinning of glaciers on the Southern Antarctic Peninsula

B. Wouters,[1]* A. Martin-Español,[1] V. Helm,[2] T. Flament,[3] J. M. van Wessem,[4] S. R. M. Ligtenberg,[4] M. R. van den Broeke,[4] J. L. Bamber[1]

Growing evidence has demonstrated the importance of ice shelf buttressing on the inland grounded ice, especially if it is resting on bedrock below sea level. Much of the Southern Antarctic Peninsula satisfies this condition and also possesses a bed slope that deepens inland. Such ice sheet geometry is potentially unstable. We use satellite altimetry and gravity observations to show that a major portion of the region has, since 2009, destabilized. Ice mass loss of the marine-terminating glaciers has rapidly accelerated from close to balance in the 2000s to a sustained rate of –56 ± 8 gigatons per year, constituting a major fraction of Antarctica's contribution to rising sea level. The widespread, simultaneous nature of the acceleration, in the absence of a persistent atmospheric forcing, points to an oceanic driving mechanism.

Ice shelves have been identified as sensitive indicators of climate change (*1*). Their retreat along the coast of the Northern Antarctic Peninsula has been noted over recent decades (*2*) and associated with a sudden and prolonged increase in discharge of the inland grounded ice (*3*–*5*), especially for those glaciers overlying deep troughs (*6*). The potential future contribution to sea-level rise of these glaciers relatively modest because their catchments are small compared with those further south (*7*). The Southern Antarctic Peninsula (SAP), including Palmer Land and the Bellinghausen Coast, rests on bedrock below sea level with a retrograde slope (deeper inland) (*8*), which is believed to be an inherently unstable configuration (*9*), permitting rapid grounding line retreat and mass loss to the ocean. Recent modeling results suggest that this marine ice sheet instability may have already been initiated for part of West Antarctica (*10*, *11*).

The SAP is home to a number of fast flowing, marine terminating glaciers, many of which are still unnamed. Laser [ICESat, 2003–2009 (*12*)] and radar [Envisat, 2003– 2010 (*13*)] altimetry identified moderate surface-lowering concentrated within a narrow strip along the coast, in particular near the grounding line of the Ferrigno Ice Stream (*14*), contrasted by widespread thickening further inland. Observations from the Gravity Recovery and Climate Experiment (GRACE) mission show that these opposing signals compensated each other, resulting in a near-zero mass balance for 2002–2010 (*15*).

The Cryosat-2 satellite, launched in April 2010, provides elevation measurements of land and sea ice at a high spatial resolution up to a latitude of 88°. In contrast to conventional altimetry missions such as Envisat, Cryosat-2's dual antenna and Doppler processing results in improved resolution and geolocation of the elevation measurement (*16*). Because of the long satellite repeat period of 369 days, it has a dense track spacing in our region of interest, which is a major advantage compared with the roughly 10-times-coarser ICESat track spacing. Two recent studies using Cryosat-2 data observed thinning along the coast of the Bellinghausen Sea (*17*, *18*). Such elevation changes may result from either a decrease in surface mass balance (SMB) (accumulation minus ablation), compaction of the firn column, or an increase in the ice flow speed (also termed dynamic thinning). Both studies attributed the surface-lowering to interannual changes in SMB, based on the strong accumulation variability observed in the Gomez ice core (70.36°W, 73.59°S) (*18*, *19*). Here, we take SMB and firn compaction into account and show that the

[1]Bristol Glaciology Centre, University of Bristol, Bristol, UK. [2]Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, Bremerhaven, Germany. [3]Laboratoire d'Etudes en Géophysique et Océanographie Spatiales, Toulouse, France. [4]Institute for Marine and Atmospheric Research, Utrecht University, Netherlands.
*Corresponding author. E-mail: bert.wouters@bristol.ac.uk

signal is due to pronounced glacier dynamic ice loss instead.

We used a pseudo-repeat track method to derive elevation changes from the Cryosat-2 measurements (July 2010 to April 2014), which makes optimal use of the available observations (*20*), allows us to observe small-scale features such as the changes of the narrow Nikitin Glacier (Fig. 1B), and compares well with trends derived from high-accuracy, high-resolution airborne laser altimetry campaigns (fig. S1). Strong negative elevation trends are found along a roughly 750-km western coastal transect between the catchments of the Jensen Nunataks and the Wesnet and Williams Ice Stream (regions denoted in Fig. 1A), which are mainly localized in areas of fast glacier flow (fig. S2 for comparison). The average observed elevation rate in our area of interest [basins 23 and 24 as defined in (*21*) and used in the ice sheet mass balance inter-comparison exercise (IMBIE) study (*22*)] equals −0.42 m/year, with catchment averages as negative as −1.15 m/year for the Fox Ice Stream (table S1). Locally, near the grouding line, thinning rates in this catchment reach values down to −4 m/year. Thinning is also pronounced in the English Coast region, with rates close to the grounding line of −2 m/year or more occurring for several of the glaciers.

Integrated over the entire region (174,101 km^2), volume losses total −72 ± 10 km^3/year (July 2010 to April 2014) (table S2). Part of this signal is due to changes in the air content of the firn column, which is caused by variability in temperature and accumulation (and thus no associated change in mass) alongside variations in SMB. To correct for these two effects, we used a firn densification model (*23*) driven by a regional climate model (*24*). The variations in SMB and firn densification rate are more widespread—and not tied to fast flowing narrow glacier areas—and are an order of magnitude too small to explain the observed elevation changes (fig. S3). After correcting the altimetry rates with the firn densification model, the link between the surface-lowering and fast flowing ice becomes even more evident, with the majority of negative trends occurring between the coastline and the 50-m/year velocity contour (fig. S4B).

The firn model prescribes a volume change of −15 ± 3 km^3/year to surface processes. Attributing the remainder to ice dynamics (at a density of ρ_{ice} = 917 kg/m^3), and adding back the modeled SMB mass anomalies (fig. S5), yields a total mass loss of −59 ± 10 gigatons (Gt)/year. Repeating this approach for elevation rates obtained from combined ICESat/Envisat observations during 2003–2009 (*20*) shows a contrasting picture, with a near-balance during 2003–2009 (3 ± 22 Gt/year), with slightly more positive values at the beginning of the observations (2003–2005, 15 ± 26 Gt/year) compared with the end (2007–2009, −10 ± 15 Gt/year). This suggests a remarkable rate of acceleration in dynamic mass loss since about 2009 that must have been near-simultaneous across multiple basins and glaciers.

The GRACE satellites measure changes in mass distribution at, and beneath, the surface (*25*). Because these gravimetric observations are insensitive to the underlying processes causing the mass redistribution (in this case, either ice dynamics or SMB, or a combination), they offer an independent method with which to validate the altimetric observations. The GRACE data shows an increase in mass loss in our region of interest (fig. S6) and are consistent with the ICESat/Envisat and Cryosat-2 observations within uncertainties at all time intervals (table S3). The region was in approximate balance for 2003–2009 (−11 ± 5 Gt/year) (Fig. 2), with first signs of mass loss appearing around 2008, but these are at least partially caused by a temporal reduction in SMB. Rapid dynamic ice loss started in 2009 and has continued unabated since (−52 ± 14 Gt/year for July 2010 to April 2014). Although the post-2009 time series is still modulated by SMB variability (for example, the short-lived down- and upward event in 2010) (Fig. 2), the current mass loss lies clearly outside the range of variability observed in the modeled cumulative SMB for 1979 to present (10 Gt). GRACE trends are sensitive to mass redistribution related to glacial isostatic adjustment, but this signal is negligible in the region (2 ± 1 Gt/year) and because it is constant over these time scales, the sudden increase in mass loss cannot be explained

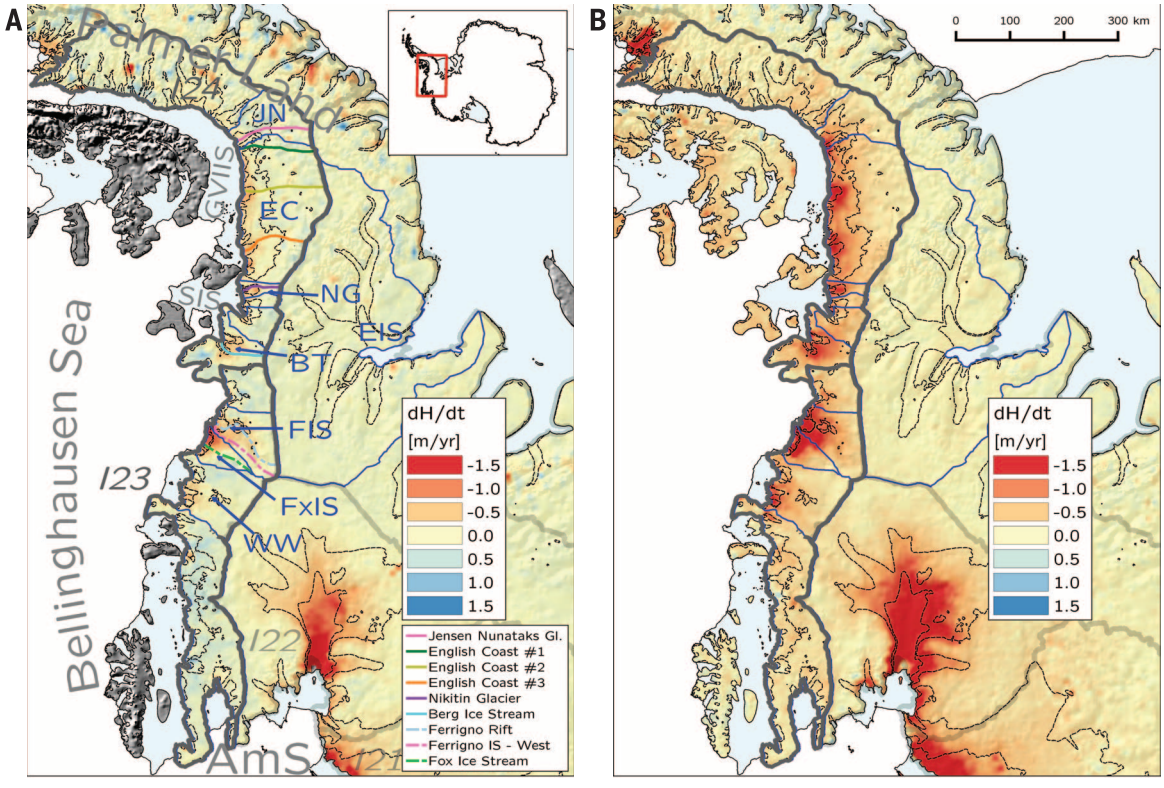


Fig. 1. Elevation rates in the Bellinghausen Sea Sector. (**A**) Envisat/ICESat (2003–2009). (**B**) Cryosat-2 (2010–2014). No correction for elevation changes due to surface processes was applied (results with this correction are provided in fig. S4). Where available, the 50- and 250-m/year velocity contours are plotted (*36*). (Inset) The location of our area of interest. The elevation rates profiles of Fig. 3 are indicated by colored lines. Glaciers basins are outlined in blue (*37*); JN, Jensen Nunataks; EC, English Coast; NG, Nitikin Glacier; BT, Berg & Thompson Ice Stream; FIS, Ferrigno Ice Stream; FxIS, Fox Ice Stream; WW, Wesnet & Williams Ice Stream; EIS, Evans Ice Stream (names of other basins are available in fig. S4). IMBIE basins are shown in gray (I23 and I24) and pale gray. Ice shelves are plotted in light blue; grounding lines are based on (*30*). GVIIS, George VI Ice Shelf; SIS, Stange Ice Shelf; AmS, Amundsen Sea.

by this source. Combining the Cryosat-2– and GRACE-derived rates yields an error-weighted mean mass loss of 56 ± 8 Gt/year for July 2010 to April 2014.

To further investigate the temporal and spatial evolution of the dynamic thinning, we sampled surface elevation rates along a number of profiles of glaciers displaying pronounced surface-lowering (locations are shown in Fig. 1A and fig. S2). As reported in earlier studies (*12*–*14*), Ferrigno Ice Stream showed thinning rates of up to 1 m/year, along the deep, subglacial rift system extending inland (*14*) during the ICESat and Envisat observation periods. No significant increase in thinning took place near the grounding line between 2003–2005 and 2007–2009, but elevation rates further upstream were slightly more negative during the latter period. In recent years, thinning near the grounding line has more than doubled and propagated ~100 km inland, which is characteristic of a dynamic thinning signal (*26*). Even larger changes are observed along the western tributary of the ice stream (Fig. 3) and the Fox Ice Stream, where locally, surface-lowering of roughly –4 m/year is now occurring at the glacier fronts, and ice drawdown stretches 75 to 100 km inland.

Further to the east, the unnamed glacier in the Jensen Nunataks region and unnamed glacier #1 in the English Coast basin were in near-balance up to 2009, whereas English Coast unnamed glacier #2 showed thinning (~1 m/year) at its front. During 2010–2014, all three glaciers showed negative elevations rates exceeding –2 m/year at their grounding lines, which become gradually less pronounced further upstream. At all nine glacier profiles surveyed, elevation rates were consistently more negative during the latter period.

In terms of larger-scale spatial variability, glacier-thinning is restricted to the western side of the southwest Peninsula. For instance, the Berg Ice Stream shows thinning up to the Peninsula's divide (–0.5 ± 0.1 m/year) (Fig. 3), with barely detectable trends on the ice on the eastern side of the divide feeding into the the Evans Ice Stream. The basin of this neighboring ice stream (118 300km^2) (Fig. 1) has been in near balance during the entire study period, with a total mass change of only 8 ± 20, 3 ± 12, and –3 ± 13 Gt/year for 2003–2005, 2007–2009, and 2010–2014, respectively.

The widespread and simultaneous speed-up of the southwest Antarctic Peninsula marine-terminating glaciers, in the absence of persistent changes in SMB in the region, points to ocean processes as the driving mechanism. Near the continental margin of the Bellinghausen Sea, warm Circumpolar Deep Water (CDW) slopes upward toward shallower depths, facilitating episodic but persistent intrusion of CDW onto the continental shelf (*27*, *28*). These water masses have direct access to the glacier fronts of the Ferrigno and Fox Ice Streams, via the Belgica Trough and Eltanin Bay (fig. S2) (*14*). The eastern glaciers of the SAP flow into the Stange ice shelf and George VI ice shelf (GVIIS), the second largest ice shelf on the Antarctic Peninsula, and particularly vulnerable to intrusion of CDW (*2*, *29*). CDW is channeled below the GVIIS through the George VI Sound, resulting in basal melt of several meters per year (*29*–*31*), which is not fully compensated by surface mass accumulation and glacier inflow (*30*, *31*). As a result, the GVIIS has been thinning during the past few decades, with recent rates on the order of –1.5 m/year near the grounding lines of glaciers feeding the southeastern flank of the GVIIS (*32*). Simultaneously, increased rifting has been reported, rendering parts of the GVIIS structurally weak, combined with a retreat of the southern ice shelf front (*29*). Using LANDSAT imagery, we estimate a loss of about 495 km^2 in the period 2000–2013, with 265 km^2 occurring in the period 2010–2013 (fig. S7) (*20*).

The recent increase in thinning of the glaciers in our region of interest coincides with a record high in in situ temperatures measured at the bed of the Bellinghausen Sea in the 2010s, which is attributed to shoaling and warming of offshore CDW (*28*). This, combined with the observed thinning and weakening of GVIIS, shows strong similarities with the recent changes observed in the Amundsen Sea sector. There, increased subglacial melt from the intrusion of CDW into the ice shelf cavities lead to thinning of the shelves, and a sustained speed-up and thinning of the feeding glaciers (*33*). Depending on the local bathymetry and subglacial topography (*34*), glacier dynamics may be strongly coupled to the evolution of the seaward ice shelf, which provides a buttressing force on the glaciers' outflow. Both models and observations suggest that a decrease in back stress of a thinning ice shelf will lead to increased ice flux and inland retreat of the grounding line (*5*, *9*, *12*, *26*, *33*). Under the right conditions (a deep trough or submarine glacier bed and/or low basal shear stress), the glacier's dynamic response may extend far upstream (*26*), which is in agreement with our observations (Fig. 3). Although estimates of grounding zone locations in our region of interest are scarce, a grounding zone retreat has indeed been observed for some of the southern glaciers feeding into the GVIIS (*29*).

Dynamic thinning may be further promoted if the glacier is grounded below sea level on a bed with retrograde slope (*9*), as seen in the Amundsen Sea sector. Along the Bellinghausen Coast, such conditions are present at some of the glaciers showing the most pronounced thinning (fig. S2). The best documented example is the Ferrigno Rift (*14*), but the Nikitin Glacier and the unnamed glaciers of the English Coast show a similar configuration. The bedrock-deepening does not extend as far inland as observed in the Amundsen Sea Sector, but a large part of this region was inferred to be vulnerable to marine instability (*8*). Even if the forcing causing the observed thinning were to cease, dynamic thinning in the region will continue until the glaciers reach a new equilibrium

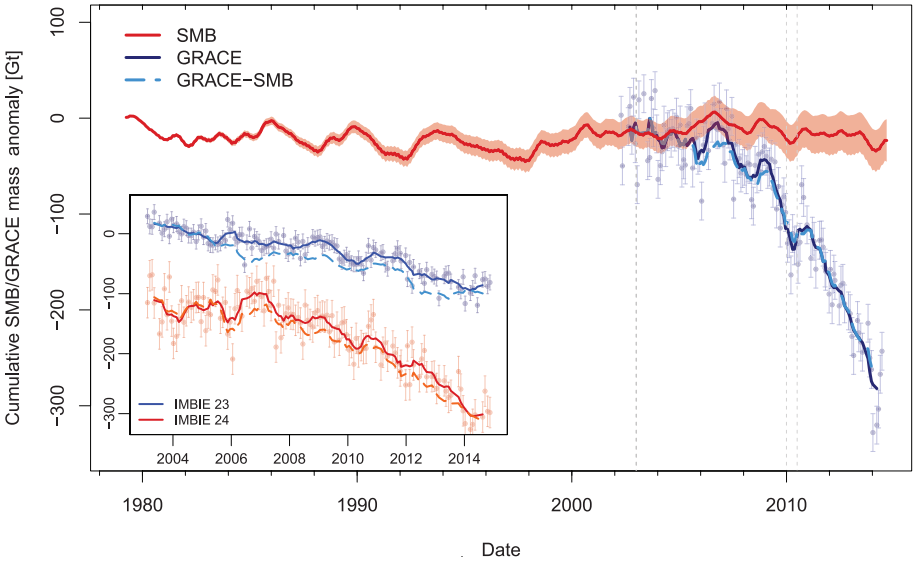


Fig. 2. Mass variations for the sum of basins 23 and 24, as observed by GRACE and modeled by RACMO2.3. Basins 23 and 24 are defined in (*21*, *22*). The faint blue dots are the monthly GRACE anomalies with 1σ error bars (*20*), and the thick blue line shows the anomalies with a 7-month running average applied so as to reduce noise. Cumulative SMB anomalies from RACMO2.3 are shown in red, with the light red area indicating the 1σ spread in an ensemble obtained by varying the baseline period (*20*). The dashed light blue line shows the estimated dynamic mass loss (GRACE minus SMB). The vertical dashed lines indicate January 2003, December 2009, and July 2010, the start and ending of the different altimetry observations. (Inset) The GRACE time series for the individual basins 23 (blue) and 24 (red), before (full lines) and after (dashed lines) applying the SMB correction.

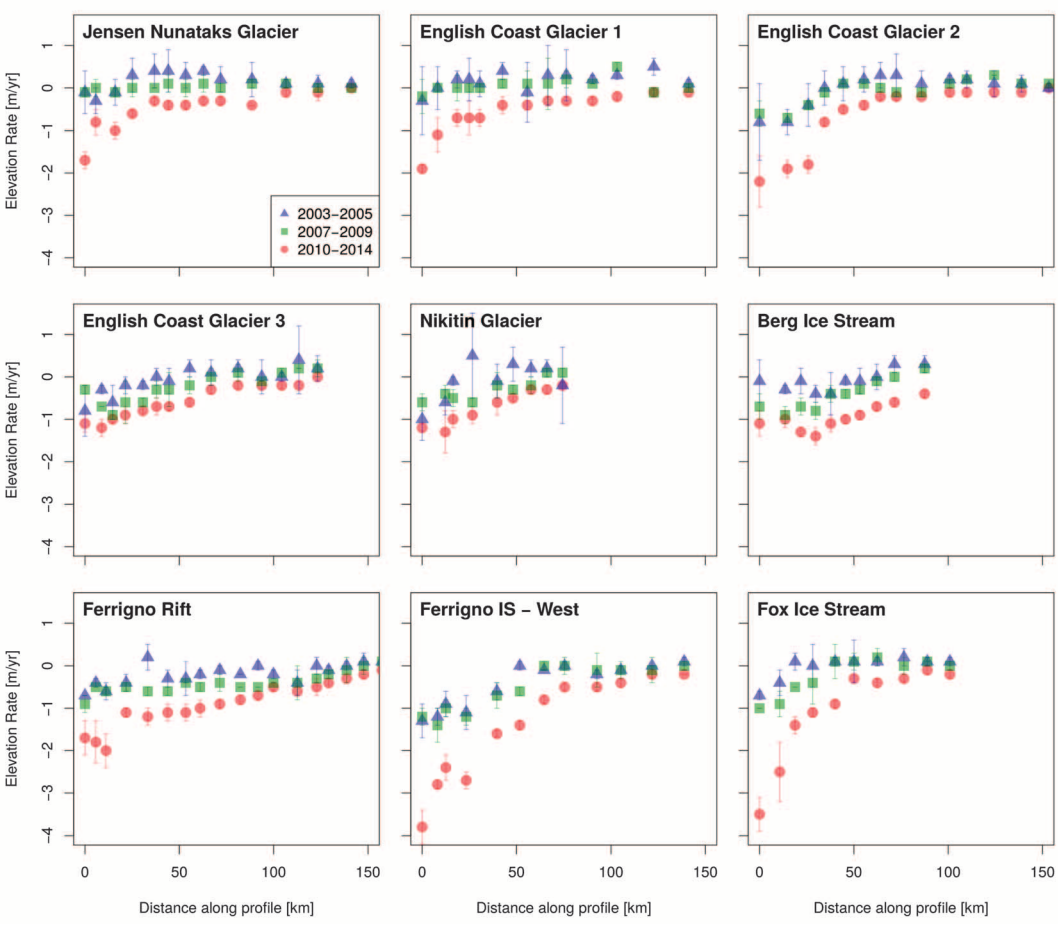


Fig. 3. Surface ice elevation rates along the profiles shown in Fig. 1, for 2003–2005, 2007–2009, and 2010–2014. 2003–2005, blue triangles; 2007–2009, green squares; 2010–2014, red circles. The altimetry observations have been corrected for surface processes, and the uncertainty bars are based on the root-square sum of the uncertainties in the altimetry data and the firn model.

state. The present losses of −56 ± 8 Gt/year are more than half of the mass loss in the Amundsen Sea Embayment [−80 to −110 Gt/year, depending on the period (*35*); IMBIE basins 21 and 22]. The Bellinghausen Coast glaciers currently add ~0.16 mm/year to global mean sea level and therefore constitute a major fraction of Antarctica's total oceanic contribution. The thinning and weakening of George VI, and other ice shelves along the western coast of the Peninsula (*32*), is most likely due to shoaling of relatively warm CDW onto the continental shelf (*12*, *28*). The intrusion of CDW will also lead to enhanced basal melting at the grounding line, resulting in steepening of the near-coast ice margin and therefore faster glacier flow. We conclude that these processes have resulted in the destabilization of the inland ice, resulting in a large and sustained mass loss to the ocean.

REFERENCES AND NOTES

1. D. G. Vaughan, C. S. M. Doake, *Nature* **379**, 328–331 (1996).
2. A. J. Cook, D. G. Vaughan, *Cryosphere* **4**, 77–98 (2010).
3. H. Rott, W. Rack, P. Skvarca, H. de Angelis, *Ann. Glaciol.* **34**, 277–282 (2002).
4. H. De Angelis, P. Skvarca, *Science* **299**, 1560–1562 (2003).
5. T. A. Scambos *et al.*, *Cryosphere* **8**, 2135–2145 (2014).
6. M. Rebesco *et al.*, *Science* **345**, 1354–1358 (2014).
7. M. Huss, D. Farinotti, *Cryosphere* **8**, 1261–1273 (2014).
8. J. L. Bamber, R. E. M. Riva, B. L. A. Vermeersen, A. M. LeBrocq, *Science* **324**, 901–903 (2009).
9. C. Schoof, *J. Geophys. Res.* **112** (F3), 3 (2007).
10. L. Favier *et al.*, *Nature Climate Change* **4**, 117–121 (2014).
11. I. Joughin, B. E. Smith, B. Medley, *Science* **344**, 735–738 (2014).
12. H. D. Pritchard *et al.*, *Nature* **484**, 502–505 (2012).
13. T. Flament, F. Rémy, *J. Glaciol.* **58**, 830–840 (2012).
14. R. G. Bingham *et al.*, *Nature* **487**, 468–471 (2012).
15. M. A. King *et al.*, *Nature* **491**, 586–589 (2012).
16. D. J. Wingham *et al.*, *Adv. Space Res.* **37**, 841–871 (2006).
17. M. McMillan *et al.*, *Geophys. Res. Lett.* **41**, 3899–3905 (2014).
18. V. Helm, A. Humbert, H. Miller, *Cryosphere* **8**, 1539–1559 (2014).
19. E. R. Thomas, G. J. Marshall, J. R. McConnell, *Geophys. Res. Lett.* **35**, 1706 (2008).
20. Materials and methods are available as supplementary materials on *Science* Online.
21. H. Zwally, M. B. Giovinetto, M. A. Beckley, J. L. Saba, *Antarctic and Greenland Drainage Systems* (GSF Cryospheric Sciences Laboratory, Greenbelt, MD, 2012).
22. A. Shepherd *et al.*, *Science* **338**, 1183–1189 (2012).
23. S. R. M. Ligtenberg, M. M. Helsen, M. R. van den Broeke, *Cryosphere* **5**, 809–819 (2011).
24. J. M. van Wessem *et al.*, *J. Glaciol.* **60**, 761–770 (2014).
25. B. Tapley, S. Bettadpur, M. Watkins, C. Reigber, *Geophys. Res. Lett.* **31**, L09607 (2004).
26. T. K. Dupont, R. B. Alley, *Geophys. Res. Lett.* **32**, 4503 (2005).
27. D. Holland, R. Thomas, B. de Young, M. Ribergaard, B. Lyberth, *Nat. Geosci.* **1**, 659–664 (2008).
28. S. Schmidtko, K. J. Heywood, A. F. Thompson, S. Aoki, *Science* **346**, 1227–1231 (2014).
29. T. O. Holt, N. F. Glasser, D. J. Quincey, M. R. Siegfried, *Cryosphere* **7**, 797–816 (2013).
30. M. A. Depoorter *et al.*, *Nature* **502**, 89–92 (2013).
31. E. Rignot, S. Jacobs, J. Mouginot, B. Scheuchl, *Science* **341**, 266–270 (2013).
32. F. S. Paolo, H. A. Fricker, L. Padman, *Science* **26**, (2015).
33. J. Mouginot, E. Rignot, B. Scheuchl, *Geophys. Res. Lett.* **41**, 1576–1584 (2014).
34. P. Dutrieux *et al.*, *Science* **343**, 174–178 (2014).
35. T. C. Sutterley *et al.*, *Geophys. Res. Lett.* **41**, 8421–8428 (2014).
36. E. Rignot, J. Mouginot, B. Scheuchl, *Science* **333**, 1427–1430 (2011).
37. The Scientific Committee on Antarctic Research, *SCAR Antarctic Digital Database*; available at www.add.scar.org, accessed 26 March 2015.

ACKNOWLEDGMENTS

B.W. and J.L.B. jointly conceived the study, interpreted the results, and wrote the article. B.W. processed the GRACE and Cryosat-2 L2 data. A.M.-E. combined the ICESat-Envisat elevation rates and derived the GVIIS front positions. V.H. developed the Cryosat-2 retracker and processed the L1B data. T.F. processed the Envisat data. J.M.v.W., S.R.M.L., and M.R.v.d.B. developed and ran the SMB and firn model. All authors commented on the manuscript. We thank J. Wahr and G. A for their help with the Glacial Isostatic Adjustment correction and elastic loading. G. Moholdt and A. Gardner are acknowledged for the fruitful discussions on altimetry. B.W. is funded by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme (FP7-PEOPLE-2011-IOF-301260). J.L.B. and A.M.-E. were supported by Natural Environment

Research Council grant NE/I027401/1. J.L.B., B.W., and A.M-E. also acknowledge support through European Space Agency contract 4000107393/12/I-NB, "REGINA." V.H. was supported by the German Ministry of Economics and Technology (grant 50EE1331). J.M.v.W., S.R.M.L., and M.R.v.d.B. acknowledge support from the Netherlands Polar Program of the Netherlands Organization for Scientific Research, section Earth and Life Sciences (NWO/ALW/NPP). The data sets used in this study can be found at http://pangaea.de.

SUPPLEMENTARY MATERIALS
www.sciencemag.org/content/348/6237/899/suppl/DC1
Materials and Methods
Figs. S1 to S9
Tables S1 to S4
References (*38*–*78*)

26 December 2014; accepted 28 April 2015
10.1126/science.aaa5727

SANITATION SUBSIDIES

Encouraging sanitation investment in the developing world: A cluster-randomized trial

Raymond Guiteras,[1] James Levinsohn,[2] Ahmed Mushfiq Mobarak[2]*

Poor sanitation contributes to morbidity and mortality in the developing world, but there is disagreement on what policies can increase sanitation coverage. To measure the effects of alternative policies on investment in hygienic latrines, we assigned 380 communities in rural Bangladesh to different marketing treatments—community motivation and information; subsidies; a supply-side market access intervention; and a control—in a cluster-randomized trial. Community motivation alone did not increase hygienic latrine ownership (+1.6 percentage points, $P = 0.43$), nor did the supply-side intervention (+0.3 percentage points, $P = 0.90$). Subsidies to the majority of the landless poor increased ownership among subsidized households (+22.0 percentage points, $P < 0.001$) and their unsubsidized neighbors (+8.5 percentage points, $P = 0.001$), which suggests that investment decisions are interlinked across neighbors. Subsidies also reduced open defecation by 14 percentage points ($P < 0.001$).

One billion people, or about 15% of the world's population, currently practice open defecation (OD), and another 1.5 billion do not have access to improved sanitation (*1*). Despite the existence of simple, effective solutions such as pour-flush latrines, poor sanitation causes 280,000 deaths per year (*2*) and may contribute to serious health problems such as stunting or tropical enteropathy (*3*–*5*).

The issue has attracted attention and resources from governments and development institutions. In 2012, the United Nations Children's Fund (UNICEF) spent USD 380 million on programs focused on water, sanitation, and hygiene for children (*1*). The World Bank's Water and Sanitation Program plans to direct USD 200 million in government and private funds to improve sanitation for 50 million people during the 2011–2015 period (*6*). In India, where over half the population practices open defecation (*7*), Prime Minister Narendra Modi declared "toilets first, temples later" during a 2013 speech and pledged to eliminate OD by 2019 (*8*–*10*).

However, disagreement remains over how best to increase sanitation coverage. Policymakers must allocate scarce resources among strategies such as demand generation (e.g., information campaigns, behavior change programming), direct provision of toilets to schools or households, or subsidizing consumers (*11*). Subsidies are particularly controversial, with practitioners concerned that subsidies may undermine intrinsic motivation or cause dependency (*12*, *13*). For example, the Government of India's Total Sanitation Campaign (TSC) used the rhetoric of "community-led," "people-centred," and "demand driven" to build one toilet for every 10 rural residents between 2001 and 2011 (*14*), but critics argue that the program as implemented was "infrastructure-centred" and "supply-led" (*15*). Recent studies of TSC find modest impacts on sanitation coverage and OD (*16*, *17*).

At the root of this disagreement is uncertainty about the reasons for low coverage. If the major constraints are poverty and the collective action problem posed by negative health externalities, then economic theory suggests that subsidies are necessary. If the key constraints are lack of information about the benefits of sanitation and the absence of strong community norms against OD, then programs such as Community-Led Total Sanitation (CLTS), which seek to change norms and create social pressure, could be sufficient without subsidies. Even when households are willing to pay for hygienic latrines, supply failures such as lack of access to markets where toilet components are sold, or lack of information about quality or installation methods, may impede adoption (*18*).

We measured the effects of alternative policies on investment in hygienic latrines using a cluster-randomized trial in 380 rural communities (18,254 households in 107 villages) in the Tanore district of northwest Bangladesh. Although sanitation coverage has increased markedly in rural Bangladesh in recent decades (*1*), progress in Tanore, located in the poorest region of the country, has been slower. At baseline, 31% of households reported that their primary defecation site was either no latrine (OD) or an unimproved latrine, and only 50% had regular access to a hygienic latrine. The intervention was conducted in 2012, and we collected follow-up data in 2013 (fig. S1).

We randomized communities to different treatments: a community motivation and health information campaign, called the Latrine Promotion Program (LPP); motivation and health information combined with subsidies for the purchase of hygienic latrines; a supply-side market access intervention linking villagers with suppliers and providing information on latrine quality and availability; and a control group receiving no interventions (*19*).

LPP was a multiday, neighborhood-level exercise to raise awareness of the problems caused by poor sanitation and to motivate the community to increase coverage of hygienic latrines. The design of LPP follows that of CLTS, an information and motivation intervention that has been implemented in over 60 countries worldwide (*20*). The nongovernmental organizations that implemented this project, WaterAid Bangladesh and Village Education Resource Center (VERC), were instrumental in the creation of CLTS (*13*). The design of LPP conformed closely to the principles of CLTS, although LPP differed in emphasizing the importance of hygienic latrines, rather than simply ending OD.

In villages assigned to the "subsidy" treatment, households in the bottom three-quarters of the wealth distribution were eligible to participate in a public lottery awarding subsidy vouchers. These vouchers provided a 75% discount on the components of any of three models of latrine, priced (after subsidy) USD 5.5, USD 6.5, and USD 12. Households were responsible for delivery and installation costs of USD 7 to 10. To study the extent of demand spillovers across neighbors, we randomized the share of lottery winners at the neighborhood level into low, medium, and high intensity, corresponding to approximately 25, 50, and 75%.

The "supply" treatment was intended to improve the functioning of markets by providing

[1]Department of Economics, University of Maryland, College Park, MD 20742, USA. [2]School of Management, Yale University, New Haven, CT 06520, USA.
*Corresponding author. E-mail: ahmed.mobarak@yale.edu

technical assistance and information. In communities assigned to the supply treatment, VERC selected a local resident with technical skills and trained him as a latrine supply agent (LSA). The LSA received a fixed salary to provide information to neighborhood residents on (i) where to purchase a hygienic latrine, (ii) how to assess the quality of a latrine offered for sale, and (iii) how to install and maintain a latrine.

These treatments were randomized in a two-stage design: First, communities were randomly assigned to treatments; then, within subsidy communities, eligible households participated in household-level lotteries for subsidy vouchers. This randomization resulted in neighborhoods being assigned to five main categories (fig. S3): (1) control (number of neighborhoods, $N = 66$); (2) LPP only ($N = 49$); (3) LPP + subsidy ($N = 115$); (4) supply only ($N = 34$); and (5) LPP + subsidy + supply ($N = 116$). Groups 1, 3, 4, and 5 represent a 2 × 2 experimental design, where the demand-side strategies (LPP plus subsidies) and the supply-side strategy are implemented either in isolation or in combination and compared to a pure control group. Adding group 2 (LPP only) allows us to separate the effect of subsidies from the LPP information and motivation campaign. The 231 subsidy neighborhoods (groups 3 and 5) were randomized in equal proportion to low, medium, and high subsidy intensity.

When we consider all treatments jointly, the randomization produced an allocation of villages that was well balanced on key characteristics, including the share of households with access to hygienic latrines (table S1). In pairwise comparisons of individual treatment arms to the control group, we find that the "supply only" group had higher rates of latrine ownership and access at baseline. Because of this imbalance, we include controls for baseline ownership (or access) in our analysis. Adding controls generally affects coefficients on the supply only treatment (*21*).

The primary outcomes of interest are household access to and ownership of a hygienic latrine, defined as a latrine that safely confines feces (*22*). For pour-flush latrines, the relevant type in this context, this typically requires a water seal to block flies and other insects and a sealed pit to store fecal matter for safe disposal. We classify a latrine as hygienic if it has an intact slab, has an intact seal, and conveys feces to a sealed pit (*23*).

We focus on hygienic latrines because the safe confinement and disposal of feces are most likely to improve health (*24*). We also report effects on access to and ownership of any latrine, including nonhygienic models, because any latrine use that replaces OD is a common policy goal. Finally, we report effects on OD because reductions in reported OD help confirm that latrines are actually being used.

Outcome data were collected in two household surveys: a baseline conducted December 2011 to February 2012 and a follow-up conducted May to July 2013 (fig. S1). Data on the presence and type of latrine come from direct observation by surveyors, with ownership status determined through interviews with the household. Access and OD are based on household self-reports. Data on village and neighborhood treatment assignment and household lottery outcomes were compiled from administrative records. Wealth was proxied by landholdings reported in the baseline survey.

We first estimate overall program effects by comparing outcomes across the randomized community-level treatments, controlling for baseline levels and union fixed effects. Estimates presented here pertain to the households eligible for subsidies (*25*).

Figure 1, A to C, presents the main results (*26*). Community-based motivation alone did not increase coverage: Relative to the control group, being assigned to an LPP-only village resulted in no change in access to any latrine [−0.5 percentage points (pp), $P = 0.82$] or in access to a hygienic latrine (−0.6 pp, $P = 0.85$). However, the combination of demand-side strategies that add subsidies targeted to the poor with community motivation did increase coverage significantly. Compared to the control group, households in LPP + subsidy villages were 7.3 pp more likely to have access to any latrine ($P < 0.001$) and 14.3 pp more likely to have access to a hygienic latrine ($P < 0.001$). These are average effects at the village level, aggregating across subsidy lottery winners and losers. In contrast, the supply-side treatment by itself did not lead to a statistically significant increase in either outcome (any latrine +2.7 pp, $P = 0.38$; hygienic latrine +3.0 pp, $P = 0.58$). Finally, adding the supply treatment to the combined demand-side strategies (LPP + subsidy) does not change the effectiveness of the subsidies. There are statistically significant increases in latrine access in both groups where subsidies are provided, and the difference between those two treatment arms is not statistically significant (any latrine +0.5 pp, $P = 0.72$; hygienic latrine −0.2 pp, $P = 0.94$).

Because 78% of households had access to a latrine at baseline, the 7.3 pp subsidy effect represents a 9.4% increase in latrine access. The effect on ownership of any latrine (12.1 pp; table S2) is larger, representing a 20% increase over the baseline ownership rate. The larger effect on ownership suggests that the intervention moved some households that were previously sharing into individual ownership. The subsidy vouchers were actually provided for investment in hygienic latrines, and the subsidy effects are largest (14 to 15 pp, or 29 to 36% increase relative to control) for those outcomes.

The LPP only and supply only treatments do not have statistically significant effects on adult OD behavior; however, adding subsidies to LPP reduces OD rates among adults by 9.0 pp ($P = 0.02$), representing a 22% reduction relative to the control group (Fig. 1C). The reductions among men and women are similar (*27*).

If one household's investment in a toilet has spillover effects on its neighbors' investment decisions, that has implications for the optimal targeting of subsidies and for the share of community members who should be subsidized. To investigate whether there is a social multiplier in sanitation investments, we analyze the effects of the share of other households in the neighborhood offered subsidies [which we randomized into low-, medium-, and high-intensity (L, M, and H) neighborhoods] on latrine investment and OD. Evidence for a social multiplier comes from comparing behavior across L, M, and H neighborhoods, holding constant each household's own lottery outcome. Figure 2 focuses on ownership rather than access, because a simple

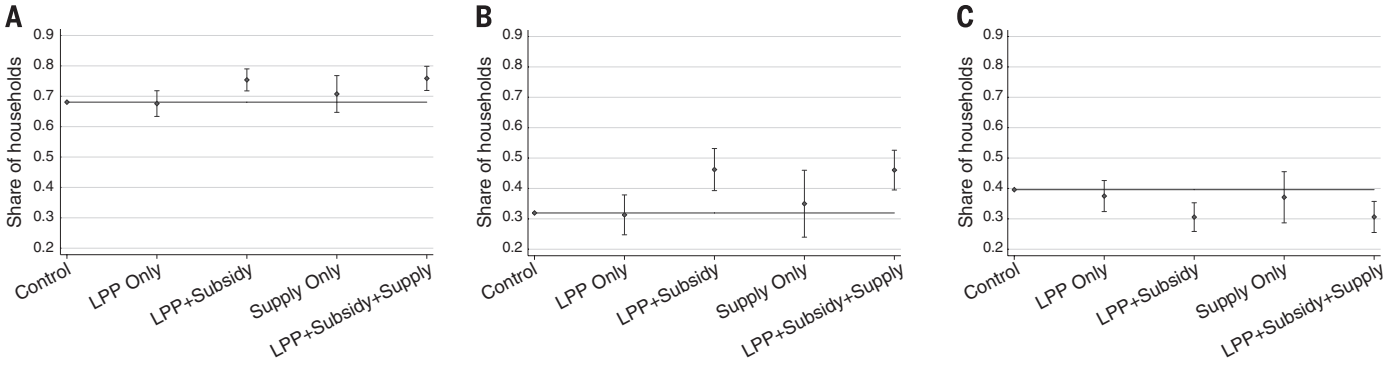


Fig. 1. Effect of supply and demand treatments on latrine access and open defecation. Figure displays the sum of the estimated coefficients and the control group means found in columns (2) and (6) of table S2 and column (2) of table S3. (**A**) Any latrine access; (**B**) hygienic latrine access; (**C**) open defecation among adults.

explanation for greater household latrine access when a larger share of neighbors receive vouchers is that the neighbors offer to share access to their toilets (*28*).

Figure 2B shows that voucher winners are more likely to own hygienic latrines than households in LPP only villages or lottery losers in subsidy villages. Furthermore, among winners, a household is more likely to convert the subsidy voucher covering half the cost of the latrine into an actual latrine investment if a larger share of its neighbors also receive vouchers. A voucher winner in a low-intensity neighborhood is 13.7 pp ($P < 0.001$) more likely to own a hygienic latrine than an eligible household in an LPP-only community. A voucher winner in a medium-intensity neighborhood is 20.9 pp ($P < 0.001$) more likely to own a hygienic latrine than an eligible household in an LPP-only community. The +7.2 pp difference between medium- and low-intensity neighborhoods is statistically significant ($P < 0.001$). Similarly, a voucher winner in a high-intensity neighborhood is 20.4 pp more likely to own a hygienic latrine than an eligible household in an LPP-only community, and the +6.7 pp difference between high- and low-intensity neighborhoods is statistically significant ($P = 0.01$). This social multiplier levels off, as there is no detectable difference in hygienic latrine ownership between winners in medium- and high-intensity neighborhoods. A similar pattern occurs in ownership of any (not necessarily hygienic) latrine (see Fig. 2A), although the estimated differences (+3.2 pp for winners in medium-intensity versus winners in low-intensity neighborhoods; +4.1 pp for winners in high-intensity versus winners in low-intensity neighborhoods) are not statistically significant ($P = 0.17$ and $P = 0.11$).

We find a similar social multiplier among eligible households that did not win a voucher. Although losing households in low-intensity neighborhoods are statistically indistinguishable from eligible households in LPP-only villages (any latrine +1.5 pp, $P = 0.56$; hygienic latrine +0.9 pp, $P = 0.70$), detectable differences emerge for losing households in medium-intensity neighborhoods (any latrine +5.8 pp, $P = 0.03$; hygienic latrine +2.7 pp, $P = 0.26$) and losing households in high-intensity neighborhoods (any latrine +5.5 pp, $P = 0.04$; hygienic latrine +6.9 pp, $P = 0.01$). The social multiplier is smaller for losing households than for winning households, which is expected because latrines were not subsidized for these households (Fig. 2C).

The more intense subsidy treatments induced not only latrine construction among neighbors but also latrine use: Households become less likely to practice OD if more of their neighbors receive subsidies (Fig. 2C). OD among adults in lottery-winning households in low-, medium-, and high-intensity neighborhoods falls by 7.2 pp ($P = 0.01$), 13.8 pp ($P < 0.001$), and 11.6 pp ($P < 0.001$) relative to adults in eligible households in control communities. These represent reductions of 18 to 35% relative to the control group mean. Even those who fail to win vouchers reduce their OD propensity (relative to the control group) by 8.8 pp ($P < 0.001$) if 50% of their neighbors win vouchers and by 8.1 pp ($P = 0.01$) if 75% of neighbors win vouchers. The decrease in OD among lottery losers in medium- and high-intensity villages is comparable to the decrease among lottery winners in low-intensity villages.

Further evidence of a social multiplier comes from the least-poor quartile of households in subsidy villages. Although they were ineligible for subsidies, they invested in latrines and reduced OD at a greater rate if a larger fraction of their poor neighbors were subsidized (*25*).

These results are consistent with a growing literature showing the importance of price as a primary barrier to adoption of health products (*29*–*31*). Current practice in sanitation sector demand-generation programming reflects a strong belief that community-based motivation is effective at moving households away from OD and toward basic latrines (*12*, *13*). However, in this context, information and motivation alone were not sufficient to increase adoption of hygienic latrines. Similarly, there was no detectable effect of an intervention providing information on the supply side (*32*). Subsidies increased coverage and reduced OD across the entire population.

This study also presents evidence of the importance of social influence and the possibility of a virtuous cycle where adoption spurs further adoption (*33*, *34*). The presence of interlinked decision-making implies that interventions will be more cost-effective if they can identify the relevant network and target that group jointly. If neighbors' decisions are interlinked, smaller subsidies targeted to multiple households in a network can generate more investment than large subsidies deployed to a few in an uncoordinated manner. Our experiments suggest that cost-effective "smart subsidy" policies should identify the threshold of investment in latrines where the social multiplier in demand is largest. The move from subsidizing 25% to subsidizing 50% of the poor produces the largest demand spillovers in our context. Asking community members to make a joint investment commitment, as in CLTS, is a potentially useful intervention, but our results suggest that this should be accompanied by targeted subsidies. Future programs could attempt to harness the interplay between subsidies and interlinked decision-making by combining financial incentives with a forum for community cooperation. More research is needed to understand the underlying mechanisms (*35*), which may include learning, changes in social norms, or technical complementarities (benefits of investment are greater when others invest).

Our study has several limitations. First, results from one study in Bangladesh may not generalize to other populations. However, the disease burden from OD is largest in the high-density rural areas of the Ganges Delta (*36*), so the results from rural Bangladesh (the most densely populated rural area of the world) are relevant for areas where the problem is most acute. Second, this study reports results for one level of subsidy (~50% of the cost of an installed latrine), and results may vary at other levels. Third, we did not include a subsidy-only treatment because the evidence suggests that providing subsidies without education is not a useful policy (*15*). We therefore cannot distinguish the effect of subsidies from the combined effect of subsidies and LPP. However, we show that LPP alone was not sufficient in this context to increase investment in hygienic latrines. Fourth, we used household self-reports of OD as a proxy for latrine use,

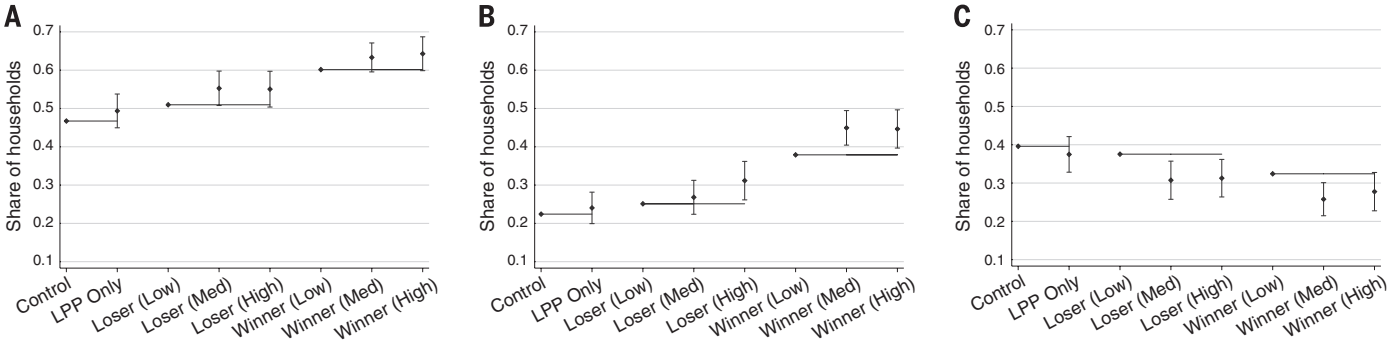


Fig. 2. Effects of the proportion of community treated on latrine ownership and open defecation for those eligible for subsidies. Figure displays the sum of the estimated coefficients and the control group means found in columns (4) and (8) of table S4 and column (2) of table S5. (**A**) Any latrine ownership; (**B**) hygienic latrine ownership; (**C**) open defecation among adults.

which may be subject to bias (*37*, *38*). Fifth, we do not measure health outcomes in this demand study, but combining our results on reductions in OD with studies that measure the relationship between OD and health outcomes (*14*, *39*–*41*) suggests that sanitation marketing interventions could plausibly produce improvements in health. Finally, the scale of this study, covering over 18,000 households and 100% samples of four subdistricts, allows us to document some of the general equilibrium changes operating via a social influence mechanism, but our results remain silent on wider general equilibrium effects operating via price mechanisms.

REFERENCES AND NOTES

1. WHO/UNICEF, "Progress on Drinking Water and Sanitation - 2014 Update," tech. rep. (WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation, Luxembourg, 2014). URL (accessed 9 April 2015): http://apps.who.int/iris/bitstream/10665/112727/1/9789241507240_eng.pdf?ua=1.
2. A. Prüss-Ustün *et al.*, *Trop. Med. Int. Health* **19**, 894–905 (2014).
3. A. D. Dangour *et al.*, *Cochrane Database Syst. Rev.* **8**, CD009382 (2013).
4. A. Lin *et al.*, *Am. J. Trop. Med. Hyg.* **89**, 130–137 (2013).
5. D. Spears, A. Ghosh, O. Cumming, *PLOS ONE* **8**, e73784 (2013).
6. World Bank WSP, "Water and Sanitation End of Year Report Fiscal Year 2013", tech. rep. (World Bank Water and Sanitation Program, Washington, DC, 2013). URL (accessed 9 April 2015): https://wsp.org/sites/wsp.org/files/publications/WSP-End-Year-Report-FY13.pdf.
7. Census Organization of India, *15th Indian Census - Houses, Household Amenities and Assets: Sanitation - Total Latrine Table*, 2011. URL (accessed 9 April 2015): www.devinfolive.info/censusinfodashboard/website/index.php/pages/sanitation/total/Households/IND.
8. S. Mojumdar, *Untitled Memorandum, no. W-11042/34/2014-NBA* (July 2014). URL (accessed 9 April 2015): www.mdws.gov.in/sites/upload_files/ddws/files/pdf/Swachh_Bharat_by_2019.pdf.
9. S. Prasad, *Swachchh Bharat by 2019* (July 2014). URL (accessed 9 April 2015): http://mdws.gov.in/hindi/sites/upload_files/ddwshindi/files/Swachchh_Bharat_by_2019.pdf.
10. Press Trust of India, "Modi's new mantra: Toilets first, temples later," *Hindustan Times* (October 2013). URL (accessed 9 April 2015): www.hindustantimes.com/newdelhi/modi-s-new-mantra-toilets-first-temples-later/article1-1130281.aspx.
11. E. Perez *et al.*, "What Does It Take to Scale Up Rural Sanitation?" Water and Sanitation Program Working Paper (July 2012). URL (accessed 9 April 2015): www.wsp.org/sites/wsp.org/files/publications/WSP-What-does-it-take-to-scale-up-rural-sanitation.pdf.
12. K. Kar, K. Pasteur, "Subsidy or self-respect?: Community led total sanitation; an update on recent developments," IDS Working Paper 257 (Institute of Development Studies, University of Sussex, Brighton, UK, 2005).
13. K. Kar, "Subsidy or self-respect? Participatory total community sanitation in Bangladesh." IDS Working Paper 184 (Institute of Development Studies, University of Sussex, Brighton, UK, 2003).
14. D. Spears, *Effects of Rural Sanitation on Infant Mortality and Human Capital: Evidence from India's Total Sanitation Campaign* (July 2012). URL (accessed 9 April 2015): www.dartmouth.edu/~neudc2012/docs/paper_86.pdf.
15. A. Hueso, B. Bell, *Water Policy* **15**, 1001–1017 (2013).
16. S. Barnard *et al.*, *PLOS ONE* **8**, e71438 (2013).
17. S. R. Patil *et al.*, *PLOS Med.* **11**, e1001709 (2014).
18. S. Trémolet, "Sanitation Markets: Using economics to improve the delivery of services along the sanitation value chain," tech. rep. (SHARE, London, UK, December 2012). URL (accessed 9 April 2015): www.shareresearch.org/LocalResources/SHARE_sanitation_markets_pathfinder_Dec_2012.pdf.
19. Materials and Methods, section S1.2.
20. R. Sigler, L. Mahmoudi, J. P. Graham, *Health Promot. Int.* **30**, 16–28 (2015).
21. Materials and methods, section S1.6.
22. S. Hanchett, M. H. Khan, L. Krieger, C. Kullmann, paper presented at the 35th WEDC International conference, "The future of water sanitation and hygiene: Innovation, adaptation and engagement in a changing world," Loughborough, UK, 2011. URL (accessed 9 April 2015): http://wedc.lboro.ac.uk/resources/conference/35/Hanchett-S-1036.pdf.
23. Materials and methods, section S1.5.
24. D. Mara, J. Lane, B. Scott, D. Trouba, *PLOS Med.* **7**, e1000363 (2010).
25. Effects on ineligibles discussed in the supplementary text, section S2.2.
26. Details in tables S2 and S3.
27. Supplementary text, section S2.1.
28. Details, including effects on access, in tables S4 and S5.
29. A. Ahuja, M. Kremer, A. P. Zwane, *Annu. Rev. Resour. Econ.* **2**, 237–256 (2010).
30. P. Dupas, *Science* **345**, 1279–1281 (2014).
31. A. M. Mobarak, P. Dwivedi, R. Bailis, L. Hildemann, G. Miller, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10815–10820 (2012).
32. Supplementary text, section S2.2.
33. K. Dickinson, S. K. Pattanayak, "Open sky latrines: Do social effects influence technology adoption in the case of a (very) impure public good?" (2009). URL (accessed 9 April 2015): www.isid.ac.in/~pu/conference/dec_09_conf/Papers/SubhrenduPattanayak.pdf.
34. S. K. Pattanayak *et al.*, *Bull. World Health Org.* **87**, 580–587 (2009).
35. A. Banerjee, A. G. Chandrasekhar, E. Duflo, M. O. Jackson, *Science* **341**, 1236498 (2013).
36. D. Spears, "How Much International Variation in Child Height Can Sanitation Explain?" Working Paper 6351 (World Bank Policy Research, 2013). URL (accessed 9 April 2015): www-wds.worldbank.org/external/default/WDSContentServer/IW3P/IB/2013/02/05/000158349_20130205082533/Rendered/PDF/wps6351.pdf.
37. T. Clasen *et al.*, *Environ. Sci. Technol.* **46**, 3295–3303 (2012).
38. M. W. Jenkins, M. C. Freeman, P. Routray, *Int. J. Environ. Res. Public Health* **11**, 8319–8346 (2014).
39. D. Coffey, "Sanitation externalities, disease, and children's anemia," Working Paper (Princeton University, March 2014). URL (accessed 9 April 2015): http://paa2014.princeton.edu/papers/140309.
40. M. Geruso, D. Spears, "Neighborhood Sanitation and Infant Mortality," Working Paper (RICE Institute, January 2015). URL (accessed 9 April 2015): www.amse-aixmarseille.fr/sites/default/files/_evenements/muslim_v2.pdf.
41. D. Spears, S. Lamba,"Effects of Early-Life Exposure to Sanitation on Childhood Cognitive Skills" (The World Bank Sustainable Development Network Water and Sanitation Program Unit, October 2013, Working Paper 6659). URL (accessed 9 April 2015): http://elibrary.worldbank.org/doi/pdf/10.1596/1813-9450-6659

ACKNOWLEDGMENTS

We are grateful to WaterAid Bangladesh and VERC for implementing the interventions and Innovations for Poverty Action for assistance running the study. We also thank the study participants for generously giving their time, as well as the field officers for their dedication to the quality of data collected. I. Ahmed, M. Ali, B. Atuesta, M. Bakhtiar, Z. Chen, L. Feeney, M. Khan, R. Mahbub, A. Moderson-Kox, L. Moore, A. Vargas, D. Wolfson, and T. Zelenska provided excellent project management and research assistance. We thank The Bill and Melinda Gates Foundation (BMGF) for generous funding. BMGF, WaterAid Bangladesh, and VERC were involved in the study design but played no role in data collection, analysis, the decision to publish, or manuscript preparation. The activities associated with this research were approved by the Yale University Human Subject Committee Protocol no. 1203009864. The authors declare no conflicts of interest. Data and replication code are publicly available at the Harvard Dataverse Network (doi:10.7910/DVN/GJDUTV).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/903/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S7
Tables S1 to S9

10 October 2014; accepted 7 April 2015
Published online 16 April 2015;
10.1126/science.aaa0491

NEUROPHYSIOLOGY

Decoding motor imagery from the posterior parietal cortex of a tetraplegic human

Tyson Aflalo,[1]* Spencer Kellis,[1]* Christian Klaes,[1] Brian Lee,[2] Ying Shi,[1] Kelsie Pejsa,[1] Kathleen Shanfield,[3] Stephanie Hayes-Jackson,[3] Mindy Aisen,[3] Christi Heck,[2] Charles Liu,[2] Richard A. Andersen[1]†

Nonhuman primate and human studies have suggested that populations of neurons in the posterior parietal cortex (PPC) may represent high-level aspects of action planning that can be used to control external devices as part of a brain-machine interface. However, there is no direct neuron-recording evidence that human PPC is involved in action planning, and the suitability of these signals for neuroprosthetic control has not been tested. We recorded neural population activity with arrays of microelectrodes implanted in the PPC of a tetraplegic subject. Motor imagery could be decoded from these neural populations, including imagined goals, trajectories, and types of movement. These findings indicate that the PPC of humans represents high-level, cognitive aspects of action and that the PPC can be a rich source for cognitive control signals for neural prosthetics that assist paralyzed patients.

The posterior parietal cortex (PPC) in humans and nonhuman primates (NHPs) is situated between sensory and motor cortices and is involved in high-level aspects of motor behavior (*1*, *2*). Lesions to this region do not produce motor weakness or primary sensory deficits but rather more complex sensorimotor losses, including deficits in the rehearsal of movements (i.e., motor imagery) (*3*–*7*). The activity of PPC neurons recorded in NHPs reflects

the movement plans of the animals, and they can generate these signals to control cursors on computer screens without making any movements (*8*–*10*). It is tempting to speculate that the animals have learned to use motor imagery for this "brain control" task, but it is of course not possible to ask the animals directly. These brain control results are promising for neural prosthetics because imagined movements would be a versatile and intuitive method for controlling external devices (*11*). We find that motor imagery recorded from populations of human PPC neurons can be used to control the trajectories and goals of a robotic limb or computer cursor. Also, the activity is often specific for the imagined effector (right or left limb), which holds promise for bimanual control of robotic limbs.

A 32-year-old tetraplegic subject, EGS, was implanted with two microelectrode arrays on 17 April 2013. He had a complete lesion of the spinal cord at cervical level C3-4, sustained 10 years earlier, with paralysis of all limbs. Using functional magnetic resonance imaging (fMRI), we asked EGS to imagine reaching and grasping. These imagined movements activated separate regions of the left hemisphere of the PPC (fig. S1). A reach area on the superior parietal lobule (putative human area 5d) and a grasp area at the junction of the intraparietal and postcentral sulci (putative human anterior intraparietal area, AIP) were chosen for implantation of 96-channel electrode arrays. Recordings were made over more than 21 months with no adverse events related to the implanted devices. Spike activity was recorded and used to control external devices, including a 17–degree-of-freedom robotic limb and a cursor in two dimensions (2D) or 3D on a computer screen.

Recordings began 16 days after implantation. The subject could control the activity of single cells through imagining particular actions. An example of volitional control is shown in movie S1. The cell is activated when EGS imagines moving his hand to his mouth but not for movements with similar gross characteristics such as imagined movements of the hand to the chin or ear. Another example (movie S2) shows EGS increasing the activity of a different cell by imagining rotation of his shoulder, and decreasing activity by imagining touching his nose. In many cases, the subject could exert volitional control of single neurons by imagining simple movements of the upper arm, elbow, wrist, or hand.

We found that EGS's neurons coded both the goal and imagined trajectory of movements. To characterize these forms of spatial tuning, we used a masked memory reach paradigm (MMR, Fig. 1A). In the task, EGS imagined a continuous reaching movement to a spatially cued target after a delay period during which the goal was removed from the screen. On some trials, motion

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA. [2]USC Neurorestoration Center and the Departments of Neurosurgery and Neurology, University of Southern California, Los Angeles, CA 90033, USA. [3]Rancho Los Amigos National Rehabilitation Center, Downey, CA 90242, USA.
*These authors contributed equally to this work. **†Corresponding author. E-mail: andersen@vis.caltech.edu**

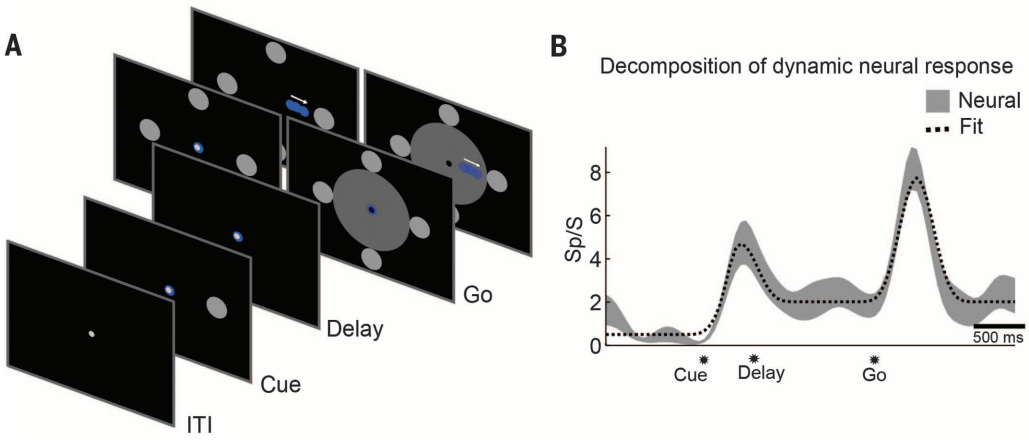


Fig. 1. Goal and trajectory coding in the PPC. (**A**) The masked memory reach task was used to quantify goal and trajectory tuning in the PPC by dissociating their respective tuning in time. EGS imagined a continuous reaching movement to spatially cued targets after a delay period. Motion of the cursor was occluded from view by using a mask in interleaved trials. (**B**) Goal and trajectory fitting. Average neural response (±SE) of a sample neuron over the duration of a trial, along with a linear model reconstruction of the time course. The linear model included components for the transient early visual response, sustained goal tuning, and transient trajectory tuning. The significance of the fit coefficients was used to determine population tuning to goal and trajectory (see Fig. 2).

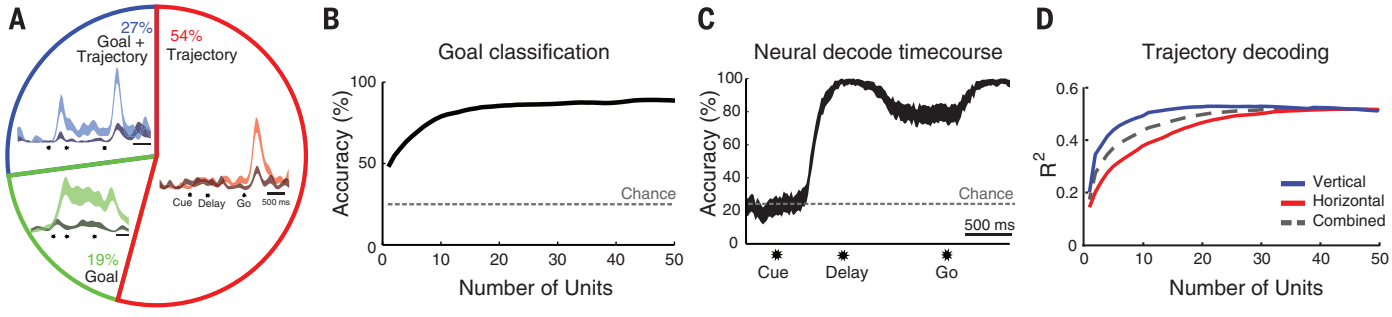


Fig. 2. Neurons in PPC encode both the goal and trajectory of movements. (**A**) The pie chart indicates the proportion of units that encode trajectory exclusively, goal exclusively, or mixed goal and trajectory. Insets show the activity (mean ± SE) for three example neurons. The lighter hue indicates response to the direction evoking maximal response; the darker hue indicates response for the opposite direction. Data taken from masked trials to avoid visual confounds (Fig. 1A). (**B**) Small populations of informative units allow accurate classification of motor goals from delay-period activity (when no visible target is present). Using a greedy algorithm, an optimized neural population for data combined across multiple days shows that >90% classification is possible with fewer than 30 units. (**C**) Temporal dynamics of goal representation. Offline analysis depicting accuracy of target classification through time [300-ms sliding window, 95% confidence interval (CI)]. Significant classification occurs within 190 ms of target presentation. (**D**) Similar to (B) but for trajectory reconstructions. All data taken from the MMR task (Fig. 1A).

of the cursor was blocked from view by using a mask. This allowed us to characterize spatial tuning for goals and trajectories (Fig. 1B) while controlling for visual confounds.

The number of recorded units was relatively constant through time, but units would appear and disappear on individual channels over the course of hours, days, or weeks (fig. S2). This allowed us to sample the functional properties of a large population of PPC neurons. From 124 spatially tuned units recorded across 7 days with the MMR task, 19% coded the goal of movement exclusively, 54% coded the trajectory of the movement exclusively, and 27% coded both goal and trajectory (Fig. 2A). Goal-tuned units supported accurate classification of spatial targets (>90% classification with as few as 30 units), representing the first known instance of decoding high-level motor intentions from human neuronal populations (Fig. 2B). The goal encoding was rapid with significant classification (shuffle test) occurring within 190 ms of cue presentation and remaining high during the delay period in which there was no visual goal present (Fig. 2C). Similarly, this population of neurons enabled reconstructions of the moment-to-moment velocity of the effector (Fig. 2D) with coefficient of determination (R^2) comparable to those reported for offline reconstructions of velocity in human M1 studies [e.g., (*12*, *13*); see also fig. S3]. In other tasks, trajectory-tuned units supported instantaneous volitional control of an anthropomorphic robotic limb at its endpoint (see movie S3).

In the MMR task, goal tuning was not directly used by the subject to control the computer interface; only the trajectory of the cursor was under brain control. To verify that goal-tuned units could support direct selection of spatial targets in closed-loop brain control, we used a direct goal classification (DGC) task (Fig. 3A). Target classification was performed by using neural activity taken during a delay period, after the visual cue was extinguished, so that neural activity was more likely to reflect intent. Online classification accuracy was significant (shuffle test); however, similar to the MMR task, aggregating neurons across days improved classification accuracy by providing a better selection of well-tuned units (Fig. 3, C and D). Goal decoding accuracy was enhanced despite the presence of more targets (six versus four) when the subject controlled the closed-loop interface using goal activity as compared to trajectory activity (Fig. 3C). Consistent with the idea that spatially tuned neural activity reflected volitional intent, decode accuracy was maintained whether the target was cued by a flashed stimulus or cued symbolically (Fig. 3, B and D).

To what degree was the spatially tuned activity specific for imagined actions of the limb? Does the activity reflect the intentions to move a specific limb, or more general spatial processes? Effector specificity was tested by asking EGS to imagine moving his left or right arm, or make actual eye movements in the symbolically cued delayed movement paradigm (Fig. 3B). We found cells that showed specificity for each effector (Fig. 4, A to C). Although the degree of specificity varied for individual units, the population showed a strong bias for imagined reaches versus saccades (Student's t test, $P < 0.05$, Fig. 4D). Whereas some neurons showed a high degree of specificity for the left and right limb,

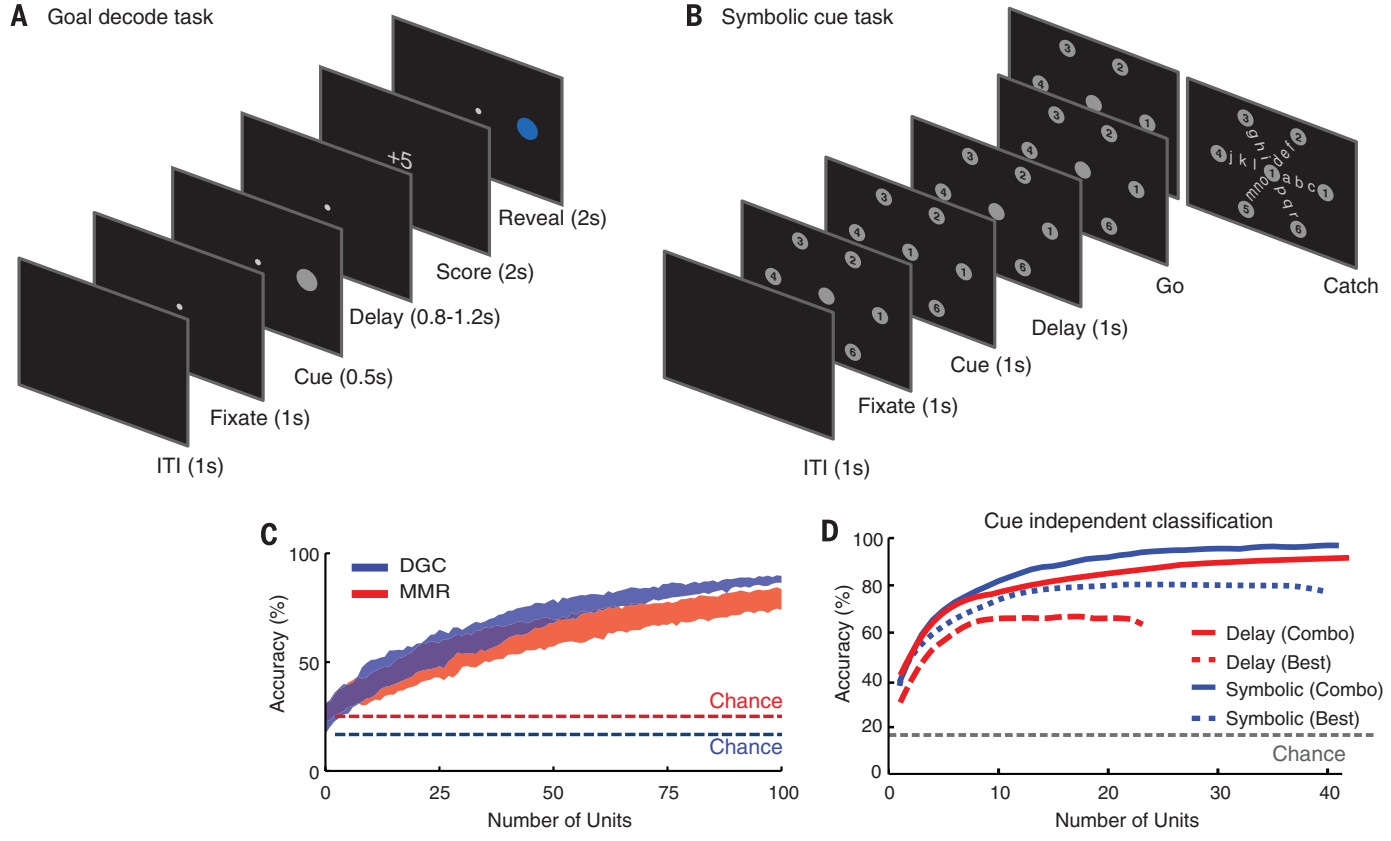


Fig. 3. Goal decoding. (**A**) Direct goal classification (DGC) task. EGS was instructed to intend motion toward a cued target through a delay period after the target was removed from the screen. Neural activity from the final 500 ms of the delay period was used to decode the location of the spatial target. EGS was awarded points depending on the relative location of the decoded and cued target. The decoded target location was presented at the end of each trial. (**B**) Symbolic task. A target grid was presented along with a number indicating the current target. The cue was removed during the delay period. A series of tones was used to cue the start and end of movements. Multiple effectors were tested in interleaved blocks. Catch trials provided a means to ensure that EGS was, on average, engaged in the task. (**C**) Estimated classification accuracy (mean with 95% CI) for variable population sizes. Populations were constructed by using randomly sampled units from the recorded population for the MMR and DGC tasks. Chance based on number of potential targets (MMR: four targets; DGC: six targets). (**D**) Greedy dropping curves show that high classification accuracy is possible whether targets are cued directly (A) or symbolically (B). Best: best single day performance; Combo: performance when combining data across days.

many reach-selective neurons were bimanual, as they frequently showed no bias for which limb EGS imagined using (Fig. 4E). The population response provided sufficient information to decode which effector EGS imagined using on a given trial (Fig. 4F).

The results show the coding of motor imagery in the human PPC at the level of single neurons and the encoding of goals and trajectories by populations of human PPC neurons. Moreover, many cells showed effector specificity, being active for imagining left-arm or right-arm movements or making actual eye movements. These results tie together NHP and human research and point to similar sensorimotor functions of the PPC in both species.

It could be argued that the results reflect visual attention rather than motor imagery. The voluntary activation of single neurons with specific imagined movements (e.g., movement of the hand to the mouth) without any visual stimulation argues against this sensory interpretation. The effector specificity also cannot be easily explained by a simple attention hypothesis.

The neural activity in delayed goal tasks is very similar to the persistent activity seen with planning in the NHP literature and attributed to the animals' intent (*14*). The PPC in NHPs codes both trajectory and goal information (*15*). The dynamics of this trajectory signal in NHPs, when compared to the kinematics of the co-occurring limb movements, suggest that the signal is a forward model of the limb movement; an internal monitor of the movement command in order to match the intended movement with actual movement for online correction (*15*). Deficits in online control in humans with PPC lesions have led investigators to propose that the PPC uses these forward models (*16*). If the trajectory signal is indeed a forward model, then EGS can generate this forward model through imagery without actually moving his limbs.

Effector specificity at the single-neuron level has been routinely reported in the PPC of NHPs (*17*). In NHPs, there is a map of intentions with areas selective for eye (lateral intraparietal area, LIP), limb (parietal reach region, PRR, and area 5d), and grasp (anterior intraparietal region, AIP) movements (*1*). Bimanual activity (left and right limb) from single PRR neurons has been reported with qualitatively similar results in the NHP (*18*). Control of two limbs across the spectrum of human behavior is challenging and requires both independent and coordinated movement between the limbs. One possibility is that units showing effector-specific and bimanual tuning could play complementary roles in independent and coordinated movements; however, more direct evidence in which EGS attempts various bimanual actions is necessary to fully test the potential for controlling two limbs from the PPC.

We have focused on the representation of motor intentions in the human PPC. Some cells appeared to code comparatively simple motor intentions, whereas others coded coordinated ethologically meaningful actions. One unexplored possibility is that the PPC also encodes nonmotor intentions such as the desire to turn on the television, or preheat the oven. As the world becomes increasingly connected through technology, the possibility of directly decoding nonmotor intentions to control one's environment may alter approaches to brain-machine interfaces (BMIs).

Neurons that constituted the recorded population would frequently change (fig. S2). This finding presents challenges for the widespread adoption of BMIs that can be addressed through a variety of techniques. One approach is the use of robust and adaptive decoding algorithms that can adapt alongside the changing neural population [e.g., (*19*)]. In the long term, the development of chronic recording technologies that can stably maintain recordings should be a priority.

This study shows that the human PPC can be a source of signals for neuroprosthetic applications in humans. The high-level cognitive aspects of movement imagery have several advantages for neuroprosthetics. The goal encoding can lead to very rapid readout of the intended movement (Fig. 2C). The PPC encodes both the goal and trajectory, which in NHPs improves decoding of movement goals when the two streams of information are combined in decoders (*10*). The bimanual representation of the limbs may allow the operation of two robotic limbs with recordings made from one hemisphere. In terms of usefulness for neuroprosthetics, it is difficult to directly compare the performance of PPC to previous studies of M1. In NHP studies, M1 has been shown to be a rich source of neural signals correlated with the trajectory of limb movements (*20*). In previous human M1 recordings, primarily the trajectory was decoded (*12*, *13*, *21*, *22*). The reported offline trajectory reconstructions from M1 populations are comparable to the values we achieved from PPC neurons (Fig. 2D) (*12*, *13*). The other aspects of encoding, e.g., goals and effectors, have not yet been examined in detail in human M1. However, it can be concluded from

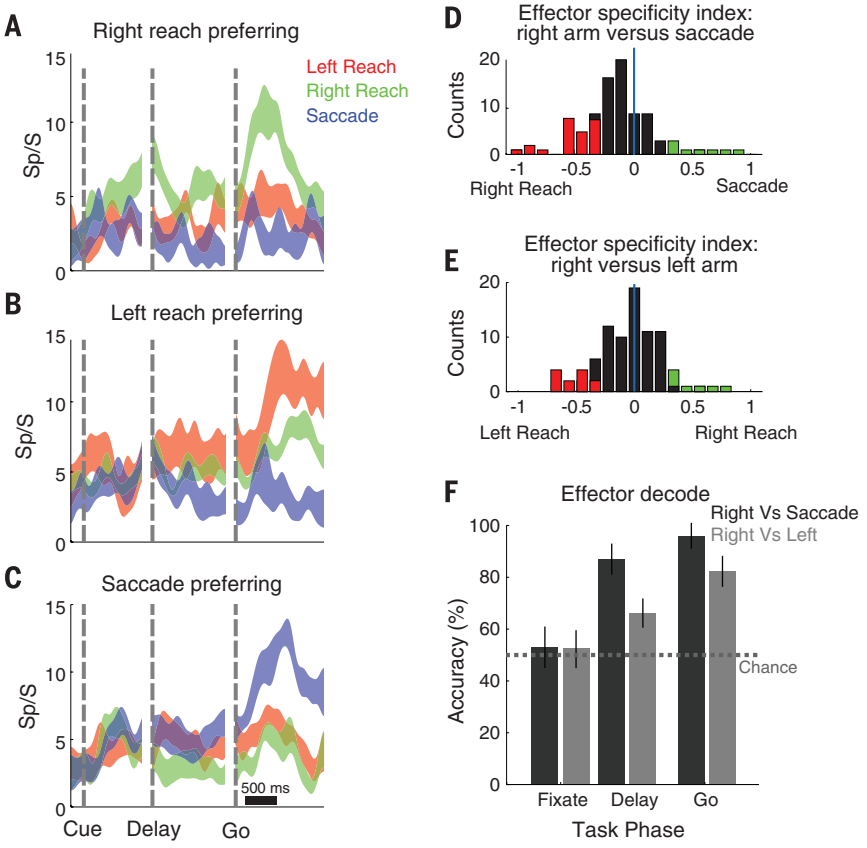


Fig. 4. Effector specificity in PPC. (**A**) Unit showing preferential activation to imagined movements of the right arm. Each trace shows the neural firing rate (mean ± SE) for the movement direction evoking the maximal response for each effector. (**B** and **C**) Same as (A), but for left arm and saccade-preferring neurons. (**D**) Population analysis. The degree of effector specificity varied across the population. Effector specificity was quantified with a specificity index based on the normalized depth of modulation (DM) for reaches versus saccades $\left(\frac{DM_{reach}-DM_{saccade}}{DM_{reach}+DM_{saccade}}\right)$. The specificity index for units that were spatially tuned to at least one effector is shown as a histogram. Colored bars indicate a significant preference for an effector. (**E**) Same as (D) but for imagined right arm versus left arm movements. (**F**) The effector used to perform the task could be decoded from the neural population (mean with 95% CI).

our study that the PPC is a good candidate for future clinical applications as it contains signals both overlapping and likely complementary to those found in M1.

REFERENCES AND NOTES

1. R. A. Andersen, C. A. Buneo, *Annu. Rev. Neurosci.* **25**, 189–220 (2002).
2. J. C. Culham, C. Cavina-Pratesi, A. Singhal, *Neuropsychologia* **44**, 2668–2684 (2006).
3. R. Balint, *Monatsschr. Psychiatr. Neurol.* **25**, 51–81 (1909).
4. M. T. Perenin, A. Vighetto, *Brain* **111**, 643–674 (1988).
5. M. A. Goodale, A. D. Milner, *Trends Neurosci.* **15**, 20–25 (1992).
6. A. Sirigu *et al.*, *Science* **273**, 1564–1568 (1996).
7. L. Pisella *et al.*, *Nat. Neurosci.* **3**, 729–736 (2000).
8. S. Musallam, B. D. Corneil, B. Greger, H. Scherberger, R. A. Andersen, *Science* **305**, 258–262 (2004).
9. M. Hauschild, G. H. Mulliken, I. Fineman, G. E. Loeb, R. A. Andersen, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17075–17080 (2012).
10. G. H. Mulliken, S. Musallam, R. A. Andersen, *J. Neurosci.* **28**, 12913–12926 (2008).
11. R. A. Andersen, S. Kellis, C. Klaes, T. Aflalo, *Curr. Biol.* **24**, R885–R897 (2014).
12. W. Truccolo, G. M. Friehs, J. P. Donoghue, L. R. Hochberg, *J. Neurosci.* **28**, 1163–1178 (2008).
13. S.-P. Kim, J. D. Simeral, L. R. Hochberg, J. P. Donoghue, M. J. Black, *J. Neural Eng.* **5**, 455–476 (2008).
14. J. W. Gnadt, R. A. Andersen, *Exp. Brain Res.* **70**, 216–220 (1988).
15. G. H. Mulliken, S. Musallam, R. A. Andersen, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 8170–8177 (2008).
16. D. M. Wolpert, S. J. Goodbody, M. Husain, *Nat. Neurosci.* **1**, 529–533 (1998).
17. L. H. Snyder, A. P. Batista, R. A. Andersen, *Nature* **386**, 167–170 (1997).
18. S. W. C. Chang, A. R. Dickinson, L. H. Snyder, *J. Neurosci.* **28**, 6128–6140 (2008).
19. S. Dangi *et al.*, *Neural Comput.* **26**, 1811–1839 (2014).
20. A. P. Georgopoulos, J. F. Kalaska, R. Caminiti, J. T. Massey, *J. Neurosci.* **2**, 1527–1537 (1982).
21. L. R. Hochberg *et al.*, *Nature* **442**, 164–171 (2006).
22. J. L. Collinger *et al.*, *Lancet* **381**, 557–564 (2013).

ACKNOWLEDGMENTS

We thank EGS for his unwavering dedication and enthusiasm, which made this study possible. We acknowledge V. Shcherbatyuk for computer assistance; T. Yao, A. Berumen, and S. Oviedo, for administrative support; K. Durkin for nursing assistance; and our colleagues at the Applied Physics Laboratory at Johns Hopkins and at Blackrock Microsystems for technical support. This work was supported by the NIH under grants EY013337, EY015545, and P50 MH942581A; the Boswell Foundation; The Center for Neurorestoration at the University of Southern California; and Defense Department contract N66001-10-4056. All primary behavioral and neurophysiological data are archived in the Division of Biology and Biological Engineering at the California Institute of Technology.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/906/suppl/DC1
Materials and Methods
Figs. S1 to S3
Movies S1 to S3
References (*23*–*26*)

20 December 2014; accepted 31 March 2015
10.1126/science.aaa5417

EPIGENETICS

Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing

Darren A. Cusanovich,[1] Riza Daza,[1] Andrew Adey,[2] Hannah A. Pliner,[1] Lena Christiansen,[3] Kevin L. Gunderson,[3] Frank J. Steemers,[3] Cole Trapnell,[1] Jay Shendure[1]*

Technical advances have enabled the collection of genome and transcriptome data sets with single-cell resolution. However, single-cell characterization of the epigenome has remained challenging. Furthermore, because cells must be physically separated before biochemical processing, conventional single-cell preparatory methods scale linearly. We applied combinatorial cellular indexing to measure chromatin accessibility in thousands of single cells per assay, circumventing the need for compartmentalization of individual cells. We report chromatin accessibility profiles from more than 15,000 single cells and use these data to cluster cells on the basis of chromatin accessibility landscapes. We identify modules of coordinately regulated chromatin accessibility at the level of single cells both between and within cell types, with a scalable method that may accelerate progress toward a human cell atlas.

Chromatin state is dynamically regulated in a cell type–specific manner (*1*, *2*). To identify active regulatory regions, sequencing of deoxyribonuclease I (DNase I) digestion products [DNase-seq (*3*)] and assay for transposase-accessible chromatin using sequencing [ATAC-seq (*4*)] measure the degree to which specific regions of chromatin are accessible to regulatory factors. However, these assays measure an average of the chromatin states within a population of cells, masking heterogeneity between and within cell types.

[1]University of Washington, Department of Genome Sciences, Seattle, WA, USA. [2]Oregon Health and Science University, Department of Molecular and Medical Genetics, Portland, OR, USA. [3]Illumina, Inc., Advanced Research Group, San Diego, CA, USA.
***Corresponding author. E-mail: shendure@uw.edu**

Single-cell methods for genome sequence (*5*), transcriptomes (*6*–*10*), DNA methylation (*11*), and chromosome conformation (*12*) have been reported. However, we presently lack technologies for genome-wide, single-cell characterization of chromatin state. Furthermore, a limitation of most such methods is that single cells are individually compartmentalized, and the nucleic acid content of each cell is biochemically processed within its own reaction volume (*13*–*16*). Processing of large numbers of cells in this way can be expensive and labor intensive, and it is difficult to work with single cells, small volumes, and low nucleic acid inputs.

We recently used combinatorial indexing of genomic DNA fragments for haplotype resolution or de novo genome assembly (*17*, *18*). Here, we adapt the concept of combinatorial indexing to intact nuclei to acquire data from thousands of single cells without requiring their individualized processing (Fig. 1A). First, we molecularly barcode populations of nuclei in each of many wells. We then pool, dilute, and redistribute intact nuclei to a second set of wells, introduce a second barcode, and complete library construction. Because the overwhelming majority of nuclei pass through a unique combination of wells, they are "compartmentalized" by the unique barcode combination that they receive. The rate of "collisions"—i.e., nuclei coincidentally receiving the same combination of indexes—can be tuned by adjusting how many nuclei are distributed to the second set of wells (fig. S1) (*19*).

We sought to integrate combinatorial cellular indexing and ATAC-seq to measure chromatin accessibility in large numbers of single cells. In ATAC-seq, permeabilized nuclei are exposed to transposase loaded with sequencing adapters ["tagmentation" (*4*, *20*)]. In the context of chromatin, the transposase preferentially inserts adapters into nucleosome-free regions. These "open" regions are generally sites of regulatory activity and correlate with DNase I hypersensitive sites (DHSs).

In the integrated method, we molecularly tag nuclei in 96 wells with barcoded transposase complexes (Fig. 1A) (*17*–*19*). We then pool, dilute, and redistribute 15 to 25 nuclei to each of 96 wells of a second plate, using a cell sorter. After lysing nuclei, a second barcode is introduced during polymerase chain reaction (PCR) with indexed primers complementary to the transposase-introduced adapters. Finally, all PCR products are pooled and sequenced, with the expectation that most sequence reads bearing the same combination of barcodes will be derived from a single cell (estimated collision rate of ~11% for experiments described here) (fig. S1).

As an initial test, we mixed equal numbers of nuclei from human (GM12878) and mouse [Patski (*21*)] cell lines, performed combinatorial cellular indexing, and sequenced the resulting

library. Although at least one mappable read was observed for most of the 9216 (96 × 96) possible barcode combinations, most barcodes were associated with very few reads. We used a conservative cutoff of 500 reads per cell (*19*), retaining 533 barcode combinations for further analysis (fig. S2A) (range: 502 to 69,847 reads per barcode combination; median: 2503). A high PCR duplication rate (~73% of mappable, nonmitochondrial reads) confirmed that the library had been sequenced to saturation. We estimate that we recovered 13 to 55% of the molecular complexity that we could expect to recover based on complexity estimates for bulk, 500-cell ATAC-seq experiments (*4*, *19*).

If each barcode combination represents either a mouse or human nucleus, then its corresponding reads should map overwhelmingly to either the mouse or human genome. Indeed, we observe that ~93% of 533 barcode combinations had >90% of their reads mapping to mouse (n = 290) or human (n = 207) (Fig. 1B). In addition, these data retain signals of chromatin accessibility in relation to nucleosome hindrance of insertion events (Fig. 1C). Furthermore, 52% of reads from mouse and 50% of reads from human single cells overlapped reference DHS maps [ENCODE (*19*, *22*)] for these cell lines (20-fold and 34-fold enrichments, respectively) (Fig. 1D and table S1).

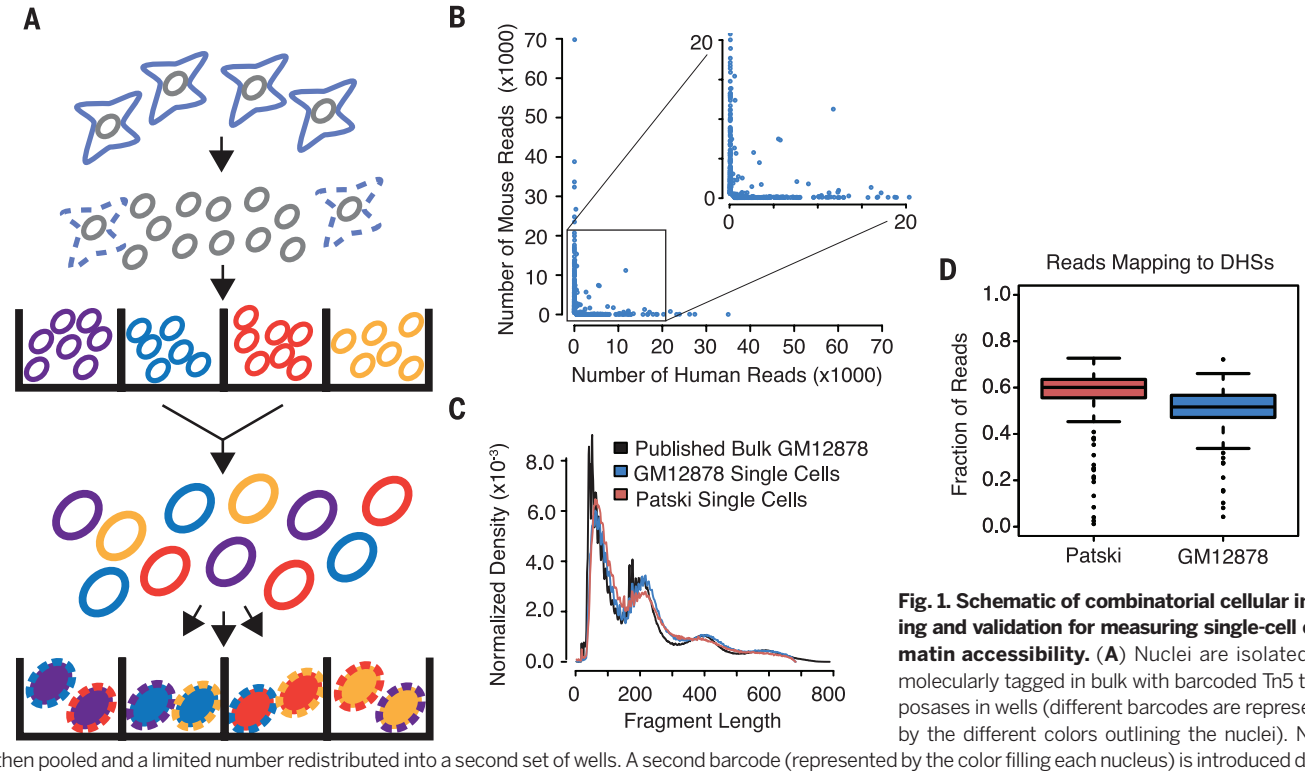


Fig. 1. Schematic of combinatorial cellular indexing and validation for measuring single-cell chromatin accessibility. (**A**) Nuclei are isolated and molecularly tagged in bulk with barcoded Tn5 transposases in wells (different barcodes are represented by the different colors outlining the nuclei). Nuclei are then pooled and a limited number redistributed into a second set of wells. A second barcode (represented by the color filling each nucleus) is introduced during PCR. (**B**) Scatterplot of number of reads mapping uniquely to human or mouse genome for individual barcode combinations. (**C**) Fragment size distribution for single-cell ATAC-seq versus published bulk ATAC-seq (*4*). (**D**) Box plot of the fraction of reads mapping to ENCODE-defined DHSs for individual Patski and GM12878 cells.

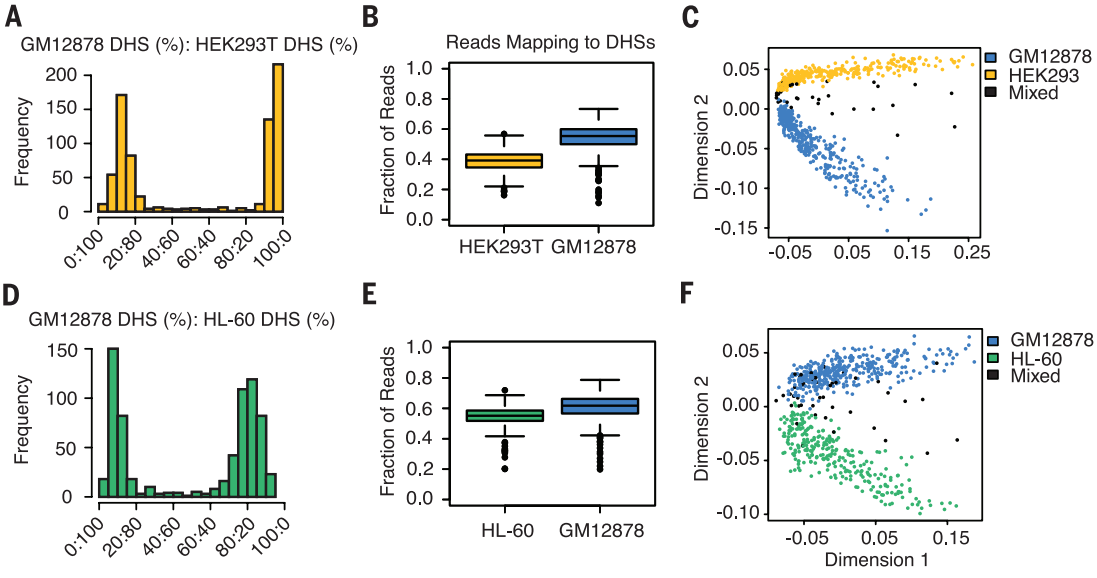


Fig. 2. Single-cell ATAC-seq deconvolutes human cell-type mixtures. (**A** to **C**) GM12878/HEK293T nuclei. (**D** to **F**) GM12878/HL-60 nuclei. [(A) and (D)] Histograms of proportions of reads mapping to cell type–specific DHSs that correspond to one cell type or the other. [(B) and (E)] Box plots of the overall fraction of reads mapping to ENCODE-defined DHSs for individual cells. [(C) and (F)] Multidimensional scaling of single-cell ATAC-seq data using pairwise Jaccard distances between cells based on DHS usage. Cell-type assignments based on proportions shown in (A) and (D).

We next sought to distinguish single cells from the same species. We mixed pairs of cell lines (HEK293T or HL-60 versus GM12878), performed combinatorial cellular indexing, and sequenced the resulting libraries to saturation (65% duplicate rate). For the mixture of HEK293T and GM12878, we recovered 748 cells with ≥500 reads (fig. S2B) (range: 502 to 28,712 reads; median: 1685 reads). Focusing on reads mapping to previously defined cell-type exclusive DHS sites (fig. S3A) (*19*, *22*), we observe a bimodal distribution, with nearly all cells assignable to one of the two cell types (~95% of 748; defined by ≥70% of reads mapping to cell type–specific DHSs corresponding to one cell type or the other) (Fig. 2A). The fraction of reads mapping to reference DHSs in single cells was again strongly enriched [41% (14-fold enrichment) for HEK293T and 52% (18-fold enrichment) for GM12878)] (Fig. 2B and table S1). About 57% of 181,379 distinct sites from the reference DHS maps were observed as accessible in at least one cell. Some fraction of these may be spurious overlaps, but this provides an upper bound on the number of DHSs for which we recovered accessibility information. Individual cells ranged in coverage of this DHS map from 29 to 5890 sites (fig. S4) (median: 429 sites).

For the mixture of HL-60 and GM12878, we recovered 700 cells (fig. S2C) (range: 500 to 21,887 reads; median: 1390 reads; 64% duplicate rate). Although both are representative of the hematopoietic lineage, 94% of cells were assignable based on the same criteria used for HEK293T/GM12878 (Fig. 2D and fig. S3B). The fraction of reads mapping to reference DHSs was again strongly enriched [55% (16-fold enrichment) for HL-60 and 59% (18-fold enrichment) for GM12878] (Fig. 2E and table S1). About 46% of 230,632 distinct sites from the reference DHS maps were observed as accessible in at least one cell, with individual cells ranging in coverage from 72 to 4687 sites (fig. S4) (median: 442 sites).

We next examined whether single cells within a heterogeneous mixture could be clustered in an unsupervised manner. Importantly, at the level of single cells, chromatin accessibility is a nearly binary phenomenon (~2 genome equivalents per cell), in contrast with the dynamic range of mRNA transcripts within single cells. Thus, we reasoned that we would require observations across each of many single cells to generate quantitative estimates for accessibility of a particular site in a particular cell type, within a heterogeneous population.

For each cell-type mixture, we defined the union of ENCODE DHSs [analogous to how RNA-seq transcript quantification relies on a catalog of transcript models (*19*)] and created a binary matrix where DHS sites were scored as "used" or "unused" in each cell. We then calculated Jaccard distances between pairs of cells on the basis of the degree of shared DHS usage. Applying multidimensional scaling to these distances, the first dimension was strongly correlated with the read depth of each cell (fig. S5) (Spearman's rho of ~0.95), whereas the second dimension separated cells consistently with our crude cell-type assignments (Fig. 2, C and F). The extent of discrimination between cell types is proportional to read depth, but even with relatively few reads, individual cells can be clustered on the basis of shared DHS usage alone. To evaluate whether our data provided reproducible and quantitative estimates of the accessibility of DHSs, we used GM12878-assigned cells from all three experiments described above as biological replicates. For each experiment, we summed the number of cells "using" each site and compared these counts between replicates (Spearman's rho's of 0.64 to 0.69, or 0.54 to 0.62 when restricted to sites observed in ≥5 cells in each replicate) and also compared them with bulk ATAC-seq measurements from 500 GM12878 cells (fig. S6) [Spearman's rho's of 0.61 to 0.7 (*4*)]. This positive correlation shows that sites that are more sensitive in bulk experiments are also more commonly observed in single cells. Furthermore, these correlations are not far from the range of 0.64 to 0.72 for replicate bulk measurements from the 500-cell ATAC-seq libraries.

To identify individual DHSs with significant differences in accessibility between different cell types (based on single-cell data from the GM12878/HL-60 mixture), we performed likelihood ratio tests within the framework of a generalized linear model. We identified 1666 sites [out of 52,479 DHSs tested (*19*)] that were differentially accessible at a false discovery rate (FDR) of 0.05. Interestingly, only about half of these sites are cell-type exclusive in the reference DHS maps (381 GM12878-exclusive and 472 HL-60-exclusive); differentially accessible DHSs are marginally enriched for GM12878-specific sites (hypergeometric $P = 0.04$) and strikingly enriched for HL-60 sites ($P = 2.2 \times 10^{-15}$). They are also larger [1184 base pairs (bp) versus 580 bp

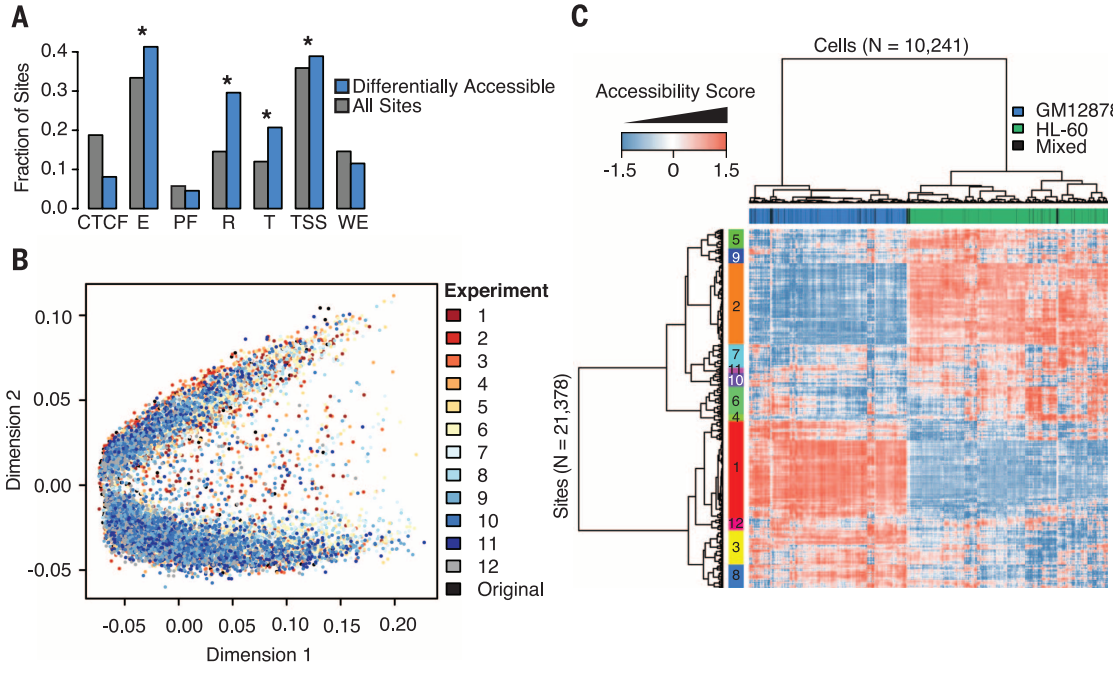


Fig. 3. Single-cell ATAC-seq identifies functionally relevant differences in accessibility between cell types. (**A**) Bar plot for relative fraction of DHSs overlapping each chromatin state (HL-60 versus GM12878). Gray bars show frequencies for all sites tested. Blue bars show frequencies for differentially accessible sites. CTCF, CTCF-enriched element; E, predicted enhancer; PF, predicted promoter flanking region; R, predicted repressed; T, predicted transcribed; TSS, predicted promoter region; WE, predicted weak enhancer. *, significant difference in proportions. Values do not add to 1 because sites can overlap multiple chromatin states. (**B**) Multidimensional scaling of chromatin accessibility data for 14,533 cells (GM12878/HL-60 mixtures from 13 experiments on four dates). (**C**) Heat map of hypersensitive site usage for 10,241 cells (columns) at 21,378 DHSs (rows) (GM12878/HL-60 mixtures). Colors indicate accessibility of sites after latent semantic indexing. Top color bar is coded by cell-type assignments (green, HL-60; blue, GM12878; black, unassigned). Left color bar indicates modules formed by clustering DHSs.

median; Wilcoxon rank sum $P = 3.4 \times 10^{-247}$], observed in more cells (10 cells versus 3 cells median; Wilcoxon rank sum $P \approx 0$), and enriched for "enhancer" (hypergeometric $P = 4.3 \times 10^{-12}$), "repressed" ($P = 1.5 \times 10^{-57}$), "transcribed" ($P = 7.4 \times 10^{-25}$), and "transcription start site" ($P = 5.1 \times 10^{-3}$) annotations in GM12878, relative to sites not identified as differentially accessible (Fig. 3A) (*19*).

We next linked differentially accessible sites defined from single cells to the genes they potentially regulate (*2*) and compared these to genes differentially expressed between GM12878 and HL-60 (*19*). Of 8268 genes linked to ≥1 DHS and expressed in both cell types, 4095 were differentially expressed and 2211 were linked to ≥1 differentially accessible DHS (FDR 0.05). Although the DHS-gene linkages are imperfect, we observe a significant overlap of differentially expressed and differentially accessible genes (1162 genes overlap; hypergeometric $P = 4.8 \times 10^{-4}$). The genes linked to DHSs identified as differentially accessible are enriched for lymphoid and myeloid lineage annotations—e.g., "cytokine signaling" and "antigen processing" (figs. S7 and S8).

To optimize combinatorial cellular indexing, we tested 12 conditions on 3 days, always with GM12878/HL-60 mixtures. We collected as many as nearly 1500 cells in a single experiment, and we improved the median read depth to >3000 per cell in some experiments (figs. S9 to S11). We merged chromatin accessibility maps for 14,533 single cells (all GM12878/HL-60) and conducted multidimensional scaling. Although the actual mixture proportion varied between experiments, the clustering of the two cell types was highly robust to experimental condition (Fig. 3B). With this full complement of cells, ~96% of 230,632 potential sites in our DHS reference map are observed as accessible in at least one cell (individual cells covering between 4 and 12,333 sites (median: 664 sites) (fig. S4).

We used latent semantic indexing to reduce the dimensionality of this matrix [after filtering out low coverage cells and rarely used sites (*19*)], yielding a heat map of chromatin accessibility for 10,241 cells at 21,378 DHSs (Fig. 3C and fig. S12). This resulted in two large clades corresponding to the two cell types, while also identifying the subset of sites underlying that separation. Additionally, we observe a number of smaller modules of DHSs that exhibit coordinately regulated chromatin accessibility. Linking these sites again to the genes they potentially regulate (*2*), the major modules are enriched for gene ontology terms consistent with the two cell types (e.g., "osteoclast differentiation" for a module more open in HL-60) (Fig. 3C and figs. S13 and S14).

To evaluate cell-to-cell variation within a cell type, we took the subset of cells classified as GM12878 and repeated latent semantic indexing (*19*), yielding a heat map of chromatin accessibility for 4118 cells at 22,755 DHSs. Hierarchical clustering identified four major subgroups of single cells and seven modules of coordinately regulated chromatin accessibility (Fig. 4A). These modules of DHSs are enriched for binding by particular transcription factors (hypergeometric FDR 0.10) (fig. S15), in some cases quite strongly, and are linked to genes associated with immune response, cell cycle regulation, and other processes (figs. S16 and S17). Importantly, although we included samples from experiments conducted on different days, the cell subtypes do not cluster by experiment (figs. S18 and S19), and the enrichments for transcription factor binding within subtype-defining modules are apparent even with subsets of the data (figs. S20 and S21). Sites in modules 1 and 2 are highly enriched for binding by transcription factors such as nuclear factor κB (NF-κB) and other factors downstream of the B cell receptor (*19*). The four GM12878 subtypes appear principally defined by the activation status of these two modules, suggesting that variability across the cells is driven by NF-κB activity. These results indicate that even within an apparently homogeneous cell type, we are able to identify subsets of cells with differences in their regulatory landscape related to cell cycle and possibly environmental signals. Focusing on individual loci within GM12878,

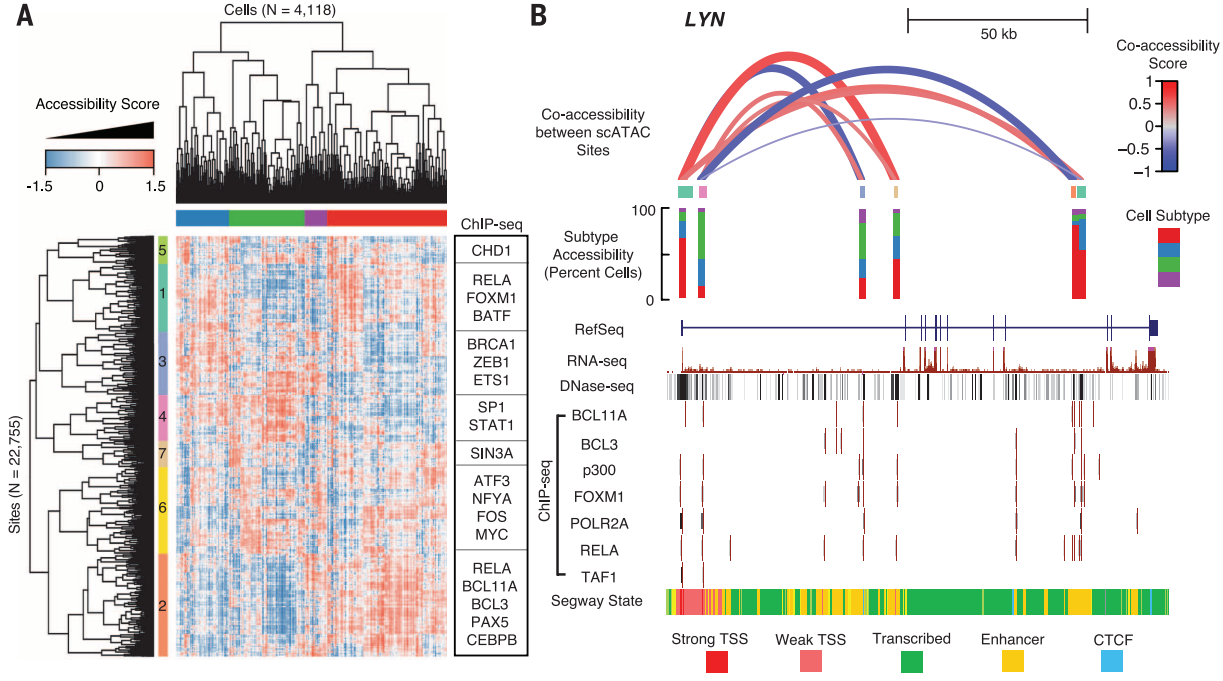


Fig. 4. Single-cell ATAC-seq identifies GM12878 subtypes. (**A**) Heat map of chromatin accessibility measures after latent semantic indexing of DHS usage shows that GM12878 cells cluster into subpopulations. Modules of coordinately accessible chromatin accessibility are significantly enriched for binding of selected transcription factors (TFs) (examples on right). (**B**) Detailed depiction of *LYN* locus. The top shows coaccessibility scores between the transcription start sites and four putative enhancers in the region, which are Pearson correlation values of latent semantic indexing–based accessibility scores between cells, for six DHSs present in this region. Height and thickness of each loop indicates the strength of correlation (red, positive; blue, negative). Middle shows in which subtypes [defined in top bar of (A)] these elements are most often accessible. Bottom shows ENCODE data for this region from the University of California–Santa Cruz browser, including transcript model, DHS peaks, chromatin immunoprecipitation sequencing (ChIP-seq) binding profiles for several TFs, and predicted chromatin state.

we observe sets of regulatory sites that exhibit patterns of coordinated regulation (e.g., *LYN*, encoding a tyrosine kinase involved in B cell signaling) (Fig. 4B), although reproducibility of these patterns across biological replicates was modest (fig. S22). Given the sparsity of the data, identifying pairs of coaccessible DNA elements within individual loci is statistically challenging and merits further development.

We report chromatin accessibility maps for >15,000 single cells. Our combinatorial cellular indexing scheme could feasibly be scaled to collect data from ~17,280 cells per experiment by using 384-by-384 barcoding and sorting 100 nuclei per well (assuming similar cell recovery and collision rates) (fig. S1) (*19*). Particularly as large-scale efforts to build a human cell atlas are contemplated (*23*), it is worth noting that because DNA is at uniform copy number, single-cell chromatin accessibility mapping may require far fewer reads per single cell to define cell types, relative to single-cell RNA-seq. As such, this method's simplicity and scalability may accelerate the characterization of complex tissues containing myriad cell types, as well as dynamic processes such as differentiation.

REFERENCES AND NOTES

1. A. B. Stergachis *et al.*, *Cell* **154**, 888–903 (2013).
2. R. E. Thurman *et al.*, *Nature* **489**, 75–82 (2012).
3. A. P. Boyle *et al.*, *Cell* **132**, 311–322 (2008).
4. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, *Nat. Methods* **10**, 1213–1218 (2013).
5. N. Navin *et al.*, *Nature* **472**, 90–94 (2011).
6. A. R. Wu *et al.*, *Nat. Methods* **11**, 41–46 (2014).
7. D. A. Jaitin *et al.*, *Science* **343**, 776–779 (2014).
8. Q. Deng, D. Ramsköld, B. Reinius, R. Sandberg, *Science* **343**, 193–196 (2014).
9. A. K. Shalek *et al.*, *Nature* **498**, 236–240 (2013).
10. C. Trapnell *et al.*, *Nat. Biotechnol.* **32**, 381–386 (2014).
11. S. A. Smallwood *et al.*, *Nat. Methods* **11**, 817–820 (2014).
12. T. Nagano *et al.*, *Nature* **502**, 59–64 (2013).
13. J. Gole *et al.*, *Nat. Biotechnol.* **31**, 1126–1132 (2013).
14. H. C. Fan, G. K. Fu, S. P. A. Fodor, *Science* **347**, 1258367 (2015).
15. A.-E. Saliba, A. J. Westermann, S. A. Gorski, J. Vogel, *Nucleic Acids Res.* **42**, 8845–8860 (2014).
16. X. Pan, *Single Cell Biol.* **3**, 106 (2014).
17. A. Adey *et al.*, *Genome Res.* **24**, 2041–2049 (2014).
18. S. Amini *et al.*, *Nat. Genet.* **46**, 1343–1349 (2014).
19. Materials and methods are available as supplementary materials on *Science* Online.
20. A. Adey *et al.*, *Genome Biol.* **11**, R119 (2010).
21. F. Yang, T. Babak, J. Shendure, C. M. Disteche, *Genome Res.* **20**, 614–622 (2010).
22. The ENCODE Project Consortium, *Nature* **489**, 57–74 (2012).
23. A. Regev, The Human Cell Atlas; www.genome.gov/Multimedia/Slides/GSPFuture2014/10_Regev.pdf.

ACKNOWLEDGMENTS

We thank the Trapnell and Shendure laboratories, particularly R. Hause, C. Lee, V. Ramani, R. Qiu, Z. Duan, and J. Kitzman, for helpful discussions; D. Prunkard, J. Fredrickson, and L. Gitari in the Rabinovitch laboratory for their exceptional assistance in flow sorting; and C. Disteche for the Patski cell line. This work was funded by an NIH Director's Pioneer Award (1DP1HG007811 to J.S.) and a grant from the Paul G. Allen Family Foundation (J.S.). C.T. is supported in part by the Damon Runyon Cancer Research Foundation (DFS-#10-14). All sequencing data are available from the NIH National Center for Biotechnology Information Gene Expression Omnibus (accession number GSE68103). L.C., K.L.G., and F.J.S. declare competing financial interests in the form of stock ownership and paid employment by Illumina, Inc. One or more embodiments of one or more patents and patent applications filed by Illumina may encompass the methods, reagents, and data disclosed in this manuscript. All methods for making the transposase complexes are described in (*18*); however, Illumina will provide transposase complexes in response to reasonable requests from the scientific community subject to a material transfer agreement. Some work in this study is related to technology described in patent applications WO2014142850, 2014/0194324, 2010/0120098, 2011/0287435, 2013/0196860, and 2012/0208705.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/910/suppl/DC1
Materials and Methods
Figs. S1 to S22
Tables S1 and S2
References (*24*–*39*)

19 March 2015; accepted 24 April 2015
Published online 7 May 2015;
10.1126/science.aab1601

VIROLOGY

A virus that infects a hyperthermophile encapsidates A-form DNA

Frank DiMaio,[1]* Xiong Yu,[2]* Elena Rensen,[3] Mart Krupovic,[3] David Prangishvili,[3]† Edward H. Egelman[2]†

Extremophiles, microorganisms thriving in extreme environmental conditions, must have proteins and nucleic acids that are stable at extremes of temperature and pH. The nonenveloped, rod-shaped virus SIRV2 (*Sulfolobus islandicus* rod-shaped virus 2) infects the hyperthermophilic acidophile *Sulfolobus islandicus*, which lives at 80°C and pH 3. We have used cryo–electron microscopy to generate a three-dimensional reconstruction of the SIRV2 virion at ~4 angstrom resolution, which revealed a previously unknown form of virion organization. Although almost half of the capsid protein is unstructured in solution, this unstructured region folds in the virion into a single extended α helix that wraps around the DNA. The DNA is entirely in the A-form, which suggests a common mechanism with bacterial spores for protecting DNA in the most adverse environments.

Extreme geothermal environments, with temperatures above 80°C, are the habitat of hyperthermophilic DNA viruses that parasitize Archaea (*1*). These viruses have more than 92% of genes without homologs in databases (*2*, *3*), distinct protein folds (*4*), and distinct mechanisms of viral egress (*5*). The high diversity of virion morphotypes may underpin virion morphogenesis and DNA packaging, which could determine the high stability of the virions. Viruses from the family *Rudiviridae* (*6*) consist of a nonenveloped, helically arranged nucleoprotein composed of double-stranded DNA (dsDNA) and thousands of copies of a 134-residue protein. To understand the mechanisms stabilizing rudiviral DNA in natural habitats of host cells, which involve high temperatures (~80°C) and low pH values (~pH 3), we used cryo–electron microscopy (cryo-EM) to analyze the rudivirus SIRV2 (*Sulfolobus islandicus* rod-shaped virus 2) (*6*), which infects the hyperthermophilic acidophilic archaeon *Sulfolobus islandicus* (*7*) (see supplementary materials and methods). Members of the archaeal genus *Sulfolobus* maintain their cytoplasmic pH neutral at pH 5.6 to 6.5 (*8*, *9*). SIRV2 is therefore exposed to a wide range of pH values: from about pH 6 in the cellular cytoplasm, where it assembles and maturates (*10*), to pH 2 to 3 in the extracellular environment. We performed our studies at pH 6. SIRV2 is stable over a wide range of temperatures: from −80°C, the temperature at which the virus can be stored for years without loss of infectivity, to 80°C, the temperature of its natural environment. The overall morphology of the virion is maintained, regardless of the use of negative-stain imaging at 75°C (*11*) or cryo-EM with a sample at 4°C before vitrification (Fig. 1A).

Electron cryo-micrographs of SIRV2 (Fig. 1A) showed strong helical striations in most of the virions with a periodicity of 42 Å. We performed three-dimensional (3D) reconstruction using the iterative helical real space reconstruction method (*12*), after first determining the helical symmetry. Only one solution (with 14.67 subunits per turn of the 42 Å pitch helix) yielded a reconstruction with recognizable secondary structure, almost all α helical, and a resolution of ~3.8 Å in the more-ordered interior, which surrounds the DNA (fig. S2). The asymmetric unit was a symmetrical dimer, the α helices of which were wrapping around a continuous dsDNA. The DNA

[1]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA. [2]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908, USA. [3]Institut Pasteur, Department of Microbiology, 25 rue du Dr. Roux, Paris 75015, France.
*These authors contributed equally to this work. **†Corresponding author. E-mail: egelman@virginia.edu (E.H.E.); david.prangishvili@pasteur.fr (D.P.)**

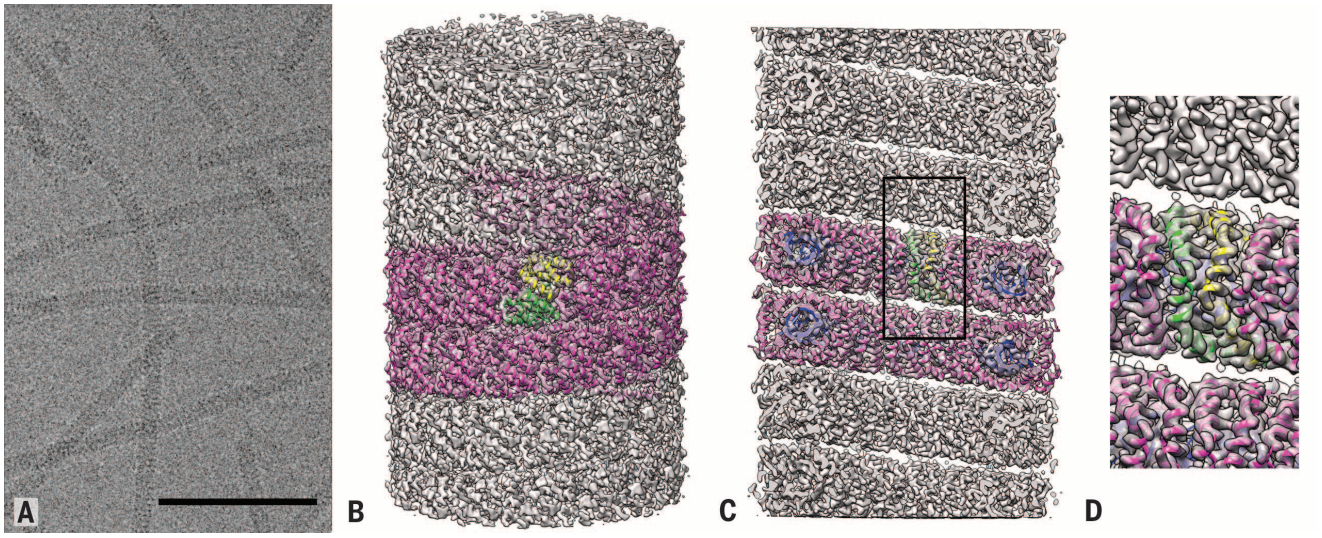


Fig. 1. Cryo-EM and 3D reconstruction of SIRV2. (**A**) Micrograph showing SIRV2 virions in vitreous ice. Scale bar, 1000 Å. (**B**) Side view of the reconstructed virion with a ribbon model for the protein (magenta). The asymmetric unit in the virion contains a protein dimer, and one is shown with one chain in yellow and the other in green. (**C**) Cutaway view showing the hollow lumen with the all α-helical protein segments that line the lumen. These α helices wrap around the dsDNA (blue) and encapsidate it. (**D**) Close-up view of the region shown within the rectangle in (C).

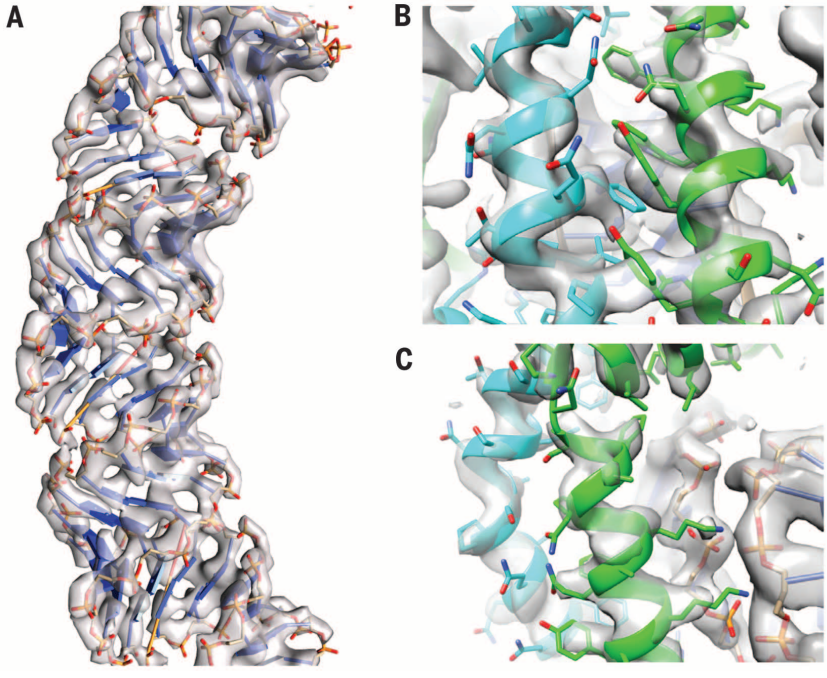


Fig. 2. The refined atomic model in the 3D reconstruction. (**A**) The density allows accurate tracing of the nucleotide chain. (**B**) The interface between helices within the asymmetric unit features large numbers of aromatic residues, which aid in correct registration of the sequence. (**C**) Side chains at the protein-DNA interface are well defined in density.

was in an A-form, in contrast to the B-form DNA (B-DNA) observed in icosahedral bacteriophages (*13*–*15*).

We used Rosetta guided by restraints from the EM data to determine and refine the atomic structure of SIRV2 (*16*). We began by docking a model of A-form DNA (A-DNA) into the map. The resulting model showed good agreement with the experimental data, where the rigid phosphate groups were well defined (Fig. 2, A and C). A dimeric crystal structure from a close homolog with 88% identity [Protein Data Bank identifier (PDB ID) 3F2E] was docked into the map. However, this model lacked 51 N-terminal residues, the first 46 of which were shown by nuclear magnetic resonance (NMR) to be unstructured in the monomer in solution and not part of the fragment crystallized (*17*). In the cryo-EM reconstruction, these residues formed helices wrapping around the DNA. Because the NMR studies were done at pH 6, the same pH used for our cryo-EM studies, the gain in structure of these residues is associated with assembly rather than a change in pH. We used RosettaCM (*18*) to build these N-terminal residues into the density map. A representative model was chosen from a well-converged, low-energy ensemble (fig. S4); this model shows good agreement with the side-chain density in the map at both protein-protein interfaces (Fig. 2B) and protein-DNA interfaces (Fig. 2C). Seven N-terminal residues could not be placed in the density.

The final structure reveals that the N-terminal residues form a helix-turn-helix motif encapsulating the A-DNA, with helices from each subunit in the asymmetric unit packing in an antiparallel configuration (Fig. 3, A and B). Proline residues in this region (Pro^{27}, Pro^{39}, and Pro^{50}) allow some helical deformation to tightly wrap the DNA. These interdigitated helices form a solvent-inaccessible surface surrounding the DNA (Fig. 3C). The DNA was confirmed as A-form (*19*), where the parameters [including an average base pair inclination of 25°, a negative slide (average = –1.6), and a negative x-displacement (average = –4.8)] match A-DNA, whereas the slide and x-displacement are positive for B-DNA. The average phosphate-phosphate distance along the DNA backbone is 5.9 Å, as opposed to ~7.0 Å for B-DNA. The diameter of the DNA is ~24 Å. At this resolution, the sugar pucker is not discernable. A slight bulge in the DNA occurs near the dimer interface, where a buried arginine side chain (Arg^{73}) interacts with one of the DNA phosphate groups, leading to some local deviations from A-DNA. Model bias was tested by starting with the B-form, which converged to the same final model (fig. S5). The DNA (Fig. 3D), whose axis is at a helical radius of ~60 Å, has three right-handed superhelical turns every 44 (= 3 × 14.67) repeats (turns) of the DNA. So there are 528 base pairs

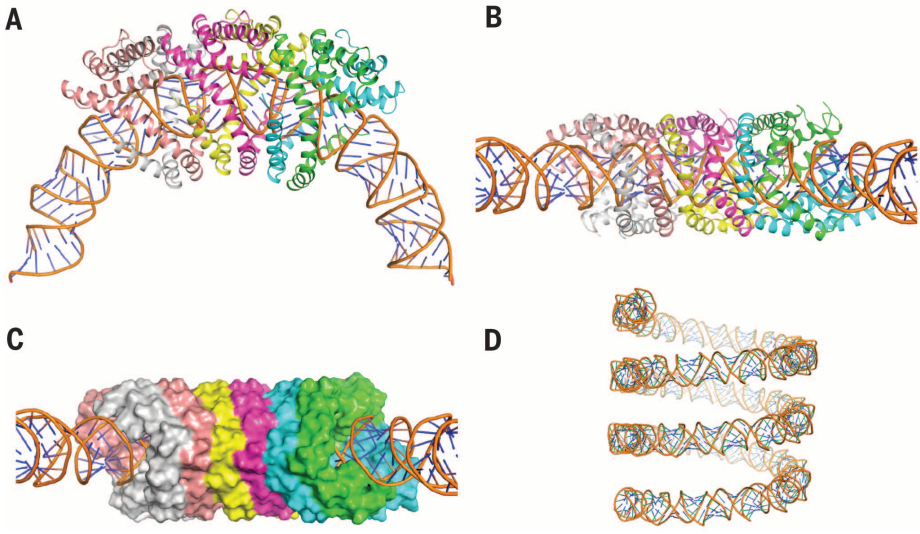
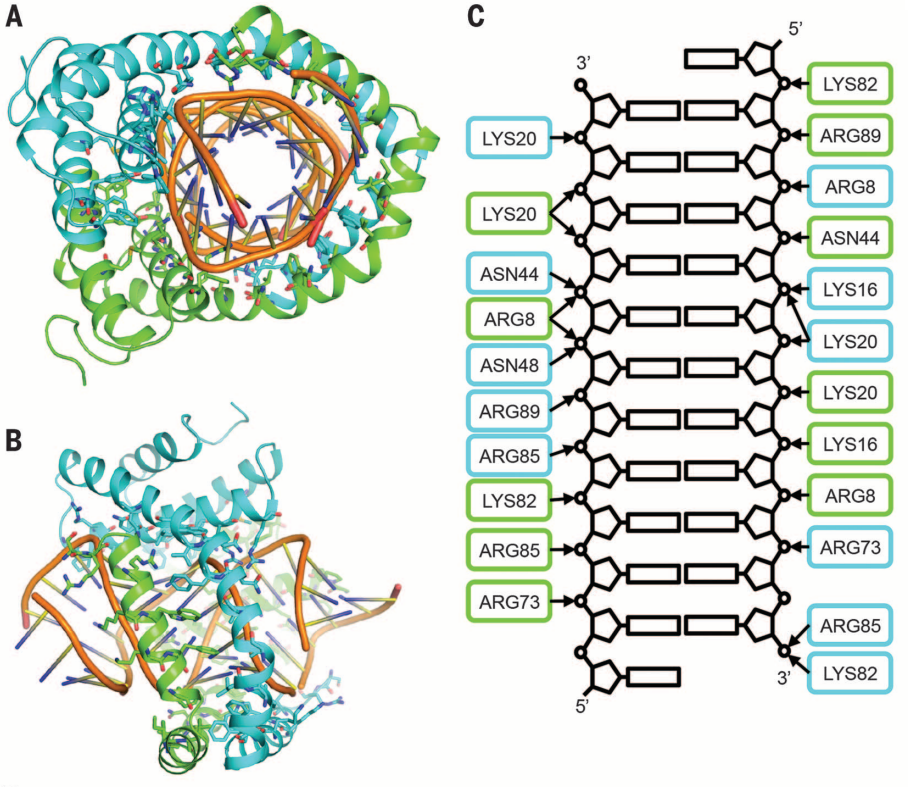


Fig. 3. The SIRV2 protein dimer helices fully encapsulate the DNA. (**A**) Three asymmetric units of the virion are shown, illustrating how the N-terminal helices wrap around the DNA, forming antiparallel helix-helix packing. (**B**) Side view. (**C**) Surface view of the protein (using a 1.4 Å probe radius). (**D**) The right-handed solenoidal supercoiling of the DNA, with three turns shown.



D

```
                10        20        40        50        70        80        90
SIRV2   MAKGHTSRSYSQRYAKWQAKFNA … VAQQNFQTNVPKF … GPNRAIYQGFGLKVARALNRIGS
SIRV1   MAKGHTKRSYSQRYAKWQAKFNA … VAQQNFQTNVPTF … GPNRAIYQGFGLKIARALNRIGS
ARV1    MAKGHTPRSYSQRYSKWSGKFTA … IAQENFQTNVPKF … GPSRAIYQGFGLKVERALNRIGS
SRV     MAKGRTPRSFTQRYAKWNAKFTA … IAQQNFQTNVPKF … GPNRAIYQGYALKVARALNRLGG
SMRV1   MAKGHTPRSFAQRYGKWQAKFTA … VAQGNFEQNSSLM … GPNRVIYQGFALKLYRAKNRIGN
```

Fig. 4. Overview of protein-DNA contacts in the virion. (**A**) The protein-DNA interface is mainly polar, with largely Arg and Lys side chains contacting phosphate groups in the DNA. (**B**) Side view. (**C**) Schematic indicating all of the polar protein-DNA contacts. The coloring of each subunit is the same as in (A) and (B). (**D**) A multiple sequence alignment with related archaeal rod-shaped viruses indicates that all of these contacts are well conserved (*29*). SIRV2, *Sulfolobus islandicus* rod-shaped virus 2; ARV1, *Acidianus* rod-shaped virus 1; SRV, *Stygiolobus* rod-shaped virus; SMRV1, *Sulfolobales* Mexican rudivirus 1. The yellow boxes indicate the residues that are interacting with the DNA backbones.

(= 44 × 12) per 47 right-handed turns, which yields an overall twist of 11.2 base pairs per turn (Fig. 3D).

DNA-protein contacts (Fig. 4, A to C) are largely polar, with nine conserved side chains (Arg^{8}, Lys^{16}, Lys^{20}, Asn^{44}, Asn^{48}, Arg^{73}, Lys^{82}, Arg^{85}, and Arg^{89}) directly interacting with the DNA phosphate groups. Additionally, the backbone of Ser^{9} makes contact with the phosphate backbone of the DNA. There are also several hydrophobic contacts with the DNA, most notably the aromatic residues Trp^{17}, Phe^{21}, Phe^{24}, and Phe^{52}, as well as Val^{37}. All of these residues are conserved in related rudiviruses (Fig. 4D), suggesting a similar method of DNA stabilization and protection. Extensive protein-DNA interactions in SIRV2 virions alleviate the necessity to package charge-neutralizing counterions (*13*, *14*). Within the dimer, protein-protein interfaces are largely hydrophobic. The interface is extensive, comprising ~17% of each monomer's surface area, with a total interface area of 1491 $Å^2$. Five aromatic residues (Tyr^{10}, Tyr^{14}, Trp^{17}, Phe^{21}, and Phe^{45}) form a well-packed barrier separating DNA from solvent. The protein-protein interface between adjacent dimers also forms a largely hydrophobic interface, with 1706 $Å^2$ of contact area between dimers on both sides. Interactions between dimers across the groove of the helix are weak and largely polar.

Small acid-soluble proteins (SASPs) are responsible for protecting DNA in Gram-positive bacterial spores (*20*) and are largely unstructured in solution (*21*–*24*), but they become α helical upon binding dsDNA (*25*, *26*). Almost half of the SIRV2 capsid protein is unstructured in solution (*17*), and this portion becomes α helical when bound to DNA in the virion. The binding of SASPs to DNA is saturable, with saturation occurring at an SASP:DNA weight ratio of ~3:1 (*27*). In the SIRV2 virion, we have now shown that the weight ratio of capsid protein to DNA is 3.5:1. The binding of the SASPs to DNA induces a dimerization of the SASPs (*25*), and the asymmetric unit in the virion contains a symmetrical dimer of the coat protein. The binding of SASPs to DNA induces

a transition from B-DNA to A-DNA (*27*), and an en masse transition of DNA from B-form to A-form can be induced in bacterial cells, suggesting that the A-form plays an unrecognized role in stabilizing DNA under adverse conditions such as dessication (*28*). Sequence analysis and structural comparison (with PDB ID 2Z3X) do not show obvious homology, suggesting that these similarities could have arisen as a result of convergent evolution.

REFERENCES AND NOTES

1. D. Prangishvili, *Annu. Rev. Microbiol.* **67**, 565–585 (2013).
2. D. Prangishvili, R. A. Garrett, E. V. Koonin, *Virus Res.* **117**, 52–67 (2006).
3. M. Krupovic, D. Prangishvili, R. W. Hendrix, D. H. Bamford, *Microbiol. Mol. Biol. Rev.* **75**, 610–635 (2011).
4. M. Krupovic, M. F. White, P. Forterre, D. Prangishvili, *Adv. Virus Res.* **82**, 33–62 (2012).
5. D. Prangishvili, T. E. Quax, *Curr. Opin. Microbiol.* **14**, 315–320 (2011).
6. D. Prangishvili *et al.*, *Genetics* **152**, 1387–1396 (1999).
7. C. Jaubert *et al.*, *Open Biol.* **3**, 130010 (2013).
8. M. Lübben, G. Schäfer, *J. Bacteriol.* **171**, 6106–6116 (1989).
9. C. Baker-Austin, M. Dopson, *Trends Microbiol.* **15**, 165–171 (2007).
10. A. Bize *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 11306–11311 (2009).
11. E. R. Quemin *et al.*, *J. Virol.* **87**, 13379–13385 (2013).
12. E. H. Egelman, *Ultramicroscopy* **85**, 225–234 (2000).
13. T. Y. Yu, J. Schaefer, *J. Mol. Biol.* **382**, 1031–1042 (2008).
14. S. A. Overman *et al.*, *Biospectroscopy* **4** (suppl.), S47–S56 (1998).
15. W. Earnshaw, S. Casjens, S. C. Harrison, *J. Mol. Biol.* **104**, 387–410 (1976).
16. F. S. DiMaio *et al.*, *Nat. Methods*, **12**, 361–365 (2015).
17. B. R. Szymczyna *et al.*, *Structure* **17**, 499–507 (2009).
18. Y. Song *et al.*, *Structure* **21**, 1735–1742 (2013).
19. X. J. Lu, W. K. Olson, *Nucleic Acids Res.* **31**, 5108–5121 (2003).
20. P. Setlow, *Trends Microbiol.* **15**, 172–180 (2007).
21. B. Setlow, P. Setlow, *J. Bacteriol.* **177**, 4149–4151 (1995).
22. C. S. Hayes, P. Setlow, *J. Bacteriol.* **179**, 6020–6027 (1997).
23. C. S. Hayes, B. Illades-Aguiar, L. Casillas-Martinez, P. Setlow, *J. Bacteriol.* **180**, 2694–2700 (1998).
24. C. S. Hayes, E. Alarcon-Hernandez, P. Setlow, *J. Biol. Chem.* **276**, 2267–2275 (2001).
25. C. S. Hayes, Z. Y. Peng, P. Setlow, *J. Biol. Chem.* **275**, 35040–35050 (2000).
26. K. S. Lee, D. Bumbaca, J. Kosman, P. Setlow, M. J. Jedrzejas, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 2806–2811 (2008).
27. S. C. Mohr, N. V. Sokolov, C. M. He, P. Setlow, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 77–81 (1991).
28. D. R. Whelan *et al.*, *J. R. Soc. Interface* **11**, 20140454 (2014).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

ACKNOWLEDGMENTS

This work was supported by NIH grant GM035269 (to E.H.E.). The map and model have been deposited in the Electron Microscopy Data Bank (accession number EMD-6310) and in PDB (ID 3J9X), respectively. We thank K. Dryden for assistance with the microscopy.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/914/suppl/DC1
Materials and Methods
Figs. S1 to S5
References (*30*–*37*)

3 December 2014; accepted 21 April 2015
10.1126/science.aaa4181

RNA STRUCTURE

Structure of the HIV-1 RNA packaging signal

Sarah C. Keane,[1] Xiao Heng,[1]* Kun Lu,[1]† Siarhei Kharytonchyk,[2] Venkateswaran Ramakrishnan,[1] Gregory Carter,[1] Shawn Barton,[1] Azra Hosic,[1] Alyssa Florwick,[1] Justin Santos,[1] Nicholas C. Bolden,[1] Sayo McCowin,[1] David A. Case,[3] Bruce A. Johnson,[4] Marco Salemi,[5] Alice Telesnitsky,[2]‡ Michael F. Summers[1]‡

The 5′ leader of the HIV-1 genome contains conserved elements that direct selective packaging of the unspliced, dimeric viral RNA into assembling particles. By using a ^{2}H-edited nuclear magnetic resonance (NMR) approach, we determined the structure of a 155-nucleotide region of the leader that is independently capable of directing packaging (core encapsidation signal; Ψ^{CES}). The RNA adopts an unexpected tandem three-way junction structure, in which residues of the major splice donor and translation initiation sites are sequestered by long-range base pairing and guanosines essential for both packaging and high-affinity binding to the cognate Gag protein are exposed in helical junctions. The structure reveals how translation is attenuated, Gag binding promoted, and unspliced dimeric genomes selected, by the RNA conformer that directs packaging.

Assembly of HIV-1 particles is initiated by the cytoplasmic trafficking of two copies of the viral genome and a small number of viral Gag proteins to assembly sites on the plasma membrane (*1*–*6*). Unspliced, dimeric genomes are efficiently selected for packaging from a cellular milieu that includes a substantial excess of nonviral messenger RNAs (mRNAs) and more than 40 spliced viral mRNAs (*7*, *8*). RNA signals that direct packaging are located primarily within the 5′ leader of the genome and are recognized by the nucleocapsid (NC) domains of Gag (*4*). Transcriptional activation, splicing, and translation initiation are also dependent on elements within the 5′ leader, the most conserved region of the genome (*9*), and there is evidence that these and other activities are temporally modulated by dimerization-dependent exposure of functional signals (*6*, *10*–*13*).

Understanding the RNA structures and mechanisms that regulate HIV-1 5′ leader function has its basis in phylogenetic, biochemical, nucleotide reactivity, and mutagenesis studies (*4*). The dimeric leader selected for packaging appears to adopt a highly branched secondary structure, in which there are structurally discrete hairpins and helices that promote transcriptional activation (TAR), transfer RNA (tRNA) primer binding (PBS), packaging (ψ), dimer initiation (DIS), splicing (SD), and dimer stability (U5:AUG) (*4*, *14*) (Fig. 1). Although nuclear magnetic resonance (NMR) signals diagnostic of TAR, PBS, ψ, DIS, U5:AUG, and polyadenylate [poly(A)] helices have been observed in spectra obtained for the full-length dimeric leader (*13*, *15*) (Fig. 1A), signals diagnostic of a putative SD hairpin have not been detected (colored magenta in Fig. 1A) (*15*), and there is little agreement among more than 20 different structure predictions for residues adjacent to the helices (*4*). For example, predictions vary for stretches of residues shown by in vivo nucleotide reactivity (*16*) and cross-linking with immunoprecipitation (*17*) to reside at or near sites of Gag binding (*4*). The TAR, poly(A), and PBS hairpins of the HIV-1 leader are not required for efficient encapsidation (*15*), and a minimal HIV-1 packaging element, the core encapsidation signal (Ψ^{CES}), exhibits NC binding properties and NMR spectral features similar to those of the intact 5′ leader and is independently capable of directing vector RNAs into viruslike particles (*15*). To gain insights into the mechanism of HIV-1 genome selection, we determined the structure of Ψ^{CES} by NMR.

Contributions of slow molecular rotational motion to NMR relaxation were minimized by substituting the dimer-promoting GC-rich loop of the Ψ^{CES} DIS hairpin by a GAGA tetraloop (Fig. 1A). This prevented dimerization (Fig. 1B) but did not affect NC binding (Fig. 1C) or nuclear Overhauser effect spectroscopy (NOESY) NMR spectral patterns (*18*), indicating that the modified RNA retains

[1]Howard Hughes Medical Institute (HHMI) and Department of Chemistry and Biochemistry, University of Maryland Baltimore County (UMBC), 1000 Hilltop Circle, Baltimore, MD 21250, USA. [2]Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI 48109-5620, USA. [3]Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ 08854, USA. [4]One Moon Scientific, Incorporated, 839 Grant Avenue, Westfield, NJ 07090, USA, and City University of New York (CUNY) Advanced Science Research Center, 85 St. Nicholas Terrace, New York, NY 10031, USA. [5]Department of Pathology, Immunology, and Laboratory Medicine, College of Medicine, and Emerging Pathogens Institute, University of Florida, Gainesville, FL 32610, USA.
*Present address: Department of Biochemistry, University of Missouri, Columbia, MO 65211, USA. †Present address: Regeneron Pharmaceuticals, Incorporated, 777 Old Saw Mill River Road, Tarrytown, NY 10591, USA. **‡Corresponding author. E-mail: summers@hhmi.umbc.edu (M.F.S.); ateles@umich.edu (A.T.)**

the structure of the native dimer. Nonexchangeable aromatic and ribose H_1', H_2', and H_3' ^{1}H NMR signals were assigned for nucleotides of the U5:AUG, lower-PBS, DIS, and ψ helices by sequential residue analysis of two-dimensional (2D) NOESY spectra obtained for nucleotide-specific ^{2}H-labeled samples (*18*–*20*) (Fig. 1D). Very-long-range A-H_2 NOEs (^{1}H-^{1}H distances up to ~7 Å) were detected in spectra of highly deuterated samples (Fig. 1E) [as observed for proteins (*21*)], facilitating assignments.

NMR signals that could not be assigned by nucleotide-specific labeling were identified by a fragmentation-based segmental ^{2}H-labeling approach that we developed, in which differentially labeled 5′ and 3′ fragments of Ψ^{CES} were prepared separately and noncovalently annealed (Fig. 2, A and B, and fig. S1). The dimer-promoting loop of the DIS hairpin served as the fragmentation site and was substituted by a short stretch of intermolecular G:C base pairs (Fig. 2A). Differential ^{2}H labeling afforded the following fragment-annealed RNAs (fr-Ψ^{CES}; denoted 5′ fragment:3′ fragment-Ψ^{CES}; D, perdeuterated fragment; superscripts denote sites of protonation, all other sites deuterated; e.g., G, fully protonated guanosines, A^{2r}, adenosines protonated at C_2 and ribose carbons): A^{2r}:U^r-Ψ^{CES}, $A^{2r}C^r$:U^r-Ψ^{CES} G^r:$A^{2r}C^r$-Ψ^{CES}, D:$A^{2r}C^r$-Ψ^{CES}, A:D-Ψ^{CES}, and D:A-Ψ^{CES} (fig. S1). Except for residues at the sites of substitution, the NMR spectra of the fr-Ψ^{CES} RNAs were consistent with those of the parent, nonfragmented RNA. For example, NOEs that correlate A124-H_2 with cytosine and uridine H_1' protons in 2D NOESY spectra obtained for nonfragmented $A^{2r}C^r$-Ψ^{CES}, $A^{2r}U^r$-Ψ^{CES}, and $A^{2r}C^rU^r$-Ψ^{CES} samples were also detected in spectra obtained for fragment-annealed A^{2r}:U^r-Ψ^{CES} and $A^{2r}C^r$:U^r-Ψ^{CES} constructs, indicating that A124 resides near a cytosine (C125) in the 5′ fragment and a uridine (U295) in the 3′ fragment (Fig. 2C). More than 80 long-range and sequential A-H_2 NOEs were identified by using the ^{2}H-edited NMR approach (Fig. 2E). The ^{1}H NMR assignments were validated by NOE cross peak pattern redundancy and database chemical shift analyses (*18*, *22*) (fig. S2).

The NMR data indicate that residues proximal to the major splice donor site do not form a hairpin but instead participate in long-range base pairing within an extended DIS stem and a short helical segment, H1 (Fig. 2E). To determine whether this secondary structure is also adopted by the native 5′ leader, we obtained NOESY data for dimeric, ^{2}H-labeled 5′ leader constructs. Adenosine-H_2 signals diagnostic of the U5:AUG, DIS, PBS, and Ψ helices were observed in spectra obtained for the native leader ([5′-L]$_2$), as expected (*15*). However, signals diagnostic of H1 were only detectable upon removal of the upper PBS loop (substituted by a GAGA tetraloop; [5′-$L^{\Delta PBS}$]$_2$), which eliminated broad upper PBS signals that overlapped with the A124-H_2 signal of H1 (Fig. 2D). This construct exhibits dimerization, NC binding, and NMR properties similar to those of the intact leader (*15*) and directs both noncompetitive (*15*) and competitive RNA packaging with near–wild-type efficiency [94 ± 4% and 93 ± 18%, respectively (reported as mean ± standard deviation)] (Fig. 2F). Thus, the secondary structure observed for Ψ^{CES}, including the H1 helix, is also adopted by the 5′ leader.

NOE-restrained structure calculations (*18*) reveal that Ψ^{CES} adopts a tandem three-way junction structure (Fig. 3, A to C, and fig. S3). The overall shape is quasi-tetrahedral, with the U5:AUG, H1, and ψ helices forming a plane that is nearly perpendicular to the plane formed by the H1, PBS, and DIS helices (Fig. 3A). Splice-site residues G289 and G290 are base-paired with C229 and U228, respectively; adjacent residues are base-paired within or near the H1-PBS-DIS (three-way-2) junction (Fig. 3, B to D); and residues of AUG are base-paired within the U5:AUG-H1-Ψ (three-way-1) junction (Fig. 3, B and D). A227 to U291 forms an extended DIS hairpin with two internally stacked but nonpaired guanosines (G272 and G273) and a G240(syn):G278(anti)-G241(anti) base triple. Sequentially stacked pyrimidines (U230*U288 and C231*C287) exhibit broad line widths indicative of millisecond time scale conformational exchange (Fig. 3E). These residues appear to function as a flexible hinge that connects the extended DIS hairpin with the tandem three-way junction (Fig. 3D). U307 to G330 forms an extended ψ-hairpin structure that contains three noncanonical base pairs [G310(anti)*A327(anti), G328*U309, G329*U308], a stacked A-A bulge

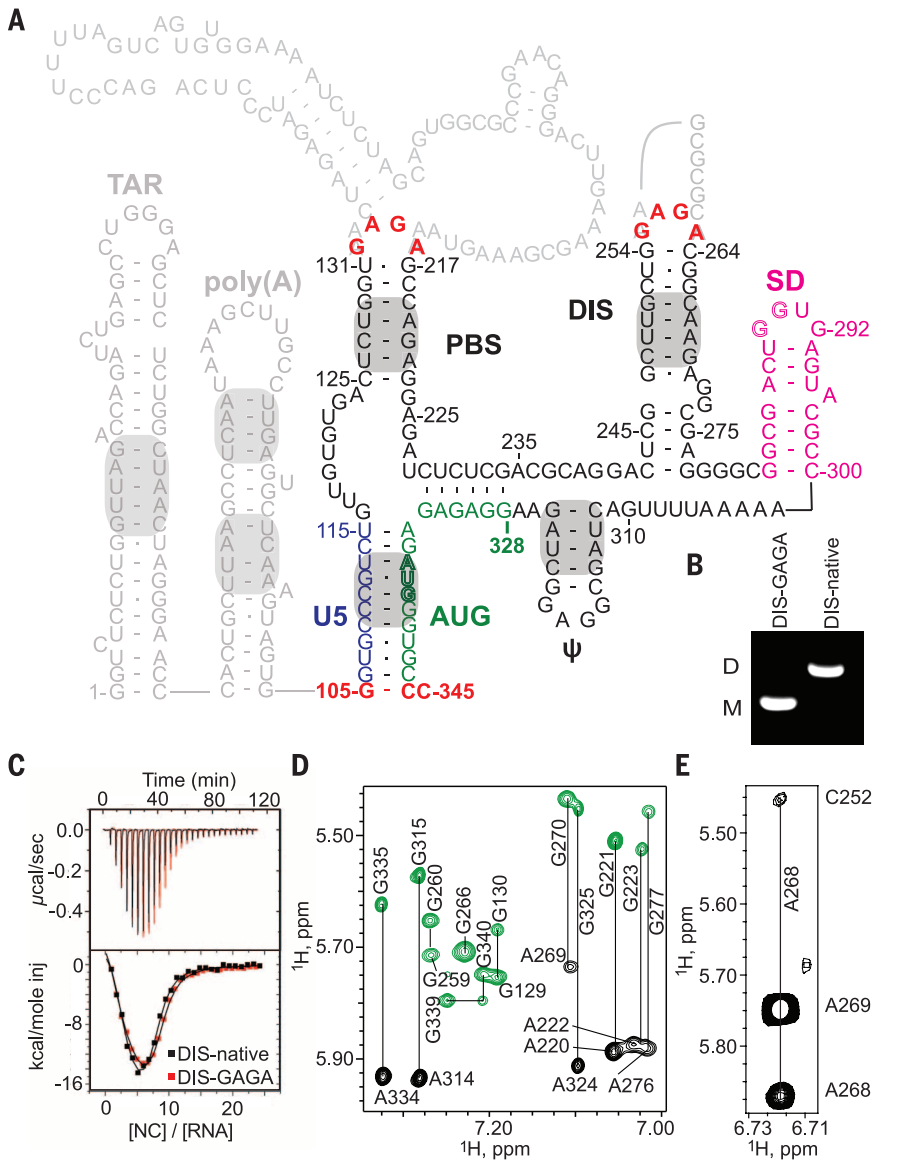


Fig. 1. HIV-1 5′ leader and Ψ^{CES} RNA construct. (**A**) Predicted secondary structure of the HIV-1 (NL4-3 strain) 5′ leader (*16*); gray shading denotes elements detected in the intact leader by NMR (*13*, *15*); dark letters denote Ψ^{CES} (nonnative residues colored red; see text). (**B** and **C**) Substitution of the native DIS loop residues (DIS-native) by GAGA (DIS-GAGA) prevents dimerization (B) but does not affect NC binding (C). ppm, parts per million. (**D**) Representative NOESY spectra for G^8A-Ψ^{CES} (black) and G-Ψ^{CES} (green); lines connect H_8 (vertical labels) and H_1' (horizontal labels) signals. (**E**) Representative very-long-range NOE (A268-H_2 to C252-H_1'; ~7 Å separation) obtained for $A^{2r}C^r$-Ψ^{CES}.

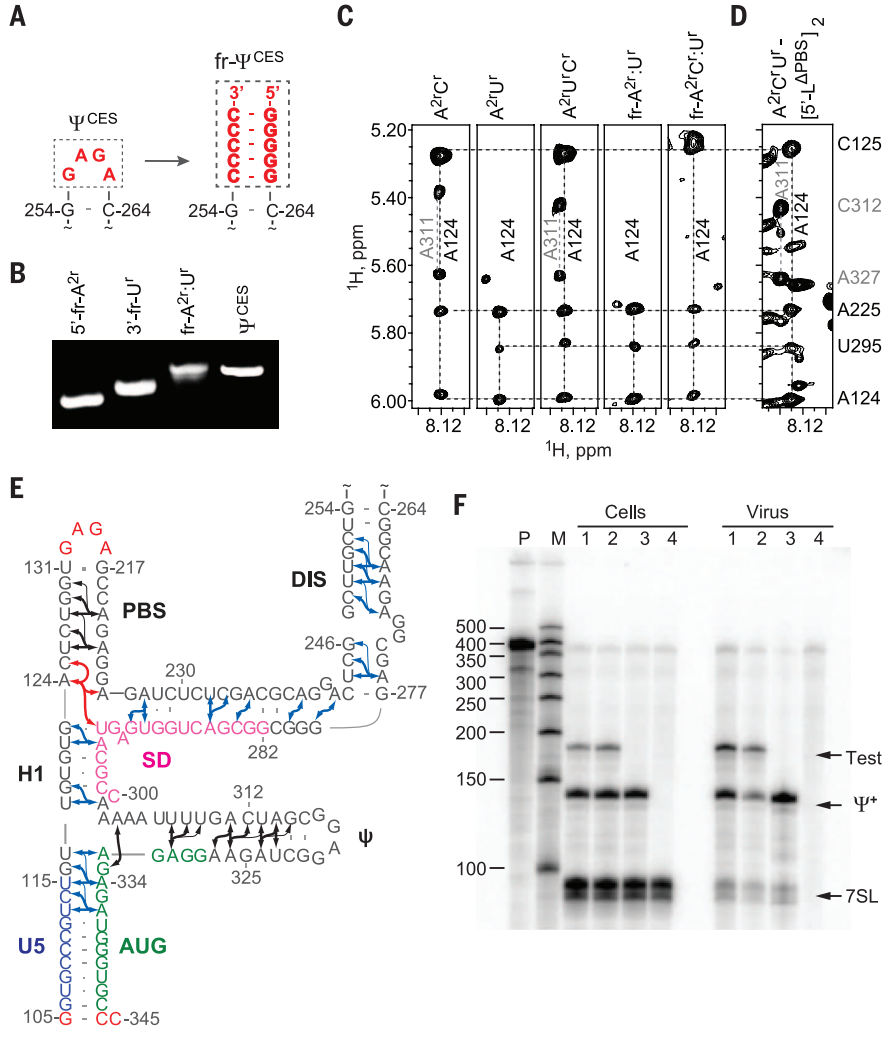


Fig. 2. Fragmentation-based ^{2}H-edited NMR approach and observed Ψ^{CES} secondary structure. (**A**) The DIS loop of Ψ^{CES} served as the fragmentation site and was substituted by a stretch of intermolecular G-C base pairs. (**B**) Fragment-annealing efficiency as measured by native agarose gel electrophoresis. (**C**) The 2D NOESY spectra of uniformly labeled $A^{2r}C^{r}$-, $A^{2r}U^{r}$-, and $A^{2r}C^{r}U^{r}$-Ψ^{CES} and segmentally labeled fr-A^{2r}:U^{r}- and fr-$A^{2r}C^{r}$:U^{r}-Ψ^{CES} samples used to make long-range NOE assignments. (**D**) Similarities in NOESY spectra obtained for $A^{2r}C^{r}U^{r}$-labeled $[5'\text{-}L^{\Delta PBS}]_2$ and Ψ^{CES} confirm that the tandem three-way junction structure is present in both constructs. (**E**) NMR-derived secondary structure of Ψ^{CES}. Black and blue arrows denote A-H_2 NOEs observable in Ψ^{CES} and fr-Ψ^{CES} samples, respectively; red arrows highlight NOEs shown in (C) and (D); thin arrows denote very-long-range NOEs. (**F**) Packaging of native HIV-1_{NL4-3} 5′-L and 5′-$L^{\Delta PBS}$ RNAs under competition conditions assayed by means of ribonuclease protection. P, undigested probe; M, RNA sizes marker. Lanes 1 and 2 show native HIV-1_{NL4-3} helper versus test vectors containing 5′-$L^{\Delta PBS}$ (*1*) or native HIV-1_{NL4-3} (*2*). Lane 3 contains HIV-1_{NL4-3} helper expressed without test RNA. Lane 4 is mock transfected cells. Samples obtained from transfected cells (Cells) or viral-containing media (Virus) are indicated. Bands corresponding to host 7SL RNA, HIV-1_{NL4-3} helper RNA (Ψ^{+}), and copackaged test RNAs (Test) are labeled.

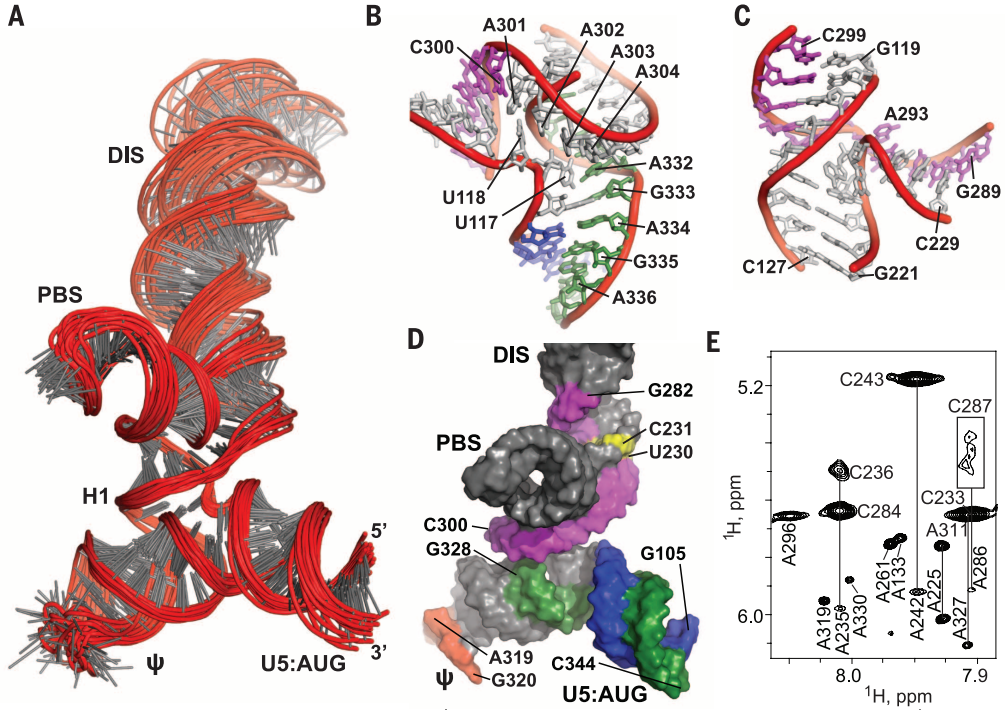


Fig. 3. Three-dimensional structure of Ψ^{CES}. (**A**) Ensemble of 20 refined structures (residues 105 to 344 shown). (**B** and **C**) Expanded views of the (B) three-way-1 and (C) three-way-2 junctions. (**D**) Surface representation of Ψ^{CES} highlighting U5 (blue):AUG (green) base pairing and the integral participation of SD residues (pink) in the tandem three-way junction structure. (**E**) Severe line broadening indicative of slow (millisecond) conformational averaging was observed for stacked, mismatched pyrimidines in the extended DIS stem [yellow in (D); broadened C287-H_1' signal boxed in (E)]. NOE patterns and sharp NMR signals also indicate that the ψ hairpin loop is unstructured [red in (D)].

[A311(anti)-A326(anti)] (Fig. 2E), and a flexible GAGG loop (Fig. 3D). Adenosines A302 to A305 exhibit pseudo A-form stacking but are not base-paired (Fig. 3B), which supports proposals that genomic adenosine enrichment occurs primarily at non–base-paired sites (*23*). A302 and A303 also make A-minor contacts with the U5:AUG helix (Fig. 3B).

To determine whether the tandem three-way junction is evolutionarily conserved, we analyzed published HIV-1 leader sequences that contained full coverage of the 5′ untranslated region (278 total sequences). Representatives from B, C, and F1 subtypes were included in the analysis (*18*). Of the 48 base-paired nucleotides at or near the three-way junction, 31 were either strictly (16 sites) or very highly (>99%, 15 sites) conserved, and 13 displayed high (90.2% to 98.9%) identity (table S2). Only 11 of 126 substitutions resulted in loss of base pairing. The remaining four sites—A^{227}, G^{279}, A^{286}, and U^{288}—exhibited significant variation, ranging from 12% (U^{288}) to 50.3% (A^{227}). Most changes mapped to terminal branches of the Ψ^{CES} phylogeny. Thus, the tandem three-way junction structure is highly conserved, and the rare variations that disrupt base pairing are due to transient polymorphisms.

The PBS, DIS, and ψ helices of Ψ^{CES} are consistent with models derived from nucleotide reactivity experiments (*4*), but the SD structure differs significantly. Recent in-gel chemical probing of resolved monomeric and dimeric leader RNAs (*24*), and probing studies under solution conditions that favor either the monomeric or dimeric species (*25*), showed that SD loop residues are relatively unreactive in the dimeric RNA, consistent with the Ψ^{CES} structure. Pseudo-free energy calculations indicate that the in-gel reactivity data for the dimeric leader (*24*) are in better agreement with the Ψ^{CES} NMR structure than the proposed model [~25% lower experimental pseudo-free energy (*18*); fig. S4]. These findings support proposals that variations in structure predictions are at least partly due to site-specific structural heterogeneity associated with the monomer-dimer equilibrium (*13*, *24*).

HIV-1 NC binds with high affinity to oligonucleotides that contain exposed guanosines (*4*, *26*, *27*). Ψ^{CES} contains five unpaired Gs (excluding the nonnative GAGA tetraloops), a GGG base triple in the DIS stem, and five additional guanosine mismatches clustered at or near the two three-way junctions (G*U, G*A, or G*G) that could serve as NC binding sites (Fig. 4A). Potential contributions of these "junction guanosines" to NC binding were tested by isothermal titration calorimetric studies of G-to-A–substituted Ψ^{CES} RNAs. Free energy calculations indicate that these substitutions, which include three G*U to A-U substitutions, should not alter the secondary structure of the RNA (*18*). Replacement of the ψ GGAG loop by GAAA eliminated one NC binding site, as expected (*27*), and substitution of the three-way-1 junction guanosines by adenosines (G116A/G333A/G328A/G329A/G331A) eliminated three additional NC sites (Fig. 4B). Mutation of the unpaired (G226, G292, and G294) and mismatched (G224) three-way-2 junction guanosines to adenosines eliminated one NC binding site (Fig. 4B). The influence of these guanosines on RNA encapsidation was evaluated by using a competitive in situ RNA packaging assay. Human embryonic kidney 293T cells were co-transfected with plasmids that produce vector RNAs containing the wild-type (Ψ^+, which also encodes for viral proteins) and mutant (Test) leader sequences (*18*). When coexpressed at similar levels, Ψ^+ and Test vector RNAs with native leader sequences were packaged into HIV-1 virus–like particles with similar efficiencies (Fig. 4C). In contrast, significant packaging defects were observed upon G-to-A mutation of the three-way-2 junction guanosines (17% ± 2%), the ψ-loop and three-way-1 junction guanosines (10% ± 2%), and all junction and ψ-loop guanosines (5% ± 1%) (Fig. 4C). Our findings indicate that the tandem three-way junction serves as a scaffold for exposing clusters of unpaired or weakly paired junction guanosines, thereby enabling their binding to the zinc knuckle domains of NC.

The Ψ^{CES} structure explains biochemical, nucleotide reactivity, and phylogenetic results and suggests a mechanism by which the 5′ leader structure regulates translation and splicing (*4*). In vitro translational activity and chemical reactivity of the AUG residues are suppressed upon dimerization (*28*), and this can be attributed to sequestration of the 5′ end of the *gag* open reading frame within the three-way-1 junction (Fig. 3D). Enhanced in vitro translational activity caused by mutations immediately downstream of the major splice donor site (ΔA296/A301U and A293C/U295C/ΔG298) can be explained by

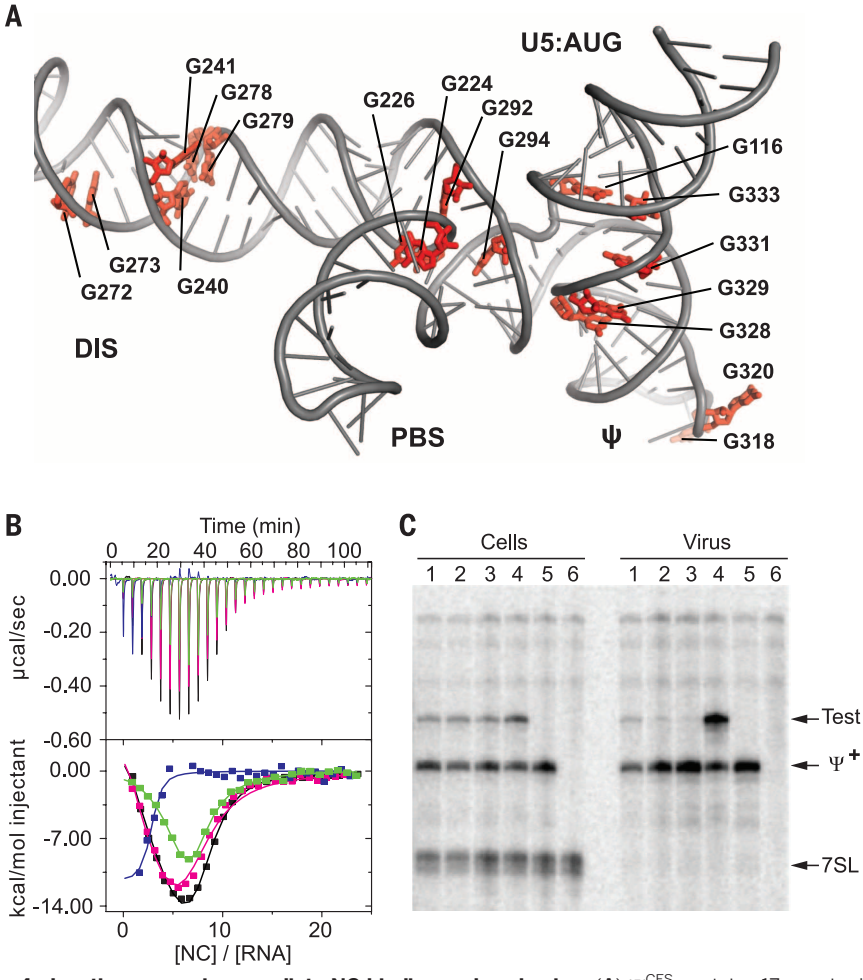


Fig. 4. Junction guanosines mediate NC binding and packaging. (**A**) Ψ^{CES} contains 17 unpaired or weakly paired guanosines (red) that serve as potential NC binding sites. (**B**) Mutation of the three-way-2 (green) or ψ (magenta) guanosines to adenosines modestly reduces NC binding (N = 7.0 ± 0.3 and 7.0 ± 0.5, respectively) relative to wild-type Ψ^{CES} (black; N = 8.0 ± 0.3). Mutation of ψ and three-way-1 guanosines to adenosines (blue) severely inhibits high-affinity NC binding (N = 2.0 ± 0.1). (**C**) Competitive packaging of HIV-1_{NL4-3} vectors containing native and mutant 5′ leader sequences, assayed by means of ribonuclease protection. Lanes 1 to 4 are native HIV-1_{NL4-3} versus test vectors containing 5′-$L^{3way2\text{-}G/A}$ (*1*), 5′-$L^{3way1\text{-}G/A}$ (*2*), 5′-$L^{3way1,2\text{-}G/A}$ (*3*), and 5′-L (*4*). Lane 5 is HIV-1_{NL4-3} helper expressed without test RNA. Lane 6 is mock transfected cells. Samples obtained from transfected cells (Cells) or viral containing media (Virus) are indicated. Bands corresponding to host 7SL RNA, HIV-1_{NL4-3} helper RNA (Ψ^+), and copackaged test RNAs (Test) are labeled.

destabilization of the H1 helix and, for ΔA296/A301U, stabilization of the SD hairpin (*29*), both of which should favor the monomer. Mutations in AUG that inhibit genome dimerization and suppress packaging (*30*, *31*) are expected to disrupt base pairing in the U5:AUG helix and ψ-hairpin stem, thereby destabilizing the tandem three-way junction structure required for Gag binding. In vitro splicing activity is also attenuated by dimerization (*12*, *32*), and this can be attributed to sequestration of the major splice-site recognition sequence within the three-way-2 junction. Antisense oligonucleotides with complementarity to the SD loop inhibit dimerization, and this is likely due to their ability to competitively block formation of the tandem three-way junction (*25*).

The Ψ^{CES} structure also explains the exquisite selectivity of HIV-1 to package its unspliced genome (*1*, *2*). Residues immediately downstream of the major splice site are base-paired within the H1 helix and are thus integral to the formation of the tandem three-way junction structure. Although unspliced and spliced HIV-1 mRNAs contain identical 5′ sequences (G1 to G289), differences in spliced mRNA sequences derived from 3′ exons would preclude formation of the packaging competent junction structure. Similarly, because SD appears to exist as a hairpin in the monomeric, unspliced 5′ leader (*12*), it is likely that monomeric genomes are also ignored during virus assembly because they do not adopt the tandem three-way junction structure.

Compared with the proteins of HIV-1, structural information for the viral nucleic acids is sparse. RNAs in general are vastly underrepresented in the structural data banks (99,000 proteins versus 2700 RNA structures), partly because of NMR technical challenges and difficulties obtaining suitable crystals for x-ray diffraction (*19*, *20*). The fr-RNA ^{2}H-edited NMR approach enables efficient segmental labeling without requiring enzymatic ligation. Given the ubiquity of hairpin elements that can serve as fragmentation or annealing sites, this method should be generally applicable to modest-sized RNAs (~160 nucleotides).

REFERENCES AND NOTES

1. J. M. Coffin, S. H. Hughes, H. E. Varmus, Eds., *Retroviruses* (Cold Spring Harbor Press, Cold Spring Harbor, NY, 1997).
2. R. Berkowitz, J. Fisher, S. P. Goff, *Curr. Top. Microbiol. Immunol.* **214**, 177–218 (1996).
3. J. Chen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 13535–13540 (2009).
4. K. Lu, X. Heng, M. F. Summers, *J. Mol. Biol.* **410**, 609–633 (2011).
5. N. Jouvenet, S. M. Simon, P. D. Bieniasz, *J. Mol. Biol.* **410**, 501–511 (2011).
6. M. Kuzembayeva, K. Dilley, L. Sardo, W.-S. Hu, *Virology* **454-455**, 362–370 (2014).
7. S. Schwartz, B. K. Felber, D. M. Benko, E. M. Fenyö, G. N. Pavlakis, *J. Virol.* **64**, 2519–2529 (1990).
8. O. A. Nikolaitchik *et al.*, *PLOS Pathog.* **9**, e1003249 (2013).
9. A. M. Lever, *Adv. Pharmacol.* **55**, 1–32 (2007).
10. J.-C. Paillart, M. Shehu-Xhilaga, R. Marquet, J. Mak, *Nat. Rev. Microbiol.* **2**, 461–472 (2004).
11. J. Greatorex, *Retrovirology* **1**, 22 (2004).
12. T. E. M. Abbink, B. Berkhout, *J. Virol.* **82**, 3090–3098 (2008).
13. K. Lu *et al.*, *Science* **334**, 242–245 (2011).
14. T. E. M. Abbink, B. Berkhout, *J. Biol. Chem.* **278**, 11601–11611 (2003).
15. X. Heng *et al.*, *J. Mol. Biol.* **417**, 224–239 (2012).
16. K. A. Wilkinson *et al.*, *PLOS Biol.* **6**, e96 (2008).
17. S. B. Kutluay *et al.*, *Cell* **159**, 1096–1109 (2014).
18. Information on materials and methods is available on *Science* Online.
19. K. Lu, Y. Miyazaki, M. F. Summers, *J. Biomol. NMR* **46**, 113–125 (2010).
20. O. Duss, P. J. Lukavsky, F. H. Allain, *Adv. Exp. Med. Biol.* **992**, 121–144 (2012).
21. L. M. Koharudin, A. M. Bonvin, R. Kaptein, R. Boelens, *J. Magn. Reson.* **163**, 228–235 (2003).
22. S. Barton, X. Heng, B. A. Johnson, M. F. Summers, *J. Biomol. NMR* **55**, 33–46 (2013).
23. F. J. van Hemert, A. C. van der Kuyl, B. Berkhout, *RNA Biol.* **10**, 211–215 (2013).
24. J. C. Kenyon, L. J. Prestwood, S. F. Le Grice, A. M. Lever, *Nucleic Acids Res.* **41**, e174 (2013).
25. J. Deforges, N. Chamond, B. Sargueil, *Biochimie* **94**, 1481–1489 (2012).
26. T. L. South, M. F. Summers, *Protein Sci.* **2**, 3–19 (1993).
27. R. N. De Guzman *et al.*, *Science* **279**, 384–388 (1998).
28. F. Baudin *et al.*, *J. Mol. Biol.* **229**, 382–397 (1993).
29. T. E. Abbink, M. Ooms, P. C. Haasnoot, B. Berkhout, *Biochemistry* **44**, 9058–9066 (2005).
30. D. T. Poon, E. N. Chertova, D. E. Ott, *Virology* **293**, 368–378 (2002).
31. O. Nikolaitchik, T. D. Rhodes, D. Ott, W.-S. Hu, *J. Virol.* **80**, 4691–4697 (2006).
32. J. A. Jablonski, E. Buratti, C. Stuani, M. Caputi, *J. Virol.* **82**, 8038–8050 (2008).

ACKNOWLEDGMENTS

This research was supported by grants from the National Institute of General Medical Sciences (NIGMS, R01 GM42561 to M.F.S. and A.T. and P50 GM 103297 to M.S., B.J., and D.A.C.). S.B., N.C.B., and S.M. were supported by a NIGMS grant for enhancing minority access to research careers (MARC U*STAR 2T34 GM008663), and S.B., J.S., N.C.B., and S.M. were supported by an HHMI undergraduate education grant. We thank the HHMI staff at UMBC for technical assistance and B. Rife (University of Florida) for advice regarding the phylogenetic analysis. The following reagent was obtained through the NIH AIDS Reagent Program, Division of AIDS, National Institute of Allergy and Infectious Diseases, NIH: pNL4-3 from M. Martin. Atomic coordinates have been deposited into the Protein Data Bank with accession code 2N1Q. NMR chemical shifts and restraints have been deposited into the Biological Magnetic Resonance Bank with accession code 25571.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/917/suppl/DC1
Materials and Methods
Figs. S1 to S4
Tables S1 and S2
References (*33*–*52*)

14 February 2015; accepted 22 April 2015
10.1126/science.aaa9266

EVOLUTION

Systematic humanization of yeast genes reveals conserved functions and genetic modularity

Aashiq H. Kachroo,[1] Jon M. Laurent,[1] Christopher M. Yellman,[1] Austin G. Meyer,[1,2] Claus O. Wilke,[1,2,3] Edward M. Marcotte[1,2,4]*

To determine whether genes retain ancestral functions over a billion years of evolution and to identify principles of deep evolutionary divergence, we replaced 414 essential yeast genes with their human orthologs, assaying for complementation of lethal growth defects upon loss of the yeast genes. Nearly half (47%) of the yeast genes could be successfully humanized. Sequence similarity and expression only partly predicted replaceability. Instead, replaceability depended strongly on gene modules: Genes in the same process tended to be similarly replaceable (e.g., sterol biosynthesis) or not (e.g., DNA replication initiation). Simulations confirmed that selection for specific function can maintain replaceability despite extensive sequence divergence. Critical ancestral functions of many essential genes are thus retained in a pathway-specific manner, resilient to drift in sequences, splicing, and protein interfaces.

The ortholog-function conjecture posits that orthologous genes in diverged species perform similar or identical functions (*1*). The conjecture is supported by comparative analyses of gene-expression patterns, genetic interaction maps, and chemogenomic profiling (*2*–*6*), and it is widely used to predict gene function across species. However, even if two genes perform similar functions in different organisms, it may not be possible to replace one for the other, particularly if the organisms are widely diverged. The extent to which deeply divergent orthologs can stand in for each other, and which principles govern such functional equivalence across species, is largely unknown.

In this study, we systematically addressed these questions by replacing a large number of yeast genes with their human orthologs. Humans and the baker's yeast *Saccharomyces cerevisiae* diverged from a common ancestor approximately 1 billion years ago (*7*). They share several thousand

[1]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, TX 78712, USA. [2]Center for Computational Biology and Bioinformatics, University of Texas at Austin, Austin, TX 78712, USA. [3]Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA. [4]Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA.
*Corresponding author. E-mail: marcotte@icmb.utexas.edu

orthologous genes, accounting for more than one-third of the yeast genome (*8*). Yeast and human orthologs tend to be recognizable but often highly diverged; amino acid identity ranges from 9 to 92%, with a genome-wide average of 32%. Although we know of individual examples of human genes capable of replacing their fungal orthologs (*9*–*12*), the extent and specific conditions under which human genes can substitute for their yeast orthologs are generally not known.

We focused on the set of genes essential for yeast cell growth under standard laboratory conditions (*13*, *14*) and for which the yeast-human orthology is 1:1 (i.e., genes without lineage-specific duplicate genes that might mask the effects). Based on the availability of full-length human cDNA recombinant clones (*15*, *16*) and matched yeast strains with conditionally null alleles of the test genes (*17*–*19*), we selected 469 human genes to study (Fig. 1A).

We first subcloned and sequence-verified each human protein coding sequence into a single-copy, centromeric yeast plasmid under the transcriptional control of either an inducible (GAL) or constitutively active (GPD) promoter (see supplementary materials and methods). We assembled a matched set of yeast strains in which each orthologous yeast gene could be conditionally down-regulated [via a tetracycline-repressible promoter (*17*)], inactivated [via a temperature-sensitive allele (*18*)], or segregated away genetically [following sporulation of a heterozygous diploid deletion strain (*13*, *19*)] (Fig. 1A and fig. S1). After verifying that the loss of the relevant yeast gene conferred a strong growth defect, we tested whether expression of the human ortholog could complement the growth defect, as illustrated for several examples in Fig. 1B (also figs. S2 to S4). When expressed in the permissive condition, 73 of the human genes exhibited toxicity; reducing the genes' expression levels allowed us to assay replacement in 66 cases (table S1).

Overall, we performed 652 informative growth assays surveying 414 human-yeast orthologs (Fig. 1, A and C). In total, 176 yeast genes (43%) could be replaced by their human orthologs in at least one of the three strain backgrounds, whereas 238 (57%) could not (table S1). We collated previously published reports of yeast gene complementation by human genes: Our assays recapitulated these cases with 90% precision and 72% recall (table S1), and incorporating the literature data for subsequent analyses brought the observed complementation rate to 47% (Fig. 1C). For randomly selected subsets of strains, we additionally validated the assays by confirming human protein expression using Western blot analysis (fig. S5), verifying complementation by tetrad dissection (table S1), and subcloning the yeast test genes into the assay vectors and confirming positive complementation (table S2).

Given that roughly half of the tested human genes successfully replaced and half did not, we next investigated factors determining replaceability. We assembled 104 quantitative features of the genes or ortholog pairs, including calculated properties of the genes' sequences (e.g., gene and protein lengths, sequence similarities, codon usage, and predicted protein aggregation potential) and properties such as protein interactions, mRNA and protein abundances, transcription and translation rates, and mRNA splicing features (table S3). We then quantified how well each feature predicted replaceability (Fig. 2A and table S3).

Notably, sequence similarity only partly predicted replaceability. This tendency was strongest for highly similar (>50% amino acid identity) or dissimilar (<20%) ortholog pairs. However, most pairs fell into an intermediate range of 20 to 50% sequence identity, which poorly predicted replaceability (Fig. 2B). Instead, replaceability was best predicted by properties of specific gene modules. In particular, proteins in the same pathway or complex tended to be similarly replaceable (Fig. 2A). Replaceable genes also tended to be shorter and more highly expressed. Using these features in a supervised Bayesian network classification algorithm (fig. S6), we achieved a high overall cross-validated prediction rate [area under the receiver operating characteristic (ROC) curve of 0.825 (Fig. 2A)] and correct prediction of 8 of 10 literature cases withheld from all computational analyses (table S4). Properties such as human gene splice-form counts, yeast 5′ and 3′ untranslated region lengths, codon adaptation measures, and yeast mRNA

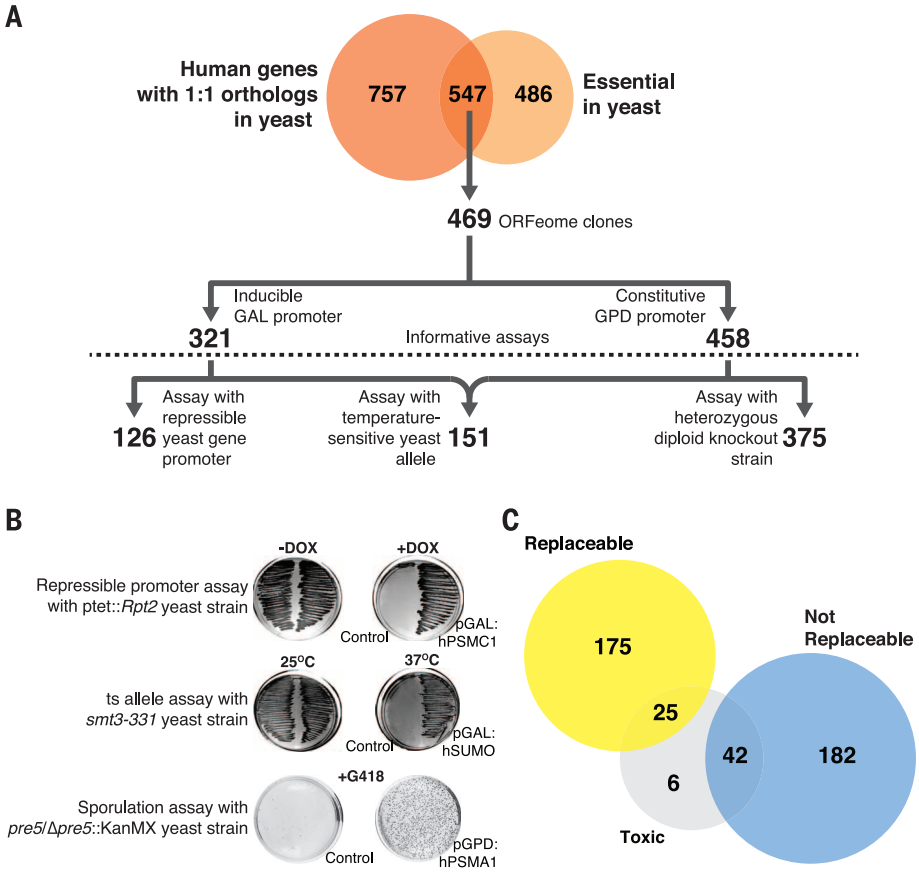


Fig. 1. Systematic functional replacement of essential yeast genes by their human counterparts. (**A**) Of 547 human genes with 1:1 orthology to essential yeast genes, 469 human open reading frames (ORFs) were subcloned into single-copy yeast expression vectors under control of either the GAL or GPD promoters. Using three distinct assay classes (repressible yeast-gene promoter, temperature-sensitive yeast allele, and heterozygous diploid knockout strain), we obtained 126, 151, and 375 informative replaceability assays, respectively. (**B**) Representative examples of the three assay classes. (**C**) Combining assays and literature, 200 human genes could functionally replace their yeast orthologs and 224 genes could not. Some human genes were toxic using GAL induction but replaced their yeast orthologs upon reducing expression.

half-lives showed little relationship with replaceability (Fig. 2A and table S3).

The strong association between replaceability and gene modules led us to investigate this phenomenon in more depth, examining replaceability as a function of specific protein complexes and pathways. Broad Kyoto Encyclopedia of Genes and Genomes (KEGG) (*20*) pathway classes showed highly differential replaceability: Metabolic enzymes (e.g., enzymes participating in lipid, amino acid, and carbohydrate metabolism) tended to be replaceable, whereas proteins involved in DNA replication and repair or in cell growth tended not to be replaceable (Fig. 2C).

Among large protein complexes and pathways, we observed both extremes of replaceability. Some were entirely nonreplaceable: For example, we did not observe a single successful replacement among 13 tested members of the TriC chaperone complex, the DNA replication initiation origin recognition complex, or its interacting minichromosome maintenance (MCM) complex (Fig. 3, A and B). In contrast, some pathways were almost entirely replaceable: Among 19 components of the sterol biosynthesis pathway (which catalyzes the conversion of acetyl–coenzyme A to cholesterol in humans and ergosterol in yeast), only the human farnesyl-diphosphate farnesyltransferase 1 (*FDFT1*) and farnesyl diphosphate synthase (*FDPS*) enzymes failed to replace their yeast orthologs. All other tested components were replaceable, suggesting that yeast and humans both retain the same essential complement of ancestral sterol biosynthesis functionality (Fig. 3C and fig. S7).

The modular nature of replaceability was particularly evident in the case of the 26*S* proteasome complex. Of 28 tested subunits, 21 human

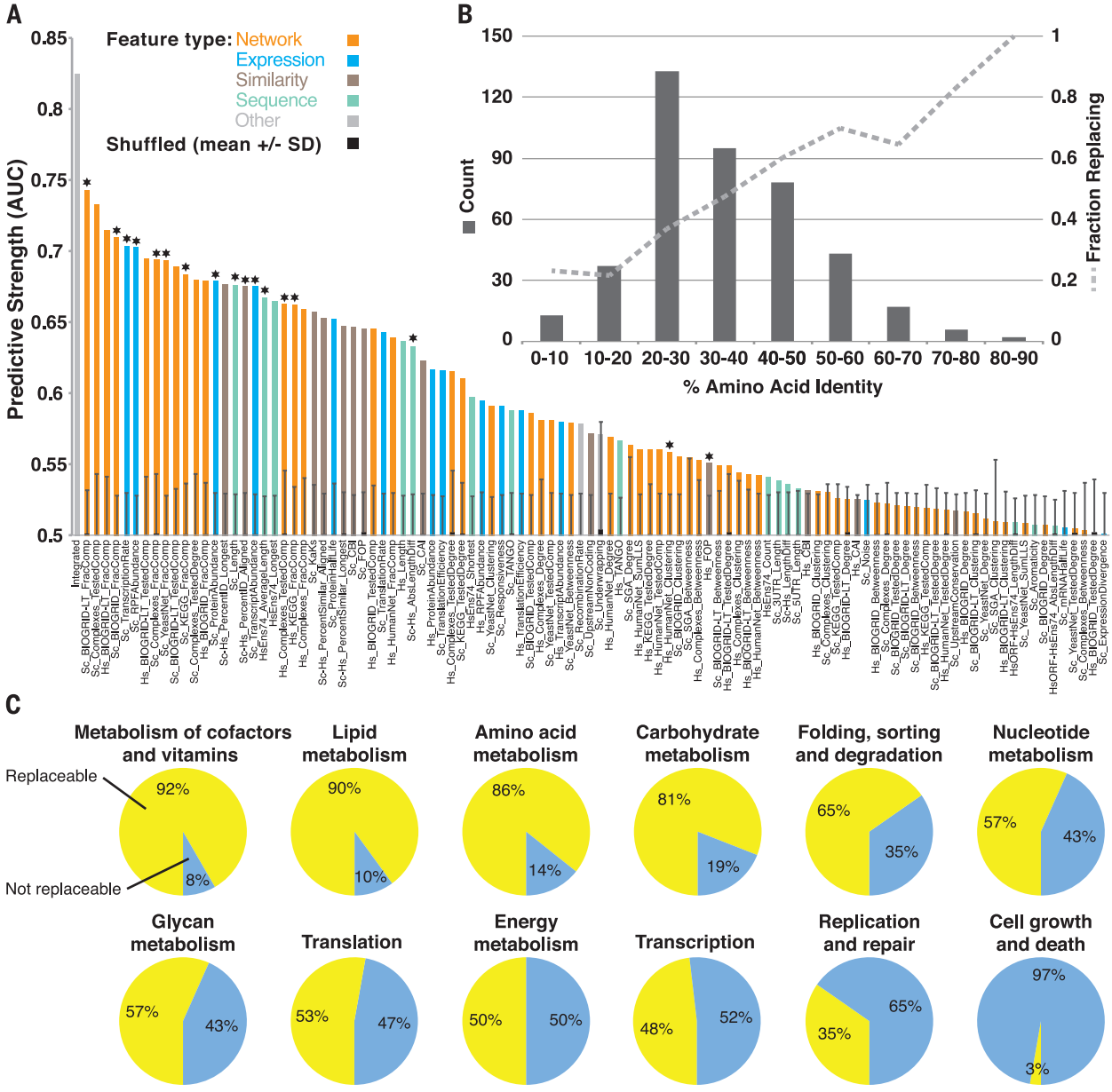


Fig. 2. Properties of gene modules can predict replaceability. (**A**) One hundred four quantitative features of proteins or ortholog pairs were evaluated for their ability to explain replaceability, assessing each feature's predictive strength as the area under a ROC curve (AUC) and determining significance by shuffling replacement status 1000 times, measuring mean AUCs ± 1 SD (error bars). AUCs above 0.58 were generally individually significant with 95% confidence. Starred features were included in the integrated classifier (leftmost bar). (**B**) Distribution of amino acid identities among the tested ortholog pairs (left *y* axis) and fraction of replaceable genes in each sequence-identity bin (right *y* axis). (**C**) Relative proportion of replaceable and nonreplaceable genes among 12 broad KEGG (*20*) pathway classes.

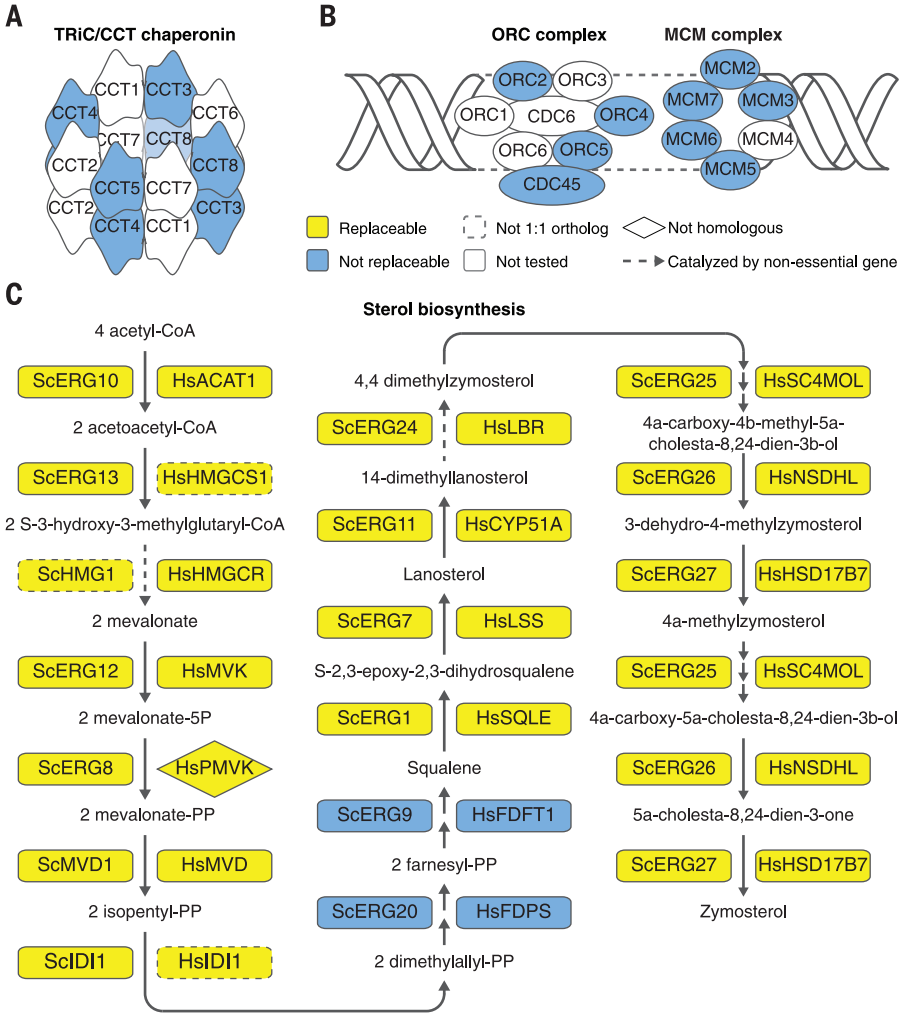


Fig. 3. The modular nature of functional replacement. (**A**) None of the four tested human TRiC/CCT chaperonin genes replaced their yeast counterparts. (**B**) Similarly, no genes tested in the origin recognition complex (ORC) or the MCM complex were replaceable. (**C**) In contrast, 17 of 19 sterol biosynthesis genes were replaceable. In two cases, the yeast gene had two human orthologs but only one could complement. Human *HMGCS1* (but not *HMGCS2*) replaced yeast *ERG13*; human *IDI1* (but not *IDI2*) replaced yeast *IDI1*. Human PMVK, a nonhomologous protein that carries out the same reaction as yeast Erg8 (*27*), complemented temperature-sensitive allele *erg8-1*.

genes replaced their yeast counterparts (Fig. 4A). However, the nonreplaceable subunits were not randomly distributed; rather, they clustered in two physically interacting groups: one consisting of the 19*S* lid components Rpn3 and Rpn12 and one consisting of the 20*S* inner-core heptameric beta-ring subunits β1, β2, β5, β6, and β7. Thus, of the two central heteroheptameric rings, all testable components of the alpha ring replaced, whereas most of the beta ring did not.

An examination of the alpha and beta subunit structures showed that subunit-subunit interfacial amino acids were conserved to similar degrees between yeast and human subunits (fig. S8A), although beta subunits exhibited elevated rates of nonsynonymous substitutions compared with alpha subunits (fig. S8B). Even when interfacial amino acids were only partly conserved, modeling human alpha subunits into the known structure of the yeast proteasome (*21*) revealed that human proteins could be sterically accommodated into the yeast intersubunit interface, as shown for human α6 (Fig. 4B) packing against yeast β6, in spite of sharing only 50% identical amino acids at the interface (fig. S8A). Only orthologous alpha subunits replaced; nonorthologs failed (fig. S9).

We further confirmed this trend across alpha and beta proteasome subunits by cloning and assaying subunits from additional organisms, including another yeast (*Saccharomyces kluyveri*), the nematode *Caenorhabditis elegans*, and several beta subunits from the frog *Xenopus laevis*. In all cases, alpha subunits complemented the loss of the yeast orthologs, whereas beta subunits generally failed to complement (Fig. 4C). The pattern of replaceability across species suggests that alpha and beta subunits experienced different evolutionary pressures, in each case operating at the level of the system of genes (the alpha or beta heteroheptamer).

To determine further why proteasome alpha subunits were replaceable but beta subunits were not, we isolated human β2 subunit mutants that complemented the yeast defect (figs. S10 to S12). A single serine-to-glycine substitution [Ser^{214}→Gly^{214} (S214G)] was sufficient to rescue growth (fig. S11). β2 subunits act as proteases, but yeast β2 catalytic activity is dispensable if the proteasome assembles with other functioning protease subunits (*22*). Notably, a catalytically dead [Thr^{44}→ Ala^{44} (T44A)] human β2 failed to complement, whereas an S214G, T44A double mutant complemented successfully (fig. S11). We conclude that the S214G mutant is competent to assemble an intact proteasome, although the subunit may not be catalytically active. Thus, native human β2 needs only one amino acid change to pack within the yeast proteasome.

Theory predicts that evolutionary divergence creates Dobzhansky-Muller incompatibilities, because evolutionarily novel mutations in one species are untested in the other species' genetic background and may be deleterious there (*23*, *24*). To better understand how proteins retain the ability to interact with their ortholog's interaction partners, even when they have diverged substantially, we developed a biochemically realistic divergence model in which we simulated the evolution of two physically interacting proteins, which both diverge over time. We considered three distinct scenarios: (i) Both thermodynamic stability and binding to the extant partner were selected at ancestral levels; (ii) binding was selected at ancestral levels but stability was not; and (iii) stability was selected at ancestral levels but binding was not. Thermodynamic stability ($\Delta G^{folding}$) and binding energy ($\Delta G^{interaction}$) were calculated using the empirical FoldX energy function (*25*). Under all scenarios, we evaluated whether an evolved member of the pair could still bind to its ancestral partner,

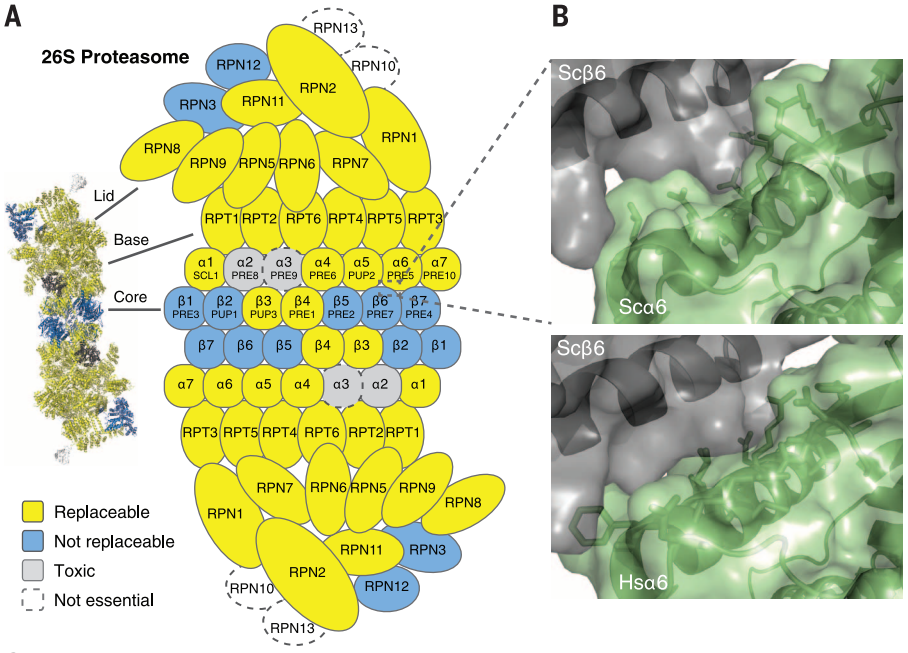
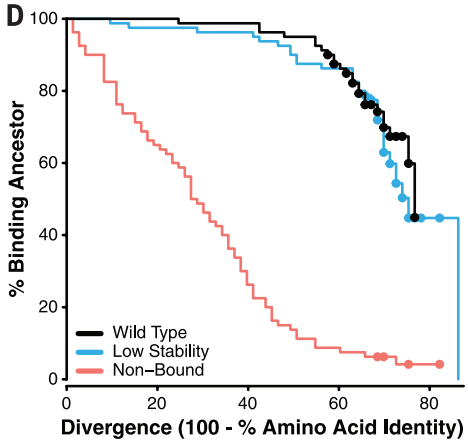


Fig. 4. Proteasome subunits are differentially replaceable. (**A**) Yeast 26*S* proteasome genes were generally replaceable, except for two interacting clusters, in the 19*S* regulatory "lid" particle and in the 20*S* core β-subunit ring. (**B**) The yeast α6-β6 subunit interface (top panel) sterically accommodates the human subunit (bottom panel, showing superposition of human α6 onto the yeast α6) despite 50% sequence identity at the interface. (**C**) Alpha subunits from diverse eukaryotes generally complemented the yeast mutant, but beta subunits did not (unlike plasmid-expressed *S. cerevisiae* genes, included as positive controls). ND, not determined. (**D**) In simulated evolution of interacting proteins Ubc9 and Smt3, if binding to the extant partner is not enforced ("Non-Bound"), a protein's ability to bind its ancestral partner decays rapidly as sequences diverge. However, if extant binding is enforced ("Wild Type" and "Low Stability"), even highly diverged proteins often still bind to their ancestral partners. (Dots indicate right-censored data; see fig. S14.)

C

Deleted 20S subunit	*S. cerevisiae*	*S. kluyveri*	*C. elegans*	*X. laevis*	*H. sapiens*
Δα1				ND	
Δα4				ND	
Δα5			ND	ND	
Δα6				ND	
Δα7				ND	
Δβ1					
Δβ2					
Δβ3				ND	
Δβ4				ND	
Δβ5					
Δβ6				ND	
Δβ7				ND	



for which binding was not enforced. We found that ancestral binding decayed rapidly under scenario (iii) but much more slowly under the other two scenarios (Fig. 4D and figs. S13 to S15). Natural selection for a protein interaction thus preserves the interaction interface in a manner consistent with binding to the ancestral partner (figs. S16 and S17), even though many lineages will eventually accumulate mutations that cause incompatibilities with the ancestral interactor.

Our data demonstrate that a substantial portion of conserved yeast and human genes perform much the same roles in both organisms, to an extent that the protein-coding DNA of a human gene can actually substitute for that of the yeast. The strong pathway-specific pattern of individual replacements suggests that group-wise replacement of the genes should be feasible, raising the possibility of humanizing entire cellular processes in yeast. Such strains would simplify drug discovery against human proteins, enable studies of the consequences of human genetic polymorphisms [as in (*26*) and fig. S7], and empower functional studies of entire human cellular processes in a simplified organism.

REFERENCES AND NOTES

1. T. Gabaldón, E. V. Koonin, *Nat. Rev. Genet.* **14**, 360–366 (2013).
2. N. L. Nehrt, W. T. Clark, P. Radivojac, M. W. Hahn, *PLOS Comput. Biol.* **7**, e1002073 (2011).
3. X. Chen, J. Zhang, *PLOS Comput. Biol.* **8**, e1002784 (2012).
4. L. Kapitzky *et al.*, *Mol. Syst. Biol.* **6**, 451 (2010).
5. C. J. Ryan *et al.*, *Mol. Cell* **46**, 691–704 (2012).
6. A. Frost *et al.*, *Cell* **149**, 1339–1352 (2012).
7. E. J. P. Douzery, E. A. Snell, E. Bapteste, F. Delsuc, H. Philippe, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15386–15391 (2004).
8. K. P. O'Brien, M. Remm, E. L. L. Sonnhammer, *Nucleic Acids Res.* **33**, D476–D480 (2005).
9. K. Dolinski, D. Botstein, *Annu. Rev. Genet.* **41**, 465–507 (2007).
10. M. E. Basson, M. Thorsness, J. Finer-Moore, R. M. Stroud, J. Rine, *Mol. Cell. Biol.* **8**, 3797–3808 (1988).
11. B. P. Cormack, M. Strubin, L. A. Stargell, K. Struhl, *Genes Dev.* **8**, 1335–1343 (1994).
12. M. J. Osborn, J. R. Miller, *Brief. Funct. Genomics Proteomics* **6**, 104–111 (2007).
13. A. H. Tong *et al.*, *Science* **294**, 2364–2368 (2001).
14. E. A. Winzeler *et al.*, *Science* **285**, 901–906 (1999).
15. The MGC Project TeamG. Temple *et al.*, *Genome Res.* **19**, 2324–2333 (2009).
16. J.-F. Rual *et al.*, *Nature* **437**, 1173–1178 (2005).
17. S. Mnaimneh *et al.*, *Cell* **118**, 31–44 (2004).
18. Z. Li *et al.*, *Nat. Biotechnol.* **29**, 361–367 (2011).
19. X. Pan *et al.*, *Mol. Cell* **16**, 487–496 (2004).
20. M. Kanehisa, S. Goto, *Nucleic Acids Res.* **28**, 27–30 (2000).
21. M. Groll *et al.*, *Nature* **386**, 463–471 (1997).
22. M. Groll *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 10976–10983 (1999).
23. H. A. Orr, *Genetics* **139**, 1805–1813 (1995).
24. J. J. Welch, *Evolution* **58**, 1145–1156 (2004).
25. R. Guerois, J. E. Nielsen, L. Serrano, *J. Mol. Biol.* **320**, 369–387 (2002).
26. N. J. Marini, P. D. Thomas, J. Rine, *PLOS Genet.* **6**, e1000968 (2010).
27. S. M. Houten, H. R. Waterham, *Mol. Genet. Metab.* **72**, 273–276 (2001).

ACKNOWLEDGMENTS

We thank M. Minnix and A. Royall for assistance with cloning and assays, K. Drew for structural modeling assistance, M. Tsechansky for TANGO assistance, and C. Boone for providing the temperature-sensitive yeast strain collection. This work was supported by Cancer Prevention and Research Institute of Texas (CPRIT) research fellowships to A.H.K. and J.M.L; NIH grant R01 GM088344, Defense Threat Reduction Agency grant HDTRA1-12-C-0007, and NSF Science and Technology Center BEACON funds (DBI-0939454) to C.O.W.; and grants from the NIH, NSF, CPRIT, and the Welch foundation (F-1515) to E.M.M.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/921/suppl/DC1
Materials and Methods
Figs. S1 to S18
Tables S1 to S6
References (*28*–*55*)
Data S1 to S3

13 October 2014; accepted 21 April 2015
10.1126/science.aaa0769

HOW ARE RESEARCHERS IMPROVING REPRODUCIBILITY?

98%

of researchers are

- ✓ performing more rigorous quality controls
- ✓ increasing sample sizes
- ✓ ensuring thorough documentation

BUT

Less than 1 in 4
are able to reproduce published work from other labs.

Learn more and be part of the change.

Educate and advocate for improved reproducibility.

Download the 2014 Survey Report
sigma-aldrich.com/translational-survey

SIGMA-ALDRICH®

HOW ARE RESEARCHERS IMPROVING REPRODUCIBILITY?

98%

of researchers are

- ✓ performing more rigorous quality controls
- ✓ increasing sample sizes
- ✓ ensuring thorough documentation

BUT

Less than 1 in 4
are able to reproduce published work from other labs.

Learn more and be part of the change.

Educate and advocate for improved reproducibility.

Download the 2014 Survey Report
sigma-aldrich.com/translational-survey

SIGMA-ALDRICH®

AAAS 2016 ANNUAL MEETING | WASHINGTON, DC FEBRUARY 11–15

Global Science Engagement

This year's meeting focuses on how the scientific enterprise can meet global challenges in need of innovation and international collaboration.

aaas.org/meetings

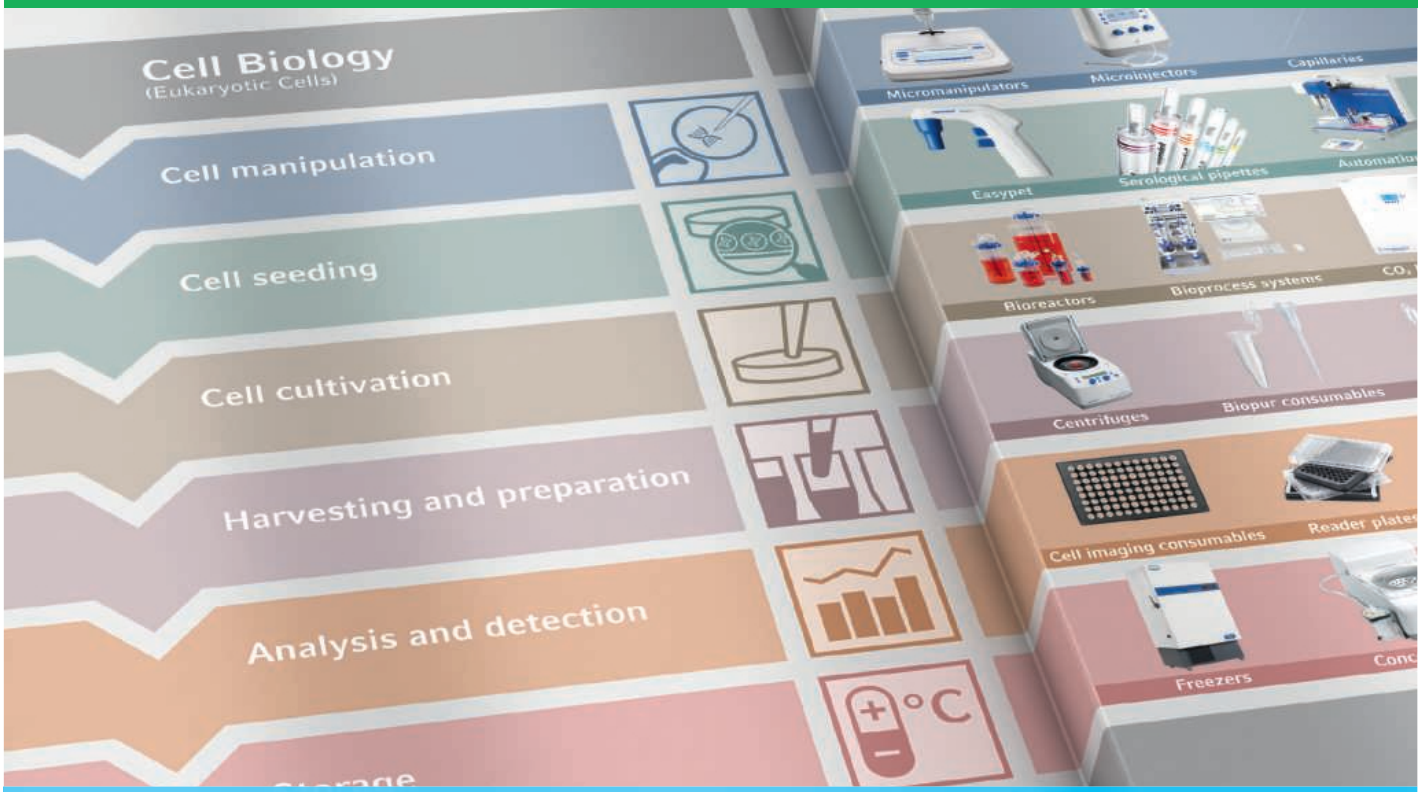
Join us in Washington, DC

The AAAS Annual Meeting is interdisciplinary and inclusive. Each year, thousands of leading scientists, engineers, educators, policymakers, and journalists gather from around the world to discuss recent developments in science and technology.

Advance registration opens in August.

eppendorf

adVantage **New Promotion March 1 – June 30, 2015**



Take Good Care of Your Cells

Smart solutions for your cell biology processes—now with up to 21 % off

As a workflow-oriented provider of lab equipment, Eppendorf offers instruments, consumables, and accessories that perfectly fit your processes in the lab and thus your daily lab work. Our comprehensive solutions are engineered with smart innovations to simplify or even eliminate cumbersome lab work.

Benefit now from substantial savings on:

> Eppendorf Xplorer® electronic multi-channel pipettes
> Centrifuge 5702 R with rotor
> Eppendorf ThermoMixer® C
> Eppendorf BioSpectrometer® fluorescence with µCuvette G1.0
> Divisible twin.tec PCR Plates 96

www.eppendorf.com/advantage

2015 MRS® FALL MEETING & EXHIBIT

November 29 – December 4, 2015 | Boston, Massachusetts

CALL FOR PAPERS

Abstract Submission Opens May 18, 2015 | Abstract Submission Deadline June 18, 2015

A Engaged Learning of Materials Science and Engineering in the 21st Century

BIOMATERIALS AND SOFT MATERIALS

B Stretchable and Active Polymers and Composites for Electronics and Medicine
C Tough, Smart and Printable Hydrogel Materials
D Biolological and Bioinspired Materials in Photonics and Electronics—Biology, Chemistry and Physics
E Engineering and Application of Bioinspired Materials
F Biomaterials for Regenerative Engineering
G Plasma Processing and Diagnostics for Life Sciences
H Multifunctionality in Polymer-Based Materials, Gels and Interfaces
I Nanocellulose Materials and Beyond—Nanoscience, Structures, Devices and Nanomanufacturing
J Wetting and Soft Electrokinetics
K Materials Science, Technology and Devices for Cancer Modeling, Diagnosis and Treatment
L Nanofunctional Materials, Nanostructures and Nanodevices for Biomedical Applications

NANOMATERIALS AND SYNTHESIS

M Micro- and Nanoscale Processing of Materials for Biomedical Devices
N Magnetic Nanomaterials for Biomedical and Energy Applications
O Plasmonic Nanomaterials for Energy Conversion
P Synthesis and Applications of Nanowires and Hybrid 1D-0D/2D/3D Semiconductor Nanostructures
Q Nano Carbon Materials—1D to 3D
R Harsh Environment Sensing—Functional Nanomaterials and Nanocomposites, Materials for Associated Packaging and Electrical Components and Applications

MECHANICAL BEHAVIOR AND FAILURE OF MATERIALS

S Mechanical Behavior at the Nanoscale
T Strength and Failure at the Micro- and Nanoscale—From Fundamentals to Applications
U Microstructure Evolution and Mechanical Properties in Interface-Dominated Metallic Materials
V Gradient and Laminate Materials
W Materials under Extreme Environments (MuEE)
Y Shape Programmable Materials

ELECTRONICS AND PHOTONICS

Z Molecularly Ordered Organic and Polymer Semiconductors—Fundamentals and Devices
AA Organic Semiconductors—Surface, Interface and Bulk Doping
BB Innovative Fabrication and Processing Methods for Organic and Hybrid Electronics
CC Organic Bioelectronics—From Biosensing Platforms to Implantable Nanodevices
DD Diamond Electronics, Sensors and Biotechnology—Fundamentals to Applications
EE Beyond Graphene—2D Materials and Their Applications
FF Integration of Functional Oxides with Semiconductors
GG Emerging Materials and Platforms for Optoelectronics
HH Optical Metamaterials—From New Plasmonic Materials to Metasurface Devices
II Phonon Transport, Interactions and Manipulations in Nanoscale Materials and Devices—Fundamentals and Applications
JJ Multiferroics and Magnetoelectrics
KK Materials and Technology for Non-Volatile Memories

ENERGY AND SUSTAINABILITY

LL Materials and Architectures for Safe and Low-Cost Electrochemical Energy Storage Technologies
MM Advances in Flexible Devices for Energy Conversion and Storage
NN Thin-Film and Nanostructure Solar Cell Materials and Devices for Next-Generation Photovoltaics
OO Nanomaterials-Based Solar Energy Conversion
PP Materials, Interfaces and Solid Electrolytes for High Energy Density Rechargeable Batteries
QQ Catalytic Materials for Energy
RR Wide-Bandgap Materials for Energy Efficiency—Power Electronics and Solid-State Lighting
SS Progress in Thermal Energy Conversion—Thermoelectric and Thermal Energy Storage Materials and Devices

THEORY, CHARACTERIZATION AND MODELING

TT Topology in Materials Science—Biological and Functional Nanomaterials, Metrology and Modeling
UU Frontiers in Scanning Probe Microscopy
VV *In Situ* Study of Synthesis and Transformation of Materials
WW Modeling and Theory-Driven Design of Soft Materials
XX Architected Materials—Synthesis, Characterization, Modeling and Optimal Design
YY Advanced Atomistic Algorithms in Materials Science
ZZ Material Design and Discovery via Multiscale Computational Material Science
AAA Big Data and Data Analytics for Materials Science
BBB Liquids and Glassy Soft Matter—Theoretical and Neutron Scattering Studies
CCC Integrating Experiments, Simulations and Machine Learning to Accelerate Materials Innovation
DDD Lighting the Path towards Non-Equilibrium Structure-Property Relationships in Complex Materials

X Frontiers of Material Research

www.mrs.org/fall2015

The **MRS/E-MRS Bilateral Conference on Energy** *will be comprised of the energy-related symposia at the 2015 MRS Fall Meeting.*

Meeting Chairs

T. John Balk University of Kentucky
Ram Devanathan Pacific Northwest National Laboratory
George G. Malliaras Ecole des Mines de St. Etienne
Larry A. Nagahara National Cancer Institute
Luisa Torsi University of Bari "A. Moro"

Don't Miss These Future MRS Meetings!

2016 MRS Spring Meeting & Exhibit
March 28 - April 1, 2016
Phoenix, Arizona

2016 MRS Fall Meeting & Exhibit
November 27 - December 2, 2016
Boston, Massachusetts

MRS MATERIALS RESEARCH SOCIETY®
Advancing materials. Improving the quality of life.
506 Keystone Drive • Warrendale, PA 15086-7573
Tel 724.779.3003 • Fax 724.779.8313
info@mrs.org • www.mrs.org

Fast GC Module

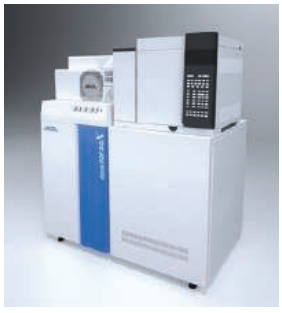
Proton Transfer Reaction-Time of Flight (PTR-TOF) Mass Spectrometry systems are capable of measuring trace gas samples in real-time with a high mass resolving power. The new "fastGC" module adds an optional chemical pre-separation step before the analysis. The module consists of a short GC column with an advanced heating concept for ultrafast heating and equally fast cooling rates that makes this pre-separation step nearly real-time. The fastGC module is integrated with the PTR-TOF and the normal sample gas inlet is utilized. This allows researchers to perform real-time measurements and add fastGC runs at time points of interest for enhanced separation and identification. The PTR-TOF 1000 is by far the most affordable, smallest, and also lightweight system available on the market. It provides all the advantages of a powerful time of flight-based solutions: the entire mass range in split seconds and higher resolution for better separation and identification.

Ionicon
For info: +43-512-214-800
www.ionicon.com/fastgc

TLC-Compact Mass Spectrometry System

The Plate Express is an automated software-controlled system designed to directly analyze thin-layer chromatography (TLC) plates and a full range of other planar surfaces by mass spectrometry. Plate Express provides a simple, automated, software-clicked means of visually pinpointing and extracting compounds from a range of TLC plate formats into Advion's expression compact mass spectrometer (CMS). The combined technique is known as TLC/CMS. Synthetic organic, natural product, and peptide chemists can quickly and confidently identify analytes in complex mixtures without additional sample preparation utilizing TLC/CMS. Plate Express is controlled by Advion's Mass Express software, vastly simplifying the process and increasing the robustness of TLC spot and planar surface analysis. A pressure-sensing feature allows for uncommon force management to be controlled, superbly sealing around the point of interest, while extending the life of the probe, further reducing the cost of consumables. The software allows methods to be developed for plates and surfaces of choice.

Advion
For info: 607-266-9162
www.advion.com



GCxGC Mass Spectrometer

The AccuTOF-GCx, the fourth generation of JEOL's gas chromatography/time-of-flight mass spectrometer systems, is designed for optimum throughput, operation, and uptime. It offers improved resolution, accuracy, and sensitivity, while retaining the power and flexibility of the previous models. In combination with comprehensive 2-D gas chromatography (GCxGC) using the Zoex thermal modulator, the GCx offers both powerful chromatographic separation and high-resolution mass spectra. An optional combination EI/FI/FD ion source eliminates the need for source exchange for these experiments. Gas chromatography/field ionization can also be used to characterize samples that would be difficult to analyze by any other technique. While hundreds or even thousands of components can be separated and detected using this type of system, interpreting the data sets can be challenging due to the unprecedented amount of information the data provides. JEOL has formed collaborations to develop new software methods and tools to simplify the analysis of GCxGC/HRMS data sets.

JEOL
For info: 978-535-5900
www.jeol.com

Mass Spectrometry Quantification Software

The new TASQ 1.0 and Pacer 2.0 software products allow users to easily screen, identify, confirm, and quantify hundreds of compounds in a single experiment. The Bruker TASQ (Target Analysis for Screening and Quantitation) software is specifically designed to exploit high resolution, accurate-mass data generated by Bruker QTOF mass spectrometers to confidently screen for trace residues in complex matrices. TASQ also efficiently exploits diagnostic ion confirmation criteria to eliminate false positive findings. Bruker's PACER 2.0 software provides extremely fast, accurate quantitative results for high throughput targeted analyses in the routine lab by building on the powerful Bruker GC and LC Triple Quad MS instruments. PACER addresses the real crunch in quantitative data review—peak integration—by using its powerful "Exception Based Review" feature set. This newest version of Pacer introduces a new, modern interface designed for simplicity and clarity, presenting information and options at the time you need them.

Bruker Corporation
For info: +49-6181-4384-100
www.bruker.com

Liquid Chromatography Systems

Two new integrated liquid chromatography (LC) systems, the Prominence-i and Nexera-i, have been added to the company's extensive line-up of high performance LC (HPLC) and UHPLC systems. Combining excellent functionality, an intuitive operating environment, and full automation, the i-Series provides excellent performance and a more efficient workflow for conventional to ultrahigh-speed analysis. Through the integration of these systems with LabSolutions software, Shimadzu fosters a new relationship between users and instrumentation. The data acquired by the Prominence-i and Nexera-i via interactive communication mode (ICM) is sent to a lab's data center by the LabSolutions network and managed uniformly by a server. ICM allows users to perform operations such as purging mobile phases and confirming analytical results from anywhere in the facility with a smart device. It also permits easy access to a system installed in a closely supervised area such as under a hood where highly active ingredients are being analyzed.

Shimadzu Scientific Instruments
For info: 800-477-1227
www.ssi.shimadzu.com

Electronically submit your new product description or product literature information! Go to www.sciencemag.org/products/newproducts.dtl for more information.

Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and governmental organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier.

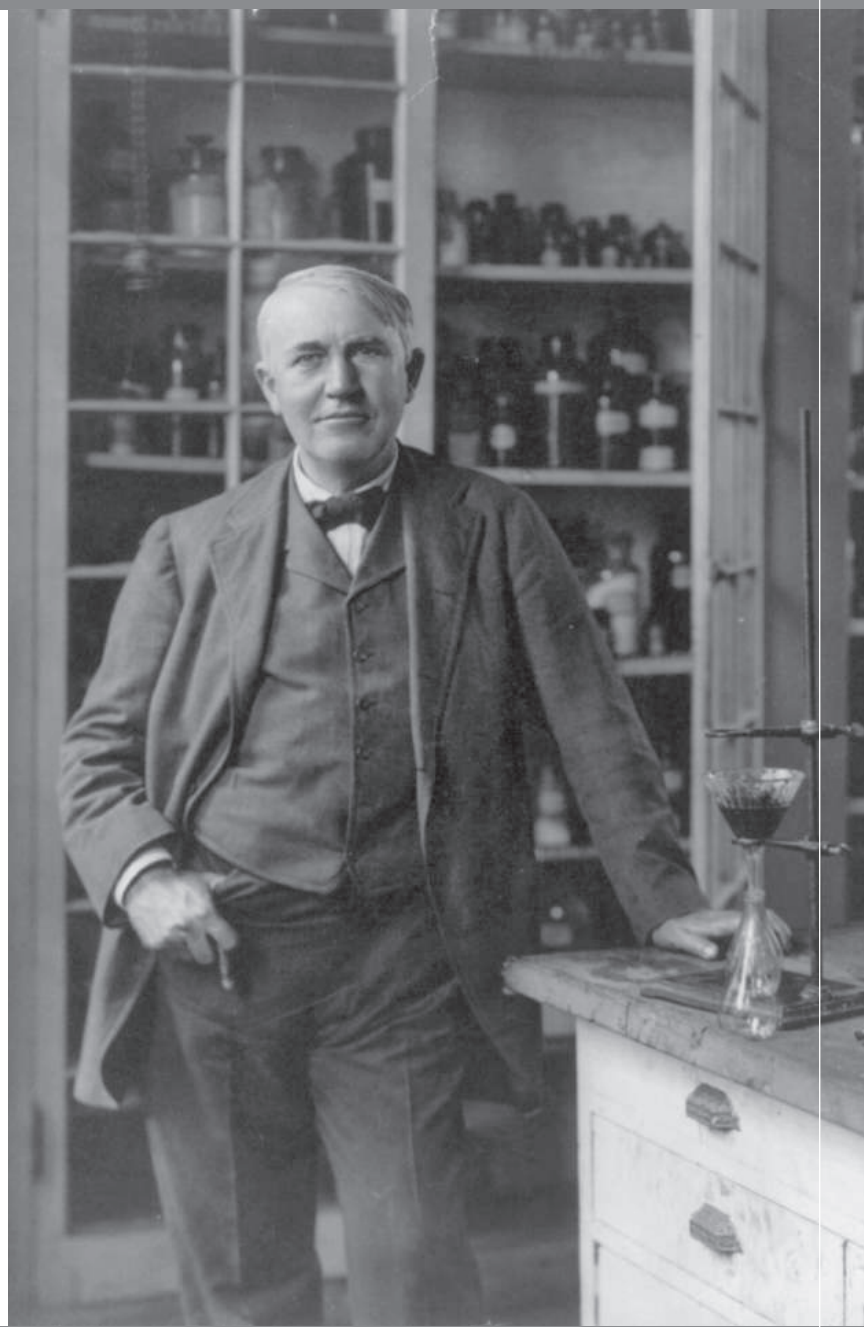
WHAT DO YOU AND THOMAS EDISON HAVE IN COMMON?

AAAS.

By investing in AAAS you join Thomas Edison and the many distinguished individuals whose vision led to the creation of AAAS and our world-renowned journal, *Science,* more than 150 years ago.

Like Edison, you can create a legacy that will last well into the future through planned giving to AAAS. By making AAAS a beneficiary of your will, trust, retirement plan, or life insurance policy, you make a strong investment in our ability to advance science in the service of society for years to come.

To discuss your legacy planning, contact Juli Staiano, Director of Development, at (202) 326-6636, or jstaiano@aaas.org, or visit www.aaas.org/1848 for more information.



1848 SOCIETY
AAAS

"I feel great knowing that I will leave behind a legacy that will be channeled through the AAAS. It also means a lot to me to be able to honor my late parents, too."

–PETER ECKEL
Member, 1848 Society and AAAS Member since 1988

B7 FAMILY ANTIBODIES

CST development scientist optimizing IHC protocols

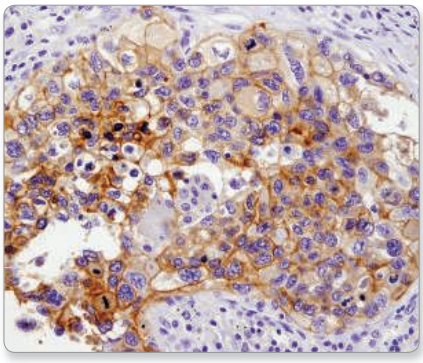
極めて重要なターゲット

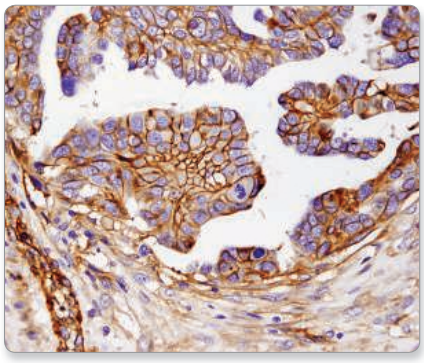
CANCER IMMUNOLOGY

実証された特異性　信頼に足る結果

CSTのPD-L1、B7-H3、B7-H4抗体は、ヒト組織の内在性タンパク質を検出し、
他のB7ファミリータンパク質には交差しません。

PD-L1 (E1L3N®) XP® Rabbit mAb #13684:
IHC analysis of paraffin-embedded human lung carcinoma using #13684.





B7-H3 (D9M2L) XP® Rabbit mAb #14058:
IHC analysis of paraffin-embedded ovarian carcinoma using #14058.

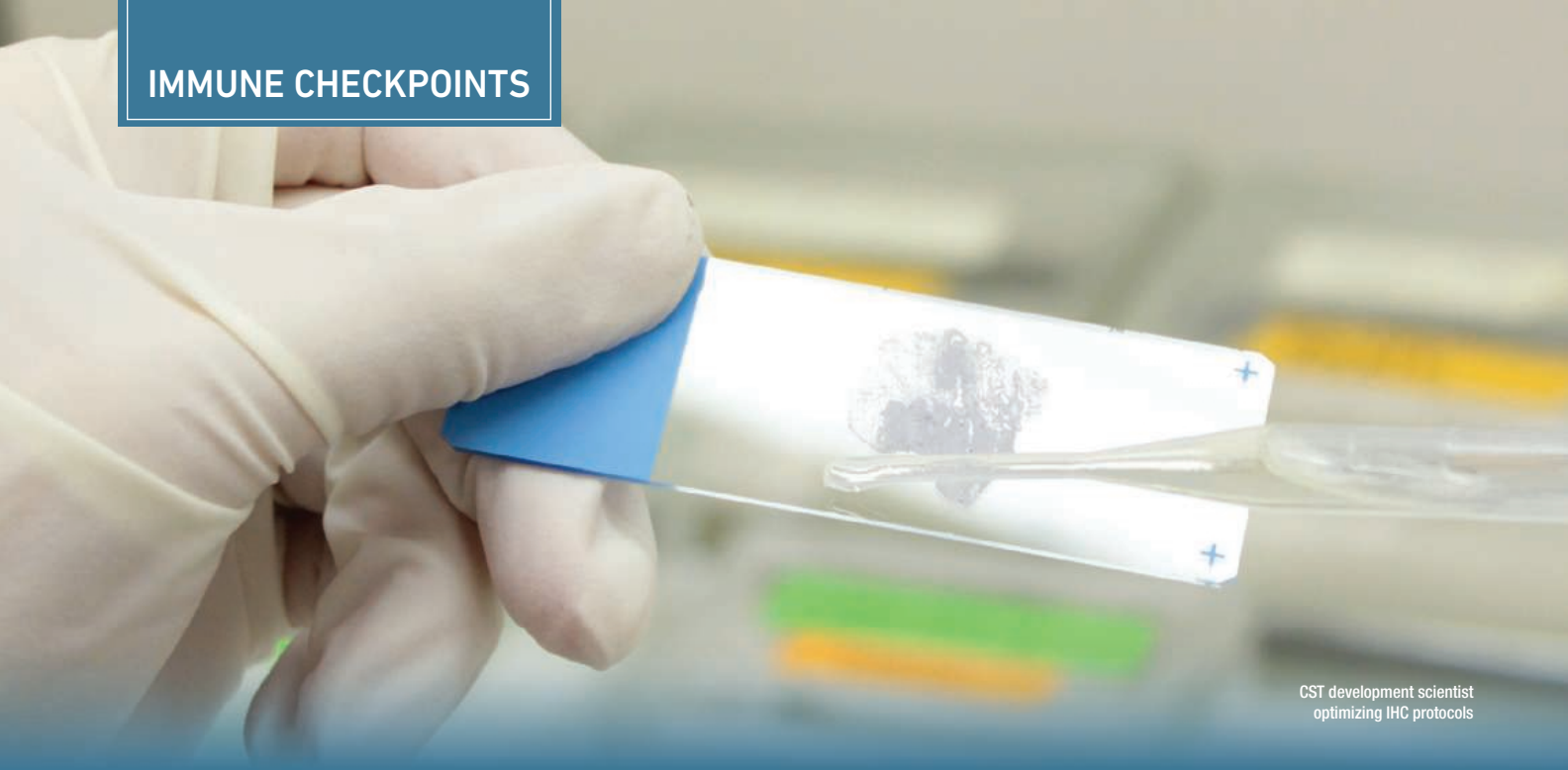
www.cstj.co.jp/B7science

腫瘍免疫ポスター！上記ご請求ページで、社内検証データや他社抗体との比較データもご覧いただけます。

15PADCANRSCIE0014JPN_00

CSTジャパン株式会社
Cell Signaling Technology®

IMMUNE CHECKPOINTS

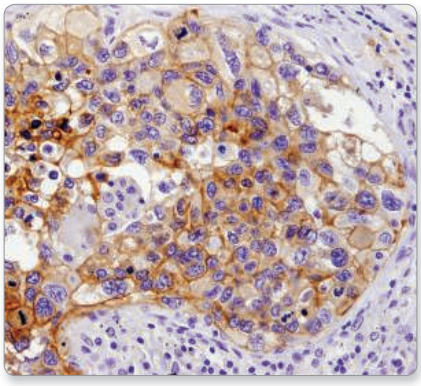


CST development scientist optimizing IHC protocols

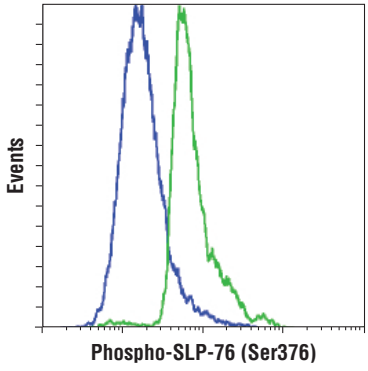
PIVOTAL TARGETS
CANCER IMMUNOLOGY

Proven specificity & sensitivity. Results you can count on.

Antibodies for PD-L1, B7-H3, B7-H4, Phospho-SLP-76 (Ser376), Phospho-Stat3 (Tyr705), and more from CST.



PD-L1 (E1L3N®) XP® Rabbit mAb #13684: IHC analysis of paraffin-embedded human lung carcinoma using #13684.



Phospho-SLP-76 (Ser376) (D9D6E) Rabbit mAb #14745: Flow cytometric analysis of Jurkat cells, untreated (blue) or treated with H2O2 (11 mM, 1 min; green), using #14745. Anti-rabbit IgG (H+L), $F(ab')_2$ Fragment (Alexa Fluor® 488 Conjugate) #4412 was used as a secondary antibody.

www.cellsignal.com/cancerscience

Visit our website to request our **Tumor Immunology Poster** and for additional validation and competitor comparison data.

For Research Use Only. Not For Use In Diagnostic Procedures.

14PADCANRIMMU0257ENG_02

Cell Signaling TECHNOLOGY®

Science Careers Advertising

For full advertising details, go to ScienceCareers.org and click For Employers, or call one of our representatives.

Tracy Holmes
Worldwide Associate Director
Science Careers
Phone: +44 (0) 1223 326525

THE AMERICAS

E-mail: advertise@sciencecareers.org
Fax: 202 289 6742

Tina Burks
Phone: 202 326 6577

Nancy Toema
Phone: 202 326 6578

Marci Gallun
Sales Administrator
Phone: 202 326 6582

Online Job Posting Questions
Phone: 202 312 6375

EUROPE / INDIA / AUSTRALIA / NEW ZEALAND / REST OF WORLD

E-mail: ads@science-int.co.uk
Fax: +44 (0) 1223 326532

Axel Gesatzki
Phone: +44 (0) 1223 326529

Sarah Lelarge
Phone: +44 (0) 1223 326527

Kelly Grace
Phone: +44 (0) 1223 326528

JAPAN

Katsuyoshi Fukamizu (Tokyo)
E-mail: kfukamizu@aaas.org
Phone: +81 3 3219 5777

Hiroyuki Mashiki (Kyoto)
E-mail: hmashiki@aaas.org
Phone: +81 75 823 1109

CHINA / KOREA / SINGAPORE / TAIWAN / THAILAND

Ruolei Wu
Phone: +86 186 0082 9345
E-mail: rwu@aaas.org

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

ScienceCareers.org

CHIEF, RNA BIOLOGY LABORATORY
CENTER FOR CANCER RESEARCH
NATIONAL CANCER INSTITUTE - FREDERICK, MARYLAND
NATIONAL INSTITUTES OF HEALTH
DEPARTMENT OF HEALTH AND HUMAN SERVICES
Application Deadline: June 15, 2015

CENTER FOR CANCER RESEARCH

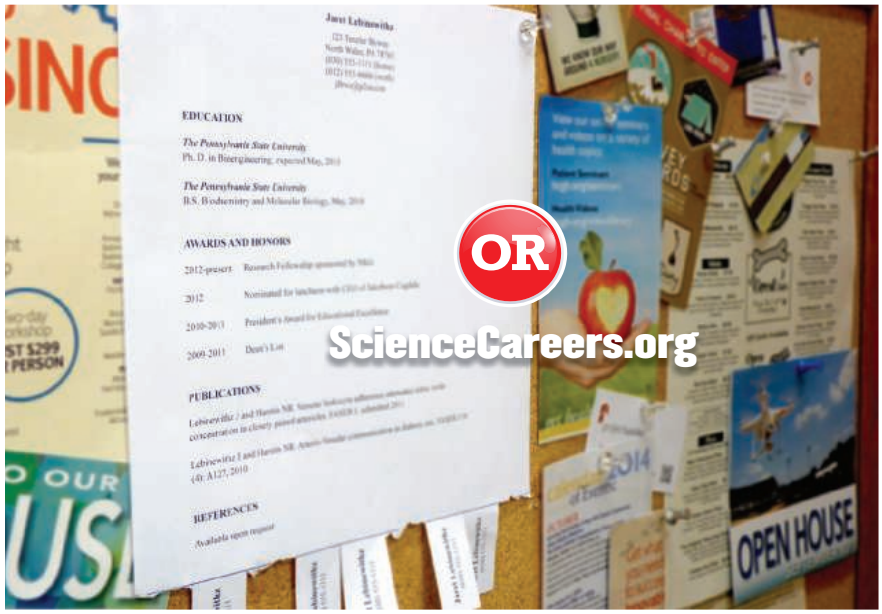
NCI is seeking an outstanding, internationally recognized scientist to serve as Chief of the RNA Biology Laboratory (RBL) in the Center for Cancer Research (CCR). The position, which is the equivalent of an academic Department Chair, is the key component of a major initiative to expand CCR's RNA Biology research at the NCI. The RBL Chief will play leading roles in developing an integrated program in RNA Biology and in the CCR RNA Initiative. In addition, the RBL Chief will direct an extensive individual research program at the Frederick campus which will complement and augment CCR expertise in chromosome biology, immunology, HIV/AIDS, cancer biology and molecular oncology, areas in which Centers of Excellence have been established. Supported with stable financial resources, the RBL will have access to a wide array of intellectual and technological assets, including high-quality technology cores dedicated to protein chemistry, natural products chemistry, biophysics, mass spectrometry, imaging, microscopy, proteomics and genomics, bioinformatics/bio-statistics, and flow cytometry, in addition to clinical support.

The National Cancer Institute (NCI) is part of the National Institutes of Health (NIH) in the Department of Health and Human Services (DHHS), a federal government agency. CCR is the largest component of the NCI Intramural Research Program, providing an environment conducive to advancing translational research and collaborative interactions through investigator-initiated and interdisciplinary team science. Additional information on CCR research priorities can be found at: http://ccr.cancer.gov.
In addition to a Ph.D., M.D./Ph.D., or equivalent doctoral degree in a relevant discipline, applicants should possess outstanding communication skills and documented leadership experience. Tenured faculty or industrial scientists of equivalent rank with a demonstrated commitment to RNA Biology should apply. Salary will be commensurate with experience and accomplishments. Applications should include a description of research interests and leadership philosophy, career synopsis, and current curriculum vitae with complete bibliography.

Review of applications will begin on or about **June 15, 2015** but applications will be accepted until the position is filled. Send applications to **Dr. Janelle Cortner, RNA Biology Laboratory Search Committee, National Cancer Institute Building 428/46, PO Box B, Frederick MD 21702**, or by email to **CCR_RNA_Biology@mail.nih.gov**

DHHS, NIH and NCI are Equal Opportunity Employers.

NATIONAL INSTITUTES OF HEALTH
NATIONAL CANCER INSTITUTE



Conduct your job search the easy way.

Target your job search using relevant resources on **ScienceCareers.org**.

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

63 KNOWLEDGE FOR A BETTER WORLD • north

Photo: Carl-Erik Eriksson

Trondheim is the ancient Viking capital of Norway. The Nidelva River flows through the city, and you can even fish for salmon during your lunch break. You'll find hiking, alpine and cross-country skiing, cycling and more within a 10-minute drive of the city centre.

Research adventure at 63 degrees north

Remote yet modern. Beautiful and peaceful. Quality and equality. All that combined with long-term government research funding and the UN's ranking as one of the best countries to live in. Welcome to Trondheim and NTNU. The best minds, world-class research, and just the right size city for a great life.

We welcome top scientists from all cultures and nations who want to contribute to our goal: Knowledge for a better world.

TOP RESEARCH DEMANDS BRILLIANT MINDS
– WE'RE ALWAYS LOOKING FOR THE BEST

Photo: Geir Mogen/Kavli Institute for Systems Neuroscience

NTNU neuroscientists May-Britt and Edvard Moser were awarded the Nobel Prize in Medicine or Physiology in December 2014.

NTNU – Trondheim
Norwegian University of
Science and Technology

www.ntnu.edu

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Professorship in Metabolism

at the Ecole polytechnique fédérale de Lausanne (EPFL)

The Faculty of Life Sciences of the Swiss Federal Institute of Technology Lausanne (EPFL) invites applications for a Professor position in the general field of **Metabolism in the context of Nutrition, human health and disease**. This search is part of an initiative to promote research in the fields of metabolism, nutrition and food. We are primarily seeking candidates for a full professor position but in exceptional cases, for candidates with outstanding credentials/track-records, more junior candidates will be considered.

Successful candidates will develop an independent and dynamic research program, participate in both undergraduate and graduate teaching, and supervise PhD students and postdoctoral fellows. We are seeking candidates with a profound knowledge in intermediary metabolism, including its biochemical and molecular regulation. Familiarity with multiscalar integrative research approaches using genetics, omics technologies, pharmacology, and physiology and that bridge model organisms from yeast to humans are of importance. A good basis of metabolic, endocrine, immune, or gastro-intestinal physiology is required.

This position is offered in an environment of biomedical research, rich for seeking deeper understanding of integrative (patho)physiological mechanisms contributing to the development of complex diseases, with as ultimate goal the development of preventive (nutritional) and therapeutic approaches. The School of Life Sciences fosters interactions with other relevant domains of the EPFL, such as the Schools of Basic Sciences, Engineering, and Information and Communication Technologies, as well as with relevant clinical departments at the Centre Hospitalier Universitaire Vaudois (CHUV), with the Faculty of Biology and Medicine of the University of Lausanne (UNIL) and with the inter-institutional Lausanne Integrative Metabolism and Nutrition Alliance (LIMNA).

Significant start-up resources, research budget and state-of-the-art research infrastructure, include metabolomics, are available, within the framework of a campus. Salaries and benefits are internationally competitive.

Applications should include a cover letter with a statement of motivation, a curriculum vitae with a list of publications, a concise (3-page) statement of research and teaching interests; and the names and contact information of at least five referees. Applications should be sent before **September 15, 2015** to:

https://academicjobsonline.org/ajo/jobs/5542

Enquiries may be addressed to:

Prof. Gisou van der Goot
Dean of Life Sciences
Email: GHI-recruit@epfl.ch

For additional information on EPFL, please consult the web sites www.epfl.ch, www.sv.epfl.ch

EPFL is committed to increasing the diversity of its faculty, and strongly encourages women to apply.

HARVARD UNIVERSITY

Professor of Psychology

The Department of Psychology seeks to appoint a tenured professor whose interdisciplinary research and teaching explores multifaceted factors that guide and affect human behavior. Areas of interest include, but are not limited to, computational cognitive neuroscience, behavioral genetics, gene by environment interactions, developmental cognitive neuroscience, neuroeconomics, or cross-disciplinary approaches to human social behavior. The successful appointee will be expected to strengthen links between the Department of Psychology and the broader scholarly community interested in human behavior. The appointment is expected to begin on July 1, 2016. The professor will teach and advise at the undergraduate and graduate levels.

Candidates are required to have a doctorate. Demonstrated excellence in teaching and research is desired. Candidates should also evince intellectual leadership and impact on the field and potential for significant contributions to the Department, University, and wider scholarly community.

Candidates should submit a cover letter, curriculum vitae, research and teaching statements to:

http://academicpositions.harvard.edu/postings/6093

Questions regarding this position can be addressed to **nock@wjh.harvard.edu**. Applications will be considered starting on **July 1, 2015**.

We are an Equal Opportunity Employer and all qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, national origin, disability status, protected veteran status, or any other characteristic protected by law.

SERRA HÚNTER PROGRAMME

For further details please visit
serrahunter.gencat.cat

The Serra Húnter Programme[1] announces an opening for 33 tenure track positions and 40 senior positions in the Catalan public universities in the following research fields:

ARCHITECTURE, ARTS, BIOLOGY, CHEMISTRY, COMPUTER SCIENCE, ECONOMICS, EDUCATION, ENGINEERING, HUMANITIES, LAW, MATHEMATICS, MEDICINE, PHARMACY, PHYSICS, PSYCHOLOGY, SOCIAL SCIENCES, AND OTHERS.

Minimum requirements are a PhD degree and a proven academic background. Only those applicants with excellent research records, leadership capabilities and, preferably, international exposure at the doctoral or post-doctoral level will be considered.

Successful applicants, unless otherwise stated, will have a permanent contract with one of the Catalan universities, and are expected to cooperate with the existing research groups and develop new research lines, complementary to those already in place. Salaries will be in line with those paid by the Catalan universities, plus (subject to negotiation) an additional amount for those candidates with outstanding research performance, or a relocation grant, if appropriate. Successful applicants will be evaluated after a three-year period, and subsequently every six years. A positive evaluation may lead to a consolidation of the additional payments.

Applications and deadline

Applications must be submitted electronically via the Serra Húnter Programme website. The website provides all the information necessary for application. The deadline is 31 May 2015.

1. The Serra Húnter Programme is funded by the Government of Catalonia and by the seven public universities of Catalonia: University of Barcelona (UB), Universitat Autònoma de Barcelona (UAB), Universitat Politècnica de Catalunya - BarcelonaTech (UPC), Pompeu Fabra University (UPF), University of Lleida (UdL), University of Girona (UdG), and Rovira i Virgili University (URV). Its main aim being to hire high-quality staff for the Catalan universities. The University of Vic – Central University of Catalonia (UVic – UCC) is an associated member of the programme.

Generalitat de Catalunya
Government of Catalonia

Faculty Positions

Candidates must have a Ph.D., M.D., or D.V.M. degree and 4 – 5 years of relevant postdoctoral training, with an exceptional record of research accomplishment. Opportunities are available for shared mentorship of trainees, and collaboration with the other campuses of The Jackson Laboratory in Maine and California.

THE JACKSON LABORATORY FOR GENOMIC MEDICINE | Farmington, Conn.

The Jackson Laboratory is inviting applications for **Assistant, Associate** and **Full Professors**. The campus is dedicated to advancing precision medicine using genomic strategies to understand the complex functional networks underlying health and disease and the development of novel diagnostics and therapeutics. We are seeking individuals to join our interactive culture of cooperation and program integration.

The Jackson Laboratory offers a uniquely collaborative scientific research environment. Faculty are supported by outstanding scientific services, unparalleled mouse and genomic resources, postdoctoral and predoctoral training programs, and numerous courses and conferences.

For inquiries, please contact **Dr. Charles Lee**, Director and Professor of The Jackson Laboratory for Genomic Medicine, at charles.lee@jax.org. Information for The Jackson Laboratory for Genomic Medicine and its current faculty may be found at www.jax.org/ct.

Areas for recruitment include:

- Cancer Biology and Cancer Genomics
- Computational Biology and Bioinformatics
- Functional Genomics and Genomic Technologies
- Systems Genomics
- Metabolic Disorders and Genetics of Aging
- Microbial Genomics, Microbiome Research and Infectious Diseases
- Neurology and Neurobiology

Applicants must apply online. Please submit a curriculum vitae and a concise statement of research interests as one document to www.jax.org/careers/faculty.html. In addition, please have three letters of reference sent to: facultyjobs@jax.org. Applicants will be reviewed on a continuing basis.

The Jackson Laboratory is an equal opportunity, affirmative employer, considering all qualified applicants and employees for hiring, placement and advancement, without regard to a person's race, color, religion, national origin, age, genetic information, military status, gender, sexual orientation, gender identity or expression, disability or protected veteran status.

The Jackson Laboratory

Leading the search for tomorrow's cures

Penn Dental Medicine

Department of Anatomy and Cell Biology
Full-Time Tenure Track or Tenured Faculty Positions,
Assistant, Associate or Professor Level

The Department of Anatomy and Cell Biology (ACB) at the University of Pennsylvania, School of Dental Medicine invites applications for full-time tenure track or tenured faculty positions at the Assistant Professor, Associate Professor or Professor Level.

We are seeking candidates with outstanding academic accomplishments in cellular and/or molecular biology to complement research at the School of Dental Medicine (SDM). A strong track record of extramural grant funding, as well as commitment in teaching basic sciences to dental students are required. Preference will be given to candidates who can foster new and enhance existing interdisciplinary, translational and collaborative research, both within SDM and across the Penn community. Departmental interests include craniofacial development and biology, stem cell and regenerative medicine, immunology, extracellular matrix biology, microbial pathogenesis and oral cancers. Applicants with a PhD or dual degree (DMD-PhD, MD-PhD, DVM-PhD) are invited to submit a statement of research, curriculum vitae and names with contact information for five references. Review of applications begin TBD and will continue until positions are filled. Anticipated start date is TBD. Applicant can apply directly to the University of Pennsylvania website **https://facultysearches.provost.upenn.edu/postings/581**

Information about the position may be obtained from the Administrator to the Chair of the Search Committee: **Amber Banayat, Department of Anatomy and Cell Biology, School of Dental Medicine, University of Pennsylvania, 240 S. 40th Street, Philadelphia, PA 19104; abanayat@dental.upenn.edu.**

The University of Pennsylvania is an Affirmative Action/Equal Opportunity Employer. All qualified applicants will receive consideration for employment and will not be discriminated against on the basis of race, color, religion, sex, sexual orientation, gender identity, creed, national or ethnic origin, citizenship status, age, disability, veteran status, or any other characteristic protected by law.

Iowa State University

College of Veterinary Medicine – Biomedical Sciences
Two Available Tenure Track Pharmacology Positions

Assistant or Associate Professor in Pharmacology

This is a tenure track faculty 9-month position with rank and salary commensurate with qualifications. The successful candidate will teach pharmacology courses to professional and graduate students, mentor graduate students and maintain a dynamic extramurally funded research program in an area of candidate's expertise. Qualified candidates may be considered for an Endowed Chair.

For more information and to apply for this job go to: **https://www.iastatejobs.com/postings/11266**

Questions? Contact: **Dr. Steve Carlson**, Search Committee Chair **515-294-0912**; **stevec@iastate.edu**.

Assistant, Associate, or Full Professor
Clinical/Translational Pharmacology

This tenure track position is part of the High Impact Hires Initiative of Big Data and is among the 29 high-impact hires targeted in this Presidential Initiative. The rank and salary will commensurate with qualifications. The successful candidate will maintain a dynamic extramurally funded research program focused on pharmacogenomics and other translational pharmacological research or pharmacokinetic-pharmacodynamic (PK-PD) mathematical modeling. The incumbent will also teach pharmacology courses to professional and graduate students and mentor graduate students in an area of candidate's expertise. Qualified candidates may be considered for an Endowed Chair.

For more information and to apply for this job go to: **https://www.iastatejobs.com/postings/11268**

Questions? Contact: **Dr. Hans Coetzee**, Search Committee Chair **515-294-7424**; **hcoetzee@iastate.edu**.

For more information about the Department of Biomedical Sciences, please visit **www.http://vetmed.iastate.edu/bms/**

Iowa State University is an Equal Opportunity/Affirmative Action Employer.

Case Western Reserve University
SCHOOL OF MEDICINE

Open Rank Protein Biophysics/Structural Biology
Faculty Position
Department of Physiology and Biophysics

We invite outstanding individuals to apply for a faculty position at any rank in the area of Protein Biophysics and/or Structural Biology. Mid-career scientists with outstanding accomplishments at the level of Associate Professor or full Professor are especially encouraged to apply. We are particularly interested in applicants who are using interdisciplinary approaches to work on basic or translational aspects of human diseases. Visit our website at http://Biophysics.case.edu. The Department and School have excellent infrastructure, particularly in x-ray crystallography and solution NMR spectroscopy (see http://Ccmsb.case.edu).

Applicants for a position as Assistant Professor should have a Ph.D. and/or M.D. degree, 3-5 years postdoctoral experience, and a strong record of scholarly activity. Competitive candidates for Associate Professor should have a strong publication record and an international reputation. Competitive candidates for Professor should have achieved records of leadership in the profession and have a substantial record of scholarly publications.

Applicants should submit a cover letter, a full *Curriculum Vitae*, including a record of prior/current funding, a brief description of their research, as well as the contact information for three professional references. Candidates at the Assistant Professor level should also submit a research plan. Please submit application materials with separate file attachments by email to: **Dr. Walter F. Boron, Chair, Department of Physiology and Biophysics, Case Western Reserve University; BiophysicsSearch@case.edu**.

"In employment, as in education, Case Western Reserve University is committed to Equal Opportunity and Diversity. Women, veterans, members of underrepresented minority groups, and individuals with disabilities are encouraged to apply."

"Case Western Reserve University provides reasonable accommodations to applicants with disabilities. Applicants requiring a reasonable accommodation for any part of the application and hiring process should contact the Office of Inclusion, Diversity and Equal Opportunity at 216-368-8877 to request a reasonable accommodation. Determinations as to granting reasonable accommodations for any applicant will be made on a case-by-case basis."

63 north

KNOWLEDGE FOR A BETTER WORLD

THE ONSAGER FELLOWSHIPS

12 tenure-track positions available at NTNU

The Norwegian University of Science and Technology (NTNU) is Norway's primary institution for educating the future's engineers and scientists. The university also has strong programmes in the social sciences, teacher education, the arts and humanities, medicine, architecture and fine art. NTNU's cross-disciplinary research delivers creative innovations that have far-reaching social and economic impact and that help contribute to a better world.

The Onsager Fellowship programme at NTNU is designed to attract the most talented scholars with an established reputation for high quality research and a commitment to learning and teaching at the university level.

APPLY FOR A TENURE-TRACK POSITION AS AN ASSOCIATE PROFESSOR IN:

- Linguistics
- Robotic vision
- Molecular biodiversity
- Medicine – bioinformatics
- Medicine – molecular biology
- Statistical machine learning
- Theoretical condensed matter physics
- Inorganic or hybrid functional materials
- Safety and reliability of complex systems
- Marine structures for the future – marine technology
- Zero emission refurbishment of the built environment
- Economics of natural resources and quantitative peace research

More info at: www.ntnu.edu/onsagerfellowship

Closing date: 25 May.

NTNU – Trondheim
Norwegian University of Science and Technology

TOP RESEARCH DEMANDS BRILLIANT MINDS
– WE'RE ALWAYS LOOKING FOR THE BEST

Faculty Positions
Infectious Diseases Research

The Public Health Research Institute (PHRI) of Rutgers New Jersey Medical School located in Newark, New Jersey, is recruiting two faculty members at the middle or senior levels to join a growing group of 20 laboratories. PHRI (www.phri.org) is a leading infectious diseases research center that emphasizes basic and translational sciences. Candidates must have training and experience of the highest quality and a funded research program addressing critical questions in cell biology, immunology and molecular biology that offer novel insights into pathogenicity, as well as innovative approaches for new vaccines, therapeutics and diagnostics. Preference will be given to programs focused on major viral and bacterial pathogens, and immunology of the lung. We will only consider candidates who have current long term NIH grant support or equivalent funding from other sources. PHRI is housed in a state-of-the-art research facility that has extensive core services, including a nationally-designated BL3 laboratory and animal facilities and an X-ray facility for structural studies. PHRI offers a robust and highly collegial research environment, generous start-up funds, and a comprehensive benefits package. Candidates should submit a curriculum vitae, a statement of research interests and accomplishments and a list of at least three references.

Any questions or applications should be sent to: **Dr. Barry Kreiswirth, Public Health Research Institute, New Jersey Medical School, Rutgers Biomedical and Health Sciences, 225 Warren Street, Newark, NJ 07103. Tel: (973) 854-3240; Fax: (973) 854-3101; Email: kreiswba@njms.rutgers.edu**

Rutgers, the State University of New Jersey, is an Equal Opportunity/Affirmative Action employer, and is compliant with the Americans with Disabilities Act (ADA). For more information, please visit http://recruitment.rutgers.edu/TheRUCommitment.htm. Women and minorities are encouraged to apply.

RUTGERS
New Jersey Medical School

THE UNIVERSITY OF MAINE™

The University of Maine's School of Marine Sciences invites applications for three, tenure-track Assistant Professorships:

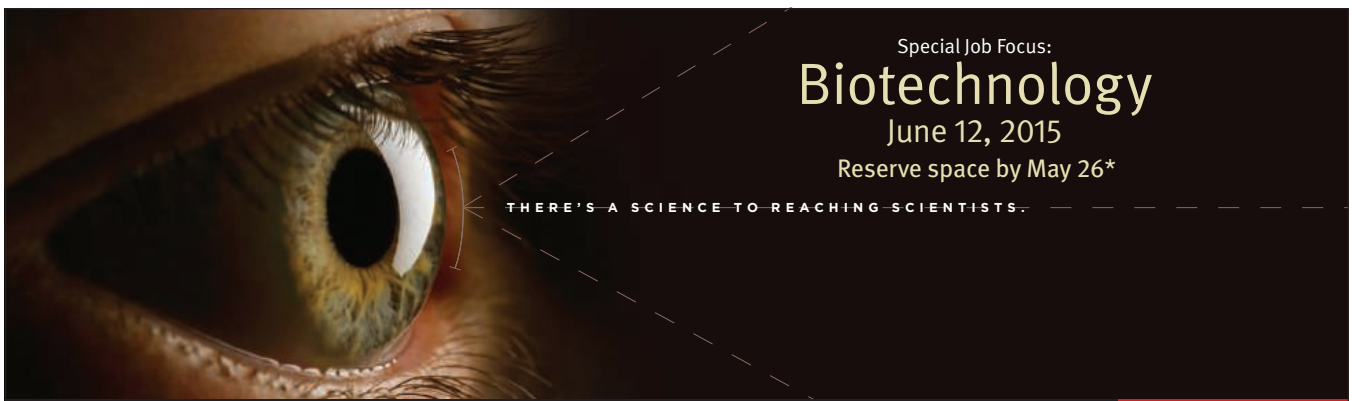
Marine Policy: We seek a scientist in any social science discipline whose research applies to the governance of marine and/or coastal systems. S/he will be expected to engage in basic and applied research with other scientists and stakeholders.

Marine Mammals: We seek experts in one of the following areas: marine mammal biology and physiology, health, ecotoxicology or population ecology, marine mammal/fisheries interactions, or environmental drivers and threats to marine mammal populations.

Marine Physiologist: We seek a scientist who works with any group of marine organisms in the fields of comparative physiology, stress physiology, developmental physiology, and/or ecological physiology.

Development of strong, externally funded research programs, contributions to undergraduate and graduate instruction, and advising of graduate students is expected for all positions. For full announcements and contact information, and to submit an application, please visit **https://umaine.hiretouch.com**. Review of applications will commence as specified (Marine Policy: July 1, 2015; Marine Mammals: July 15, 2015; Marine Physiology: August 1, 2015) and continue until positions are filled.

The University of Maine is an EEO/AA Employer. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, national origin, sexual orientation, age, disability, protected veteran status, or any other characteristic protected by law.



Why choose this biotechnology section for your advertisement?

- Relevant ads lead off the career section with special Biotechnology banner
- Bonus distribution to:

  BIO International Convention
  June 15–18, 2015, Philadelphia, PA

  BIO Career Fair
  June 18, 2015, Philadelphia, PA.

* Ads accepted until June 8 on a first-come, first-served basis.

SCIENCECAREERS.ORG

ScienceCareers
FROM THE JOURNAL SCIENCE AAAS

To book your ad: **advertise@sciencecareers.org**

The Americas: 202-326-6582 **Europe/RoW:** +44-0-1223-326500
Japan: +81-3-3219-5777 **China/Korea/Singapore/Taiwan:** +86-186-0082-9345

GREEN TALENTS

International Forum for High Potentials in Sustainable Development

DO YOU HAVE WHAT IT TAKES?

Are you an up-and-coming scientist with original ideas and a strong focus on sustainable development? Does your research have the potential to change the world? We challenge you to convince our high-ranking expert jury and become one of the 25 Green Talents of 2015!

As the theme of Germany's Science Year 2015 is the "City of the Future" we very much welcome applications from candidates who work in and around this topic. However, the competition is open to all subject areas and offers equal chances of success.

MORE INFORMATION AT **WWW.GREENTALENTS.DE**
PLEASE SUBMIT YOUR ONLINE APPLICATION BEFORE
02 JUNE 2015, 12 PM CET.

AN INITIATIVE OF THE

Federal Ministry of Education and Research

FONA Research for Sustainable Development BMBF

TEXAS BIOMEDICAL RESEARCH INSTITUTE

EMPLOYMENT OPPORTUNITY
FACULTY POSITION – VIROLOGY & IMMUNOLOGY

The Department of Virology & Immunology at the Texas Biomedical Research Institute located in San Antonio, TX, invites applications for a faculty-level position at the **ASSISTANT SCIENTIST** level equivalent to an Assistant Professor. The Department has faculty with research programs focusing on hepatitis viruses, human and simian immunodeficiency viruses, and emerging viruses such as hemorrhagic fever viruses. Major strengths of the Institute are the extensive nonhuman primate resources, which include macaques, marmosets, chimpanzees and baboons.

There is a strong pre- and postdoctoral training program and a close association with the University of Texas Health Science Center at San Antonio, including a role in graduate education. All candidates must have as a minimum a doctoral degree in the biological sciences or an M.D. degree, and have completed at least two years of relevant postdoctoral research. Please include a CV with letter outlining research experience and career goals with application.

All areas of virology will be considered. The Department is particularly interested in a person with expertise in host-pathogen interactions with an innovative program in molecular pathogenesis or cellular immunology.

Apply online at **http://www.txbiomed.org/about/employment**. Application packets are accepted electronically or in hard copy. A completed application packet is a requirement for all positions. Incomplete applications will not be accepted.

EOE

David Geffen School of Medicine

Chair, Department of Neurology

The University of California Los Angeles invites applicants for the position of Professor and Chair, Department of Neurology, David Geffen School of Medicine (DGSOM). Reporting to the Vice Chancellor for Health Sciences, the Chair will provide vision, leadership and strategic direction in meeting the educational, research, and clinical missions of the Department. Responsibilities include overall management, academic planning, budget, personnel, resource allocation and program development.

Candidates must have an outstanding record of leadership, research and clinical excellence, and a demonstrated commitment to education. Additional requirements include a proven track record of management in academics, national leadership in professional organizations, national recognition for scholarship, M.D. degree or equivalent, eligibility for California medical licensure, and documented experience and expertise in mentoring junior faculty.

The Department of Neurology at the DGSOM is relentless in its pursuit of innovation, strategic growth and success. Founded by Augustus S. Rose, M.D. in 1951, the department has grown to its current size with 106 faculty with primary appointments, 11 with secondary appointments, 5 active emeriti faculty, and 59 voluntary faculty throughout the local region. The department is integrated with seven affiliated hospitals including Harbor/UCLA Medical Center, Olive View/UCLA Medical Center, Cedars-Sinai Medical Center, the Greater Los Angeles Veterans Administration Medical Center, and Charles Drew University. These affiliations provide the ability to serve a diverse community throughout the region.

The department is organized into disease-specific and method-specific programs, including all of the major categories of neurological diseases. The department enjoyed a #1 ranking in NIH funding for nine consecutive years and currently is in the top five nationally. The faculty lead comprehensive clinical programs at the Ronald Reagan UCLA (RRUCLA) and Santa Monica UCLA Medical Centers. *US News & World Report* has recognized RRUCLA as Best in the West and one of the top five Best Hospitals in the nation, and it recognized UCLA in the top 10 for best adult neurology/neurosurgery care. The department has a strong tradition in the development of clinician-scientists and is home to 125 trainees. The Neurology residency training program is rated in the top 10 nationally and attracts applicants from the finest institutions in the nation. The faculty are educational leaders who chair many of the courses at national meetings and are the authors of many noted textbooks on subdisciplines in neurology.

Confidential review of applications, nominations and expressions of interest will begin immediately and continue until an appointment is made. Compensation for the position is highly competitive. All qualified candidates, including women and minorities, are encouraged to apply.

Electronic submission of materials is preferred. A letter of interest, curriculum vitae and the names of 3 references should be submitted online to: **https://recruit.apo.ucla.edu/apply/JPF00997**

Alan M. Fogelman, M.D.
Search Committee Chair
Tel: 310-825-6058
Email: afogelman@mednet.ucla.edu

The University of California is an Equal Opportunity/ Affirmative Action Employer. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, national origin, disability, age or protected veteran status. For the complete University of California nondiscrimination and affirmative action policy see: UC Nondiscrimination and Affirmative Action Policy.

POSITIONS OPEN

FACULTY POSITIONS-MEDICAL SCHOOL

The Saint James School of Medicine, an international medical school (**website: http://www.sjsm.org**), invites applications from candidates with teaching and/or research experience in any of the basic medical sciences for its Caribbean campuses. Faculty positions are currently available in Pathology, Histology, and Anatomy. Applicants must be M.D., D.O., and/or Ph.D. Teaching experience in the U.S. system is desirable but not required. Retired persons are encouraged to apply. Attractive salary and benefits. Submit curriculum vitae electronically to **e-mail: jobs@mail.sjsm.org** or online at **website: http://www.sjsm.org.**

Southwest Jiaotong University, P.R.China Anticipates Your Working Application

Southwest Jiaotong University (SWJTU), founded in 1896, situates itself in Chengdu, the provincial capital of Sichuan. It is a national key multidisciplinary "211" and "985 Feature" Projects university directly under the jurisdiction of the Ministry of Education, featuring engineering and a comprehensive range of study programs and research disciplines spreading across more than 20 faculties and institutes/centers. Boasting a complete Bachelor-Master-Doctor education system with more than 2,500 members of academic staff, our school also owns 2 first-level national key disciplines, 2 supplementary first-level national key disciplines (in their establishment), 15 first-level doctoral programs, 43 first-level master programs, 75 key undergraduate programs, 10 post-doctoral stations and more than 40 key laboratories at national and provincial levels.

Our university is currently implementing the strategy of "developing and strengthening the university by introducing and cultivating talents". Therefore, we sincerely look forward to your working application.

More information available at http://www.swjtu.edu.cn/

I. Positions and Requirements

A.High-level Leading Talents

It is required that candidates be listed in national top talents programs such as ***Program of Global Experts, Top Talents of National Special Support Program, "Chang Jiang Scholars", China National Funds for Distinguished Young Scientists and National Award for Distinguished Teacher.*** Candidates are supposed to be no more than 50 years old. The limitation could be extended in the most-needed areas of disciplinary development. Candidates who work in high-level universities/institutes and reach the above requirements are supposed to be no more than 45 years old.

B. Young Leading Scholars

Candidates are supposed to be listed in or qualified to apply for the following programs:

- ***National Thousand Young Talents Program***
- ***The Top Young Talents of National Special Support Program (Program for Supporting Top Young Talents)***
- ***Science Foundation for the Excellent Youth Scholars***

Candidates should have good team spirit and leadership, outstanding academic achievements, broad academic vision and international cooperation experience and have the potential of being a leading academic researcher.

C. Excellent Young Academic Backbones

Candidates under 40 years old are expected to graduate from high-level universities/institutes either in China or other countries. Those who are professors, associate professors and other equal talents from high-level universities/institutes overseas could be employed as professors and associate professors as well.

D. Excellent Doctors and Post Doctoral Fellows

Candidates under 35 years old are supposed to be excellent academic researchers from high-level universities either in China or other countries.

II. Treatments

The candidates will be provided with competitive salaries and welfares that include settling-in allowance, subsidy of rental residence, start-up funds of scientific research, assistance in establishing scientific platform and research group as well as international-level training and promotion . As for outstanding returnees, we can offer further or specific treatments that can be discussed personally.

III . Contact us:

Contacts: Ye ZENG & Yinchuan LI
Telephone number: 86-28-66366202
Email: ***talent@swjtu.edu.cn***
Address: Human Resources Department of SWJTU, the western park of high-tech zone, Chengdu, Sichuan, P.R.China, 611756

"National Thousand (Young) Talents Program", "Chang Jiang Scholars Program" and Other Faculty Positions Available in Nankai University

Located in the city of Tianjin, 30-minute away from Beijing by high-speed train, Nankai University is one of the key national universities directly under the jurisdiction of the Ministry of Education of China, and also belongs to "211 Project" and "985 Project". It is considered as the center for both education and academic research with a series of excellent achievements. In terms of the quality and the quantity of research papers, projects, funds, and prizes, Nankai University has been at the forefront of the national universities in China. Premier Enlai Zhou, the world-wide known mathematician Shiing-shen Chern, physicist Ta-you Wu and playwright Yu Cao are all alumni of Nankai University.

Nankai University is providing the following honorable positions for outstanding talents:

1. Professors and Associate Professors of "The National Thousand (Young) Talents Program", "The State (Young) Special Support Plan", and other high-level talent programs: In addition to the requirements defined by the programs, like "The National Thousand Talent Program" (http://www.1000plan.org/), the applicants with good health conditions should be well-established and highly innovative scientists with strong academic records and leadership. The applicants for the Young Programs should be able to demonstrate their potential to be outstanding scientists in the future with the support of Nankai University.

2. Distinguished Professors and Visiting Professors of "Chang Jiang Scholars Program": In addition to the requirements defined by the program (http://www.changjiang.edu.cn/), the applicants with good health conditions should be internationally known scholars with excellent achievements in their research fields, strong leadership in guiding a first-class research team and high capability in organizational management.

3. Hundred Young Academic Leaders Program of Nankai University: This program targets excellent young scholars, less than 40-year-old, in humanity, social sciences and natural sciences fields, breaking the limitation of professional title. The applicants can apply for this program before or after officially working in Nankai University and the selected scholars will be provided all-round support, including performance-based pay of the highest level of professors, research fund, experimental and working conditions.

4.Other positions (Professor/Associate Professor/Lecturer/Postdoctoral Researcher/Visiting Professor)

Pease visit ***http://www.nankai.edu.cn/s/12/t/27/64/72/info25714.htm*** for further details.

Salary, start-up package and benefits: The recruited faculty at different academic levels will be supported with competitive salary, the start-up package (competitive start-up funds, newly renovated office/lab and experienced assistants), housing allowance, medical insurance and other possible benefits. All of the above offers are negotiable.

Contact us: Applicants should send their curriculum vitae in both English and Chinese, the first page of 5 publications, statement of research interests/plans and at least three references to: Ms. Yang and Mr. Wang, Office of Human Resources, Nankai University, 94 Weijin Road, Tianjin, China, 300071; Tel(Fax): 0086-22-2350-8595;
Website: http://rsc.nankai.edu.cn; Email: nkuniversity@nankai.edu.cn.

"Hundred Young Academic Leaders Program" is now open for application, and please visit ***http://recruitment.nankai.edu.cn/webhr/login_nk.jsp*** for more information.You will be contacted after we receive your application.

"Nankai University is an Equal Opportunity Employer."

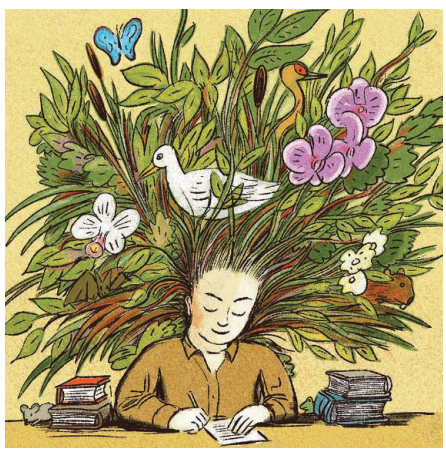
By **Gretchen Meyer**

Playing the field

I grew up in the New York suburbs, a world of small lawns and scattered parks, but I spent most of my summers in rural upstate New York, roaming the woods. The creatures I found there fascinated me, so I went to college to study biology. One summer, I spent 2 weeks at a field station—Great Gull Island in Long Island Sound, off the coast of Connecticut—banding terns. My time there planted the seed that matured into a rewarding career and, ultimately, took me back to the woods.

Great Gull Island is owned by the American Museum of Natural History (AMNH). It's a former military base used mainly to study the common and roseate terns that nest there every summer. It isn't far offshore, but to me it felt remote. The boat came twice a week, on Fridays to bring food and weekend visitors and on Sundays to take the weekenders home. The buildings were old, with no electricity or running water. I didn't care.

After graduating from college, I drifted a while, then returned to New York and took a waitressing job. I wanted to get back to biology, so I volunteered at AMNH. Great Gull Island was normally used only during summer, but I learned about a fall project studying bird migration along the Atlantic coast. I quit my job and returned to the island.



"I lived in the rainforest and woke to the sound of howler monkeys."

During the day, I removed birds from mist nets and banded and released them. At night, I played cards with the other staff members by candlelight. I watched the seasons change. Every evening at dusk, a snowy owl emerged from the old gun emplacement where it hunkered down during the day and flew off over the ocean.

I enrolled in a master's degree program at the Yale School of Forestry and Environmental Studies. While I was there, I took a semester off to work at another field station, on Barro Colorado Island in Panama, a former hilltop that was marooned when the Panama Canal flooded the area. I lived in the rainforest and woke to the sound of howler monkeys. I spent days watching lizards and recording their behavior. Once, I returned to my room at the end of the day and found it invaded by army ants.

I moved to Cornell University for a Ph.D. studying interactions between herbivorous insects and host plants. Then came a search for faculty positions, which led me to a teaching post at a liberal arts college. What I enjoyed most about that work, I found, was taking students outside: getting them on snowshoes, sending them out to catch grasshoppers, taking them canoeing. But full-time teaching had never been my dream, and when I saw an ad for a field station job, I applied.

Now I'm manager of the University of Wisconsin-Milwaukee Field Station in Saukville, Wisconsin. I've been here 15 years. It's the perfect job. I have a research program. I teach and advise. I facilitate research projects and support the teachers who bring classes here. I participate in the management of the land, at the field station and other university-owned areas: a regionally significant wetland, a virgin prairie remnant, an abandoned iron mine that's an important site for hibernating bats.

My work is structured by seasons. As the snow melts, I track bud break and leaf development. In summer and fall, I administer workshops on topics ranging from creative writing to sedge and grass identification. Winter means hikes on the frozen wetlands that, at nearly 900 hectares, are a large part of the property. I've surveyed rare orchids and an endangered dragonfly, participated in prescribed burns at our prairie site, and taken students to watch through night vision telescopes as bats exit from our hibernaculum. I've watched sandhill cranes court outside my office window and seen them later shepherding their growing chicks.

Experiences like these aren't just scientific; they're natural, and human. I've found a career that allows me to have such experiences and also share them with others. I like to think I'm helping inspire the next generation of students, as I was inspired by my own field station visits years ago. ■

Gretchen Meyer is manager and staff biologist at the University of Wisconsin-Milwaukee Field Station. For more on life and careers, visit ScienceCareers.org. Send your story to SciCareerEditor@aaas.org.

ILLUSTRATION: ROBERT NEUBECKER