PUBLIC HEALTH

# The Tobacco Strategy Entrenched

Carl F. Cranor

Ever wonder why it has been so slow and difficult to reduce the health risks from tobacco, secondhand smoke, lead, beryllium, or chromium? David Michaels's excellent *Doubt Is Their Product* provides part of the explanation, showing numerous ways in which "the product defense industry" uses scientific (and pseudoscientific) arguments to undermine public health protections, corrupt the scientific record, and mislead the public.

**Doubt Is Their Product**
How Industry's Assault on Science Threatens Your Health
***by David Michaels***
Oxford University Press, Oxford, 2008. 384 pp. $27.95, £14.99. ISBN 9780195300673.

The book's title announces its central theme. A tobacco industry strategy memo argues, "Doubt is our product since it is the best means of competing with the 'body of fact' that exists in the minds of the general public. It is also the means of establishing a controversy." (*1*) The aim: to sow doubt in the minds of the public, judges, and even regulatory scientists (if they are susceptible) about the scientific basis for greater public health or environmental protections (think global warming) or tort law actions. Because of the tobacco industry's success in obfuscating, slowing, reducing, and blocking regulatory actions, its approach has been adopted by others, has become institutionalized in presidential administrations, and has been used as talking points by some politicians. Fostering doubt and controversy and demanding high degrees of certainty postpone legal actions, keep products in commerce longer, and perhaps delay improved protections indefinitely. They can also leave the public or work force at risk.

As Michaels (an epidemiologist at George Washington University) explains, this clever strategy permits people to oppose public health rules without arguing the policy point and without being labeled anti–public health. It also uses a common science term that might resonate with some in the scientific community. Scientific articles usually note uncertainties about the research subject and the need for further studies.

Industries and their supporters have also demanded "proof" (more at home in mathematics than science) before agencies can increase public health protections or plaintiffs can successfully receive tort law compensation for injuries suffered. However, even though scientists may not understand all aspects of a problem, public health agencies need to act on the weight of the best science available at the time.

The doubt strategy is most at home in postmarket legal contexts. Public health agencies face the burden of establishing scientific and legal cases that will withstand appellate court scrutiny before they can successfully provide increased health and environmental protections or withdraw drugs or pesticides from the market. Companies that emphasize scientific uncertainties appear to be scientific angels; they only seek to preserve the integrity and certainty of the relevant fields against hasty regulatory action based on incomplete science.

Tort plaintiffs face similar burdens. They must show that the defendant's products or actions more likely than not can and did cause injuries from which the plaintiff suffers. The U.S. Supreme Court's *Daubert* decision requires judges to review the scientific basis of expert testimony; they may bar litigants from trial if the science is insufficient (*2*). Fostering doubt may sway judges, who are typically less informed about science than are regulatory scientists. The product defense industry has also helped to persuade some judges that they should review and exclude scientific studies individually without reviewing the total body of relevant evidence on which scientists rely, a most unscientific way to review the basis of expert testimony.

The book presents examples of product defense experts who have accepted funding to reach predetermined conclusions, misrepresented scientific claims, hidden their affiliations, written articles while using others' names, or had scientific papers ghost-written by lawyers.

What should be done? Among the author's recommendations is to require the testing of chemicals before workers and the public are exposed. If all products were subject to premarket testing for safety and impartial agency review before commercialization, this removes some incentives to raise doubt about the science. Drug and pesticide manufacturers rarely point out that their science is too uncertain to permit their products into the market. Michaels might have said more about additional legal changes that would reduce the influence of the doubt and uncertainty arguments, e.g., shifting legal burdens to the manufacturer once its product's safety was called into question.

Michaels also recommends a number of disclosures: of any and all research sponsors,

CREDIT: CEDENO KEN/CORBIS SYGMA

**Desire for doubt.** In April 1994, the CEOs of several tobacco companies told a committee of the U.S. House of Representatives that "Cigarette smoking is not addictive." In January 1998, their successors (shown) would only admit "Under some definitions cigarette smoking is addictive."

The reviewer, the author of *Toxic Torts: Science, Law, and the Possibility of Justice*, is at the Department of Philosophy, University of California Riverside, 900 University Avenue, Riverside, CA 92521, USA. E-mail: carl.cranor@ucr.edu

of what manufacturers know about the toxicity of their product (with penalties for covering up or lying), and of hazards in the public's midst (like community right-to-know laws). These recommendations are not panaceas but make good first steps.

In addition, *Doubt Is Their Product* reminds one of deeper risks that threaten scientific fields and democratic deliberation. When science affects commercial interests, there are substantial temptations for researchers or their employers to substitute the ethics of the marketplace for the ethics of careful, objective evaluation of the data to understand the world, environmental threats, and health risks. Such substitution can result in the corruption of the scientific literature and the breaking of incremental links in chains of evidence on which researchers and the public depend, and it also tends to undermine properly informed political and judicial decisions. The scientific community and the public need to be on guard against such abuses; Michaels's history of these events sounds an alert that must not be ignored.

**References**

1. Brown and Williamson Tobacco Company, *Smoking and Health Proposal* (Brown and Williamson document no. 680561778–1786, 1969); http://legacy.library.ucsf.edu/tid/mvs40f00.
2. *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579 (1993).

10.1126/science.1162339

THE GONZO SCIENTIST

## Chasing the Biggest Shadow of All

Choosing which extreme sport to pursue in one's life is difficult. Most people are content with the likes of bungee jumping, ice climbing, or street luge, but not scientists. In addition to thrills, they want their sport to produce useful data. I tried out an extreme scientific sport last month: eclipse chasing. The objective is to take very sensitive equipment to very remote locations, very punctually.

Online
**sciencemag.org**
For more on this episode, go to www.gonzoscientist.org

The roots of the sport go back to ancient China, where astronomers experienced, in the words of the late television anchorman Jim McKay, both "the thrill of victory" (prestige in the emperor's court) and "the agony of defeat" (beheading for miscalculation). Eclipse chasing has come a long way since then (more data, less beheading)—and made headlines around the world in 1919. On 29 May that year, after struggling with biting insects and tropical storms on a volcanic island, a British team recorded starlight bent around the eclipsed Sun by gravity, an observation that was widely trumpeted as confirming Albert Einstein's theory of relativity.

CREDIT: RAY JAYAWARDHANA

For my first taste of eclipse chasing, I joined a team of scientists (*1*) hoping for a rendezvous with an eclipse 1 August in the wild west of Mongolia. To get to the site, we made a night trek over the Altai Mountains, which nearly killed us when our driver nodded off at the wheel. On the day itself, we worked in the intense heat and dust of the Gobi desert, which actually did kill a telescope motor and camera. But just as the eclipse was getting started, I drove a few kilometers away with astrophysicist Ray Jayawardhana, to take part in a shamanistic ritual that involved a hundreds-strong chorus of screaming, shouting, and clapping at the sky. We found ourselves surrounded by terrified Mongolian locals convinced that a monstrous god called Rah was eating the Sun. But that is another story.

While Rah captured the Sun, our team captured gigabytes of data. Like astro-paparazzi, we harvested hundreds of digital images through a pair of telescopes—a refractor and a reflector fixed to a motor-driven astrograph built by team member Kosmas Gazeas—during the 2 hours of partial and 2 minutes of total eclipse. And we weren't the only ones ogling the darkened sky. A team led by Jay Pasachoff, an astrophysicist at Williams College, Massachusetts, was observing to the north in Siberia (*2*). And to the south, *Science*'s Beijing correspondent, Richard Stone, was watching in western China with researchers from the National Astronomical Observatories, Chinese Academy of Sciences, and other institutions (*3*).

But how useful are all those data? With orbiting telescopes like Hinode trained on the Sun—and capable of creating their own eclipse anytime by simply occluding the Sun's photosphere with a metal disk, can ground-based observation add anything? "I get that question all the time," comments Pasachoff. In fact, he says, data produced by earthly eclipse chasers are more valuable than ever. The space telescopes, put in place at enormous cost, provide only part of the picture. By design, "the spacecraft can't observe a huge region around the Sun, the whole inner and middle corona." Studying the dynamics of these superhot solar gases should lead to better modeling of solar wind and answer a nagging riddle: Why is the corona hundreds of times hotter than the Sun's surface? Not only are the eclipse chasers equipped with "more modern and efficient" charge-coupled device cameras, explains Pasachoff, but "the resolution on the corona that we get by processing eclipse images is finer than that obtainable by any spacecraft." To understand the Sun, astronomers still need the Moon to cover it.

Pasachoff, who has seen 47 solar eclipses, wants to rename the sport. Rather than a chaser, "I am an eclipse preceder," he says. After all, successfully predicting and getting to the site of an eclipse is the name of the game. For next year's eclipse, a blockbuster event in the International Year of Astronomy (*4*), record numbers of people are expected to chase—or rather, precede—the 22 July solar eclipse in Asia. So prepare your telescopes and book your tickets now.

**References and Notes**

1. The team comprised four astrophysicists—Ray Jayawardhana (University of Toronto), Kosmas Gazeas (Harvard-Smithsonian Center for Astrophysics), Kazuhiro Sekiguchi (National Astronomical Observatory, Japan), and Katrien Kolenberg (University of Vienna)—and remote-sensing researcher Tuvjargal Norovsambuu (National University of Mongolia).
2. www.williams.edu/astronomy/eclipse/eclipse2008.
3. R. Stone, *Science* **321**, 759 (2008).
4. www.astronomy2009.org.

**–JOHN BOHANNON**

10.1126/science.1164877

MEDICINE

# Life Cycle of Translational Research for Medical Interventions

From the initial discovery of a medical intervention to a highly cited article is a long road, and even this is not the end of the journey.

Despina G. Contopoulos-Ioannidis,[1] George A. Alexiou,[2] Theodore C. Gouvias,[2] John P. A. Ioannidis[2,3,4]*

Despite a major interest in translational research (*1*–*3*), development of new, effective medical interventions is difficult. Of 101 very promising claims of new discoveries with clear clinical potential that were made in major basic science journals between 1979 and 1983, only five resulted in interventions with licensed clinical use by 2003 and only one had extensive clinical use (*4*). Drug discovery faces major challenges (*5*–*8*). Moreover, for several interventions supported by high-profile clinical studies, subsequent evidence from larger and/or better studies contradicts their effectiveness or shows smaller benefits (*9*). The problem seems to be even greater for nonrandomized studies (*9*). Here, we present the results of an empirical evaluation of the life-cycle phases of translational research for selected medical interventions.

We examined key milestones in the life cycle of translational research for all the interventions claimed to be effective in at least one study that received over 1000 citations in the literature in 1990–2004, on the basis of the Web of Science. These are the most-cited papers in the literature of medical interventions (*10*). Because they have received the greatest attention, they provide easily identifiable scientific milestones. Citation counts are a widely accepted coinage of recognition. Of course, several blockbusters may go through an industrial discovery–testing–production process that does not involve any particular highly cited paper in the peer-reviewed literature. In these cases, it is not as clear-cut to isolate one or a few studies that are indisputable milestones in the translational process.

Of 49 articles with >1000 citations, we excluded articles where the intervention was ineffective, as well as those assessing management strategies rather than specific interventions, and we selected only the earliest article whenever two or more highly cited studies with >1000 citations had been published on the same intervention and indication. Thirty-two interventions for specific indications were thus evaluated, and we could place the milestone of when their first highly cited clinical study was published showing effectiveness (tables S1 and S2). We considered this an important time point in the translational process and estimated how long a time ("translation lag") it had taken from the initial discovery of each intervention to reach that point. Highly cited status does not necessarily mean that these interventions continue to be considered as effective as proposed in the original highly cited papers. By the end of 2006, the effectiveness of 19 interventions had been replicated by other subsequent studies ($n$ = 14) or had remained unchallenged ($n$ = 5), whereas the other 13 had been either contradicted ($n$ = 5) or found to have had initially stronger effects ($n$ = 8) when larger or better controlled subsequent studies were performed (table S1).

### Translation Lag

To place each discovery in time, we identified the year when the earliest journal publication on preparation, isolation, or synthesis appeared or the earliest patent was awarded (whichever occurred first). Overall, the median translation lag was 24 years (interquartile range, 14 to 44 years) between first description and earliest highly cited article (see the chart, page 1299). This was longer on average (median 44 versus 17 years) for those interventions that were fully or partially "refuted" (contradicted or having initially stronger effects) than for nonrefuted ones (replicated or remaining unchallenged) ($P$ = 0.004).

In a secondary analysis, we defined the time of discovery as the first description (publication or awarded patent) of any agent in the wider intervention class (those with similar characteristics and mode of action). Early translational work may be performed with different agents in the same class compared with those that eventually get translated into postulated high-profile clinical benefits. Analyses using information on the wider class of agents showed even longer translation lag, with median of 27 (interquartile range, 21 to 50) years and similar prolongations of the translation lag for refuted interventions.

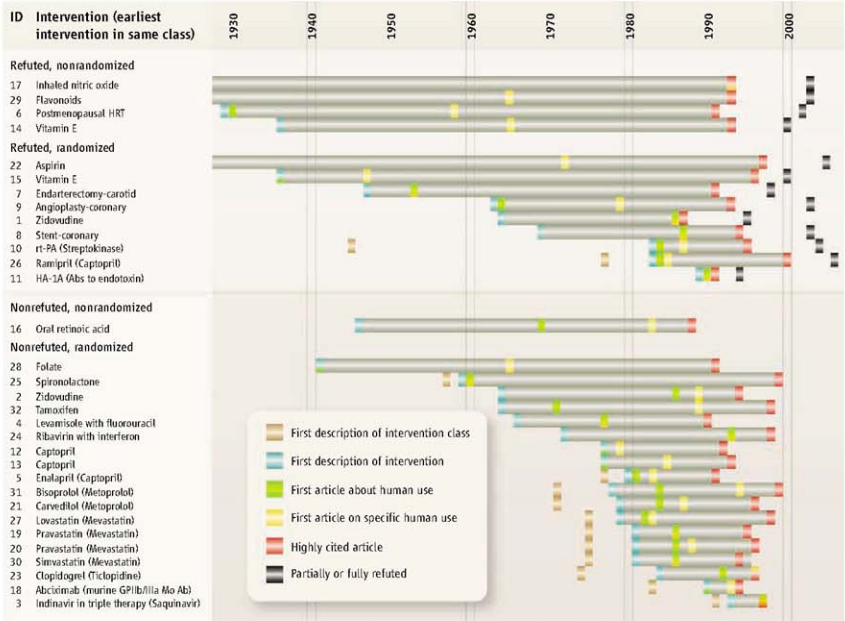
Among the 18 nonrefuted interventions that had a highly cited randomized trial to support them, the median translation lag was 16.5 years (range 4 to 50 years) in the main analysis [22 years (range 6 to 50 years) considering the wider class]. The fastest successful translation occurred for indinavir (as part of triple antiretroviral therapy) and abciximab, both of which took only 4 years from their patenting to the publication of a highly cited randomized trial. Both of these fast successes involved multidisciplinary work spanning molecular to clinical research on protease inhibitors and integrins, respectively.

We also tried to identify the first published article that described the use of each intervention in humans and the first published article that described the use of each intervention in humans for the specific intervention described eventually in the highly cited study (*11*). There was a very large variability in the timing of the first human study and of the first human study for the specific indication (see the chart, page 1299). The range for the time from first discovery to first human use was 0 to 28 years. The range for the time from first discovery to first specific human use was 0 to 221 years.

We observed that most highly cited claims that were eventually refuted had a very slow translation history preceding them [e.g., flavonoids, vitamin E, and estrogens were already available for many decades before observational (nonrandomized) studies claimed implausibly large survival benefits in the 1990s]. We conclude that claims for large benefits from old interventions require extra caution as they are likely to be exaggerated. Given the considerable refutation rate of even the most highly cited interventions, extensive replication and confirmation of proposed treatment benefits are indicated. New drug discovery is probably essential for common diseases where the existing drug armamentarium has been already extensively screened. Conversely, for uncommon and neglected diseases, the existing drug options may remain largely untested, and old drugs may find interesting new uses (*12*–*14*).

[1]Department of Pediatrics, University of Ioannina School of Medicine, Ioannina, 45110, Greece, and Department of Pediatrics, George Washington University School of Medicine and Health Sciences, Washington, DC 20037, USA. [2]Clinical Trials and Evidence-Based Medicine Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, 45110, Greece. [3]Department of Medicine, Tufts Medical Center, and Institute for Clinical Research and Health Policy Studies, Tufts University School of Medicine, Boston, MA 02111, USA. [4]Biomedical Research Institute, Foundation for Research and Technology–Hellas, Ioannina 45110, Greece.

*Author for correspondence (at the address in footnote 2). E-mail: jioannid@cc.uoi.gr; john.pa.ioannidis@gmail.com




**Milestones for the 32 interventions.** First description of agent in wider class, tan box (when the agent used in the highly cited article is not the same as the first described in its class); first description, cyan box; first human-use article, green box; first specific–human use article, yellow box; earliest highly cited publication, red box; realization of full or partial refutation (for contradicted or initially stronger effects), black box. Whenever two or more milestones coincide in the same year, the respective colors are superimposed on that box. Folate, flavonoids, and vitamin E were already in human use at the time of first description. Extending beyond the illustrated time range were the first description for nitric oxide in 1772 and its first human use in 1800; and the first description of flavonoids in 1898, aspirin in 1853, and of the wider class of antiendotoxins in 1896. Details for these interventions can be found in tables S1 to S5, listed by the ID number. Ab, antibody; GP, glycoprotein; HA-1A, human IgM monoclonal antibody against endotoxin A; HRT, hormone replacement therapy; mo Ab, monoclonal antibody; rt-PA, recombinant tissue plasminogen activator.

### Recommendations for the Future

Our analysis documents objectively show the long length of time that passes between discovery and translation. As scientists, we should convey to our funders and the public the immense difficulty of the scientific discovery process. Successful translation is demanding and takes a lot of effort and time even under the best circumstances; making unrealistic promises for quick discoveries and cures may damage the credibility of science in the eyes of the public. The following are some recommendations for improving the system, based on our analyses:

• Discovery of new substances and interventions remains essential, but proper credit and incentives should be given to accelerate the testing of these applications in high-quality, unbiased clinical research and the replication of claims for effectiveness.

• Multidisciplinary collaboration with focused targets and involving both basic and clinical sciences should be encouraged.

• Proof of effectiveness for new interventions requires large, robust randomized clinical trials.

• Translational efforts for common diseases should focus more on novel agents and new cutting-edge technologies; for these ailments, it is unlikely that genuine major benefits from interventions already known for a long time have gone unnoticed.

### References and Notes

1. E. A. Zerhouni, *JAMA* **294**, 1352 (2005).
2. F. M. Marincola, *J. Transl. Med.* **1**, 1 (2003).
3. J. P. Ioannidis, *J. Transl. Med.* **2**, 5 (2004).
4. D. G. Contopoulos-Ioannidis, E. Ntzani, J. P. Ioannidis, *Am. J. Med.* **114**, 477 (2003).
5. P. Cuatrecasas, *J. Clin. Invest.* **116**, 2837 (2006).
6. G. Duyk, *Science* **302**, 603 (2003).
7. B. Booth, R. Zemmel, *Nat. Rev. Drug Discov.* **3**, 451 (2004).
8. D. G. Hackam, D. A. Redelmeier, *JAMA* **296**, 1731 (2006).
9. J. P. Ioannidis, *JAMA* **294**, 218 (2005).
10. Methods and details for the collection and analysis of data are available as supporting material on *Science* Online along with its supplementary tables S1 to S5. These interventions included 18 drugs, two monoclonal antibodies, one hormonal therapy, four vitamins or food products, and three surgical or device interventions. Three drugs and one vitamin appear two times each in the list, as they were used for two different indications. For more information and references, see tables S1 and S2.
11. For more details, see Methods in the supporting online material and tables S3 to S5
12. C. R. Chong, D. J. Sullivan, *Nature* **448**, 645 (2007).
13. S. Zhu *et al.*, *Nature* **417**, 74 (2002).
14. J. D. Rothstein *et al.*, *Nature* **433**, 73 (2005).

10.1126/science.1160622

**Supporting Online Material**
www.sciencemag.org/cgi/content/full/321/5894/1298/DC1

GENETICS

# Enhancing Gene Regulation

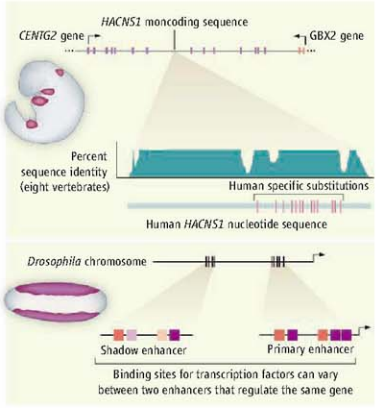
Bioinformatic approaches reveal functional changes and the evolution of regulatory sequences that control gene expression.

Gregory A. Wray and Courtney C. Babbitt

Nearly half a century has passed since François Jacob and Jacques Monod demonstrated that specific noncoding sequences are required to activate genes that metabolize lactose in the bacterium *Escherichia coli* (*1*). In a prescient observation, they noted that mutations in these regulatory sequences might play a role in the evolution of organismal traits. They further argued that gene function is not only based on the biochemical activity of its product but also on how the gene's expression is regulated. This idea was expanded in 1975 in an influential paper by Mary-Claire King and Alan Wilson (*2*), who proposed that trait differences between humans and chimpanzees are primarily due to regulatory changes in gene expression. Decades elapsed, however, before it was feasible to begin testing these ideas in detail. Two papers in this issue, by Prabhakar *et al.* on page 1346 (*3*), and by Hong *et al.* on page 1314 (*4*), demonstrate the power of combining bioinformatic approaches with experimental tests to characterize such regulatory regions.

A major impediment to studying the evolutionary importance of mutations in regulatory regions is simply knowing where to look. DNA sequences that regulate the transcription of genes occupy no fixed position relative to coding DNA regions and are often diffuse and widely dispersed. Even when the position of a regulatory element is known, there is the added challenge of identifying which mutations have functional consequences. Within coding sequences, the genetic code imposes familiar regularities: Mutations that change protein structure can be identified exhaustively and unambiguously. By contrast, identifying functional mutations within regulatory regions requires experimental tests of putative regulatory elements from different species or individuals—a costly and time-consuming process. Bioinformatic methods offer a way to identify promising functional noncoding regions and to narrow the focus for experimental tests.

One approach is to search genomes for highly conserved blocks of noncoding sequence (on the assumption that conservation implies function) and then scan for instances of rapid sequence divergence on just one branch of a phylogeny (which implies a functional change in a single species) (*5*). Prabhakar *et al.* use this approach to identify a noncoding region they call *human-accelerated conserved noncoding sequence 1* (*HACNS1*) (see the figure). To test the function of this region, they genetically engineered mouse embryos to express a construct composed of human *HACNS1*, the promoter element of a heat shock gene, and a reporter gene. Their results show that human *HACNS1* drives expression in the mesenchyme of the early developing forelimb, and later developing hindlimb, in these mouse embryos. A comparison of expression patterns driven by macaque, chimpanzee, and human orthologs of *HACNS1* revealed that consistently strong forelimb expression is a unique property of the human version. By testing various combinations of human and chimpanzee *HACNS1* sequences, the authors narrowed down the relevant functional mutations to an 81–base pair region containing 13 substitutions that arose during human evolution. This concentration of substitutions is highly unusual relative to the genome as a whole, implying positive selection on this region during human origins.




**Identification of enhancer elements.** (**Top**) *HACNS1* is a noncoding region of conservation (percent identity) among eight vertebrate species, with 13 human-specific substitutions (vertical red lines). *HACNS1* drives the expression of a reporter gene (purple) in the limbs of a developing mouse embryo. (**Bottom**) Primary enhancers (red and purple boxes) near a gene are conserved relative to more distant "shadow" enhancers, which appear to be less functionally constrained (pink and light purple boxes). These enhancers drive gene expression (purple) in the *Drosophila* embryo.

What genes does *HACNS1* regulate? This conserved region lies within an intron of *CENTG2*, which encodes a guanosine triphosphatase activating protein that regulates endosomes (membrane-bound vesicles that transport materials into a cell). It is also ~300 kb downstream of the next nearest gene, *GBX2*, which encodes a transcription factor that is expressed, among other locations, in developing limbs. If *GBX2* is indeed the target of *HACNS1* regulation, the implications are fascinating. Because of *GBX2*'s role in limb development, the authors note that changes in its expression could have altered human limb anatomy—producing, for instance, specializations of the hand that facilitate tool use, or modifications of the foot associated with bipedalism.

Another bioinformatic approach to identifying regulatory elements is to search for clusters of potential transcription factor binding sites (*6*). Hong *et al.* examined data from chromatin immunoprecipitation combined with microarray technology—so-called ChIP-chip analysis—from the fruit fly (*Drosophila melanogaster*) genome for targets of regulation by the transcription factor Dorsal and known cofactors (*4*). Surprisingly, they found that many target genes of Dorsal contain not one but two clusters of transcription factor binding sites, implying the presence of multiple regulatory regions with similar function (see the figure). Although most experimentally verified enhancers lie within a few kilobases of the gene they regulate, some of the

Department of Biology and Institute for Genome Science and Policy, Duke University, Box 90338, Durham, NC 27708, USA. E-mail: gwray@duke.edu

CREDIT: ADAPTED BY P. HUEY/SCIENCE

secondary enhancers lie tens of kilobases away. The authors used transgenic flies to show that both nearby and more distant clusters of binding sites drive similar patterns of reporter gene expression for two genes, *brk* and *sog*.

The evolutionary role of these "shadow enhancers" is unclear. One possibility suggested by Hong *et al.* is that they provide an opportunity for natural selection to tinker with regulatory sequences, while the primary enhancers maintain essential gene function. Consistent with this hypothesis, the sequences of "shadow enhancers" evolve more rapidly than those of primary enhancers, which suggests that they operate under fewer functional constraints. The intriguing possibility that "shadow enhancers" more commonly confer expression differences among species could be tested with the type of comparative experimental approach that Prabhakar and colleagues applied to *HACNS1*.

Although identifying functional changes in regulatory sequences remains a serious challenge, these two papers demonstrate the power of combining bioinformatics and experimental tests. However, most regulatory elements are neither highly conserved among species nor composed of clusters of the same binding motif. Indeed, few of the well-documented cases connecting noncoding mutations to trait evolution in humans and flies (*7*–*9*) involve regions that would have been identified as functional on the basis of sequence conservation or motif clustering.

The challenge now is to develop methods that can recognize functional changes within a much greater proportion of regulatory elements. As phylogenetic sampling of sequenced genomes grows denser, it is becoming possible to carry out unbiased surveys of regulatory change based on genome-scale functional assays (*10*), quantitative genetics (*11*), and tests for positive selection (*12*). Decades after Jacob and Monod first speculated about the evolutionary importance of regulatory mutations, we are in a position to begin testing their ideas in earnest.


**References**

1. F. Jacob, J. Monod, *J. Mol. Biol.* **3**, 318 (1961).
2. M. C. King, A. C. Wilson, *Science* **188**, 107 (1975).
3. S. Prabhakar *et al.*, *Science* **321**, 1346 (2008).
4. J.-W. Hong, D. A. Hendrix, M. S. Levine, *Science* **321**, 1314 (2008).
5. S. Prabhakar, J. P. Noonan, S. Pääbo, E. M. Rubin, *Science* **314**, 786 (2006).
6. M. Markstein, P. Markstein, V. Markstein, M. S. Levine, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 763 (2002).
7. N. Gompel, B. Prud'homme, P. J. Wittkopp, V. Kassner, S. B. Carroll, *Nature* **433**, 481 (2005).
8. N. S. Enattah *et al.*, *Nat. Genet.* **30**, 233 (2002).
9. M. T. Hamblin, A. Di Rienzo, *Am. J. Hum. Genet.* **66**, 1669 (2000).
10. A. P. Boyle *et al.*, *Cell* **132**, 311 (2008).
11. E. E. Schadt *et al.*, *Nature* **422**, 297 (2003).
12. R. Haygood, O. Fedrigo, B. Hanson, K. D. Yokoyama, G. A. Wray, *Nat. Genet.* **39**, 1140 (2007).


10.1126/science.1163568

ASTRONOMY

# The Universe Measured with a Comb

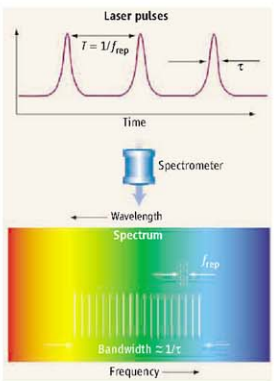
A technique for wavelength calibration promises to revolutionize observational astrophysics, in areas including planet searches and cosmology.

**Sebastian Lopez**

In 1962, Allan Sandage predicted that an expanding universe should cause a drift in the redshift of cosmological objects, but noted: "With present optical techniques there is apparently no hope of detecting such small changes in redshifts for time intervals smaller than $10^9$ years" (*1*). Future extremely large telescopes (with diameters of 30 to 40 m), equipped with powerful spectrometers, could in principle enable such a measurement. However, measuring a systematic change in radial velocity of only 1 cm $s^{-1}$ per year over the course of about 20 years—a measurement referred to as the Sandage-Loeb experiment (*2*–*4*)—would still be impossible if it were not for the recent development of a new and exquisite wavelength calibration technique called "laser frequency combs" (*5*). On page 1335 of this issue, Steinmetz *et al.* apply this technique for the first time to an astrophysical experiment (*6*), and the results look promising.

The Doppler effect provides astronomers with a precise method to measure radial velocities of stars and galaxies using the observed shift in wavelength (or frequency) of their spectral features relative to laboratories on Earth: The higher the radial velocity, the stronger the effect. And when the light entering the spectrometer comes from distant objects like galaxies or quasars—thus crossing cosmological distances to reach the telescope—their spectra provide information about the geometry and history of the universe as a whole.

Because the universe expands, distant objects can always be assigned with a redshift, a quantity that cosmologists relate to distance and time by fitting various parameters to cosmological models. The past decade has seen a series of breakthroughs in cosmology. The Wilkinson Microwave Anisotropy Probe (WMAP) mission delivered images of the cosmic microwave background that support a




**The basics of a laser frequency comb.** A mode-locked laser creates femtosecond pulses at gigahertz frequencies, $f_{rep}$ (**top**), that are synchronized with an atomic clock. A spectrum of the pulses (**bottom**) is composed of many modes that are uniformly spaced in wavelength (or frequency) and cover a spectral bandwidth given roughly by the inverse of the pulse duration. Each mode's wavelength (or frequency) does not have to be measured, but instead is given by a mathematical relation that includes $f_{rep}$, known a priori with very high accuracy. Laser frequency combs could therefore become the perfect wavelength calibration technique for astrophysical experiments that require high accuracy and long-term stability.

CREDIT: ADAPTED BY P. HUEY/SCIENCE

Departamento de Astronomía, Universidad de Chile, Casilla 36-D, Santiago, Chile. E-mail: slopez@das.uchile.cl

flat universe, and studies measuring distances to Type Ia supernovae or the large-scale distribution of galaxies, among others, have established that the universe not only expands, but that the expansion is accelerating—presumably due to the effect of an unknown component in the mass-energy budget of the universe called "dark energy." Many projects aim to elucidate what dark energy really is, but all of them rely on a given cosmological model; only the Sandage-Loeb experiment could track the history of the expansion directly, without any previous assumption on the geometry of the universe.

What has been the impediment to fully exploiting present instrument capabilities? Traditional spectral calibration techniques use a crowd of emission or absorption lines at known laboratory wavelengths as reference to map the detector pixels into wavelengths. However, calibration units are subject to uncertainties that unavoidably degrade the wavelength solution: Lines are not evenly distributed in the spectral range of interest, have a wide range of intensities, and sometimes appear blended. These systematic effects become the perennial stumbling block for precision spectroscopy. They limit the capabilities of current high-resolution spectrometers and hinder experiment repeatability, crucial for any long-term monitoring.

The recently developed laser frequency combs (*3*, *6*–*9*) may offer the solution. Such a comb is the spectrum of a femtosecond "mode-locked" laser that delivers pulses at repetition rates $f_{rep}$ of ~1 GHz (determined by the round-trip time in the laser cavity). When these pulses pass through a spectrometer, a regular train of modes is produced in the frequency domain, each of them evenly separated by $f_{rep}$ (see the figure) and spanning a spectral bandwidth given by the inverse of the pulse duration. Because time—and thus frequency—is the most accurately measured quantity in physics thanks to atomic clocks, each mode's frequency (or wavelength) is accurately known a priori and can be used as a perfect ruler to calibrate astronomical spectra.

Steinmetz *et al.* now report the first astronomical spectrum (of the Sun) calibrated with a laser frequency comb. Besides slightly outperforming current best standards of accuracy using just a small bandwidth, the team was also able to characterize the stability of the instrument in an unprecedented fashion. Use of larger bandwidths should allow wavelengths to be measured with a stable precision of 1 part in 10 billion, opening a new era in astronomical spectroscopy.

Full implementation of laser frequency combs in large telescopes will require coverage of the entire optical range. Once this challenge is overcome, at least two other astrophysical experiments besides the future Sandage-Loeb test should benefit from the use of this technique.

First, some astronomers have wondered whether the atomic physics responsible for the redshifted absorption lines seen in the spectra of distant quasars has remained the same over cosmological times. The values of fundamental constants or combinations of constants—like the proton-to-electron mass ratio or the fine-structure constant—determine the relative positions of the lines in the quasar spectra. Thus, one could in principle compare the value of those constants then ("at high redshift") and now (on Earth) to determine whether they have remained constant. By choosing particular methods and sets of lines, different groups have arrived at diverging conclusions. After a decade of research, the debate is now centered on the systematic effects inherent to the observations. Laser frequency combs could help to identify the origin of these systematic effects.

Second, the precision and stability offered by laser frequency combs could greatly help astronomers looking for exoplanets (which orbit stars other than the Sun). Such planets imprint small changes on the radial velocity of their solar system, and monitoring the radial velocities of bright stars has thus become the most reliable way of finding planets. However, the smaller the planet, the smaller the drift in radial velocity. Discovering Earth-like planets orbiting solar-like stars in the "habitable zone" (where life could exist) requires a precision of about 5 cm $s^{-1}$ and a stability of about 1 year. This should be an easy task for this new technique.

**References and Notes**


1. A. Sandage, *Astrophys. J.* **136**, 319 (1962).
2. A. Loeb, *Astrophys. J.* **499**, L111 (1998).
3. M. T. Murphy *et al.*, *Mon. Not. R. Astron. Soc.* **380**, 839 (2007).
4. J. Liske *et al.*, *Mon. Not. R. Astron. Soc.* **386**, 1192 (2008).
5. T. W. Hänsch and J. L. Hall received the 2005 Nobel Prize in Physics "for their contributions to the development of laser-based precision spectroscopy, including the optical frequency comb technique."
6. T. Steinmetz *et al.*, *Science* **321**, 1335 (2008).
7. Th. Udem, R. Woltzwarth, T. W. Hänsch, *Nature* **416**, 233 (2002).
8. C. Araujo-Hauck *et al.*, *Messenger* **129**, 24 (2007).
9. C.-H Li *et al.*, *Nature* **452**, 610 (2008).


10.126/science.1163194

MEDICINE

# The Cart Before the Horse

**Janet D. Rowley[1] and Thomas Blumenthal[2]**

Chimeric RNAs, transcribed from malignancy-associated chromosomal translocations, can also arise from RNA trans-splicing in normal cells.

Since the identification of specific regions in human chromosomes that undergo recurring structural rearrangements (translocations) and cloning of the associated breakpoint genes, the fusion of genes has been viewed as a unique event in abnormally growing, usually malignant, cells (*1*, *2*). However, the study by Li *et al.* on page 1357 in this issue (*3*) indicates quite the contrary, turning at least one paradigm of cancer cytogenetics on its head.

Li *et al.* report that in normal human endometrial tissue, there is a low amount of a messenger RNA (mRNA) that corresponds to sequences from two genes, *JAZF1* on chromosome band 7p15 and *JJAZ1/SUZ12* on chromosome band 17q21. Moreover, this chimeric mRNA is identical to that seen in 50% of human endometrial stromal sarcomas, in which there is a 7;17 chromosomal translocation that results in a gene fusion, even though no translocation is detected in normal endometrial cells. The product encoded by the chimeric mRNA is a fusion protein that is expressed in cultured cells, and consequently, could confer cellular resistance to programmed cell death and increased growth (under conditions where expression of the endogenous *JJAZ1* gene was suppressed). Further, this fusion mRNA is expressed in a cyclical manner in normal endometrial cells, most readily detected at the beginning and end of the menstrual cycle when concentrations of estrogen and progesterone are low.

How is this fusion mRNA made in the absence of a corresponding gene fusion?

[1]Department of Medicine, University of Chicago, 5841 South Maryland Avenue, MC 2115, Chicago, IL 60637, USA. E-mail: jrowley@medicine.bsd.uchicago.edu [2]Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309, USA. E-mail: tom.blumenthal@colorado.edu

Li *et al.* propose that the fusion mRNA is produced by trans-splicing of RNA (see the figure) in which nucleotides at the 3′ end of *JAZF1*-encoding precursor mRNA are replaced with those of *JJAZ1*-encoding precursor mRNA. A different form of trans-splicing is common in several lower animal phyla (*4*), and a few mRNAs have been shown to be assembled from separate transcripts in insects (*5*). And although trans-splicing in mammalian cells has been reported, the resulting chimeric RNAs do not perform obvious functions and are usually not present in large enough amounts to do so.

It has not been clear how two separate mRNAs are spliced together (*6*). Perhaps they are brought together through the pairing of nucleotides within noncoding sequences (introns) between two transcripts, or maybe each contains a binding site for proteins that can form dimers or higher-order multimers. Although sloppiness by the spliceosome—the cellular machine that removes introns from precursor RNA—could be an explanation, only specific pairs of precursor mRNAs engage in trans-splicing. Perhaps RNAs from different genes are trans-spliced because they are transcribed in the same geographic location. Alternatively, trans-splicing could occur more frequently than we realize, but most cases go undetected.

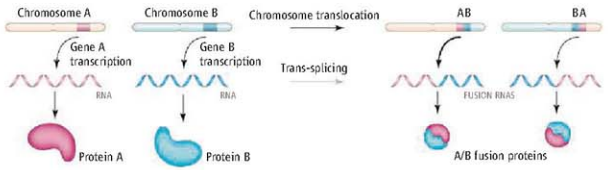
What makes the study by Li *et al.* especially interesting is that trans-splicing is clearly regulated. The mRNA fusion appears only in cells from endometrial tissue. Its expression is increased by hormones and hypoxia, with much higher expression in late secretory and early proliferative stages of the menstrual cycle. The key question is whether the chimeric RNA is transcribed from some undetected rearranged copies of the two genes. However, Li *et al.* show that there is no such gene rearrangement in cells producing the trans-spliced mRNA and that this trans-splicing event can be duplicated in vitro. In addition, they show that a nontransformed human endometrial stromal cell line had no rearranged DNA or visually abnormal chromosomes.

Given the absence of any detectable rearranged DNA in cells producing the chimeric RNA, the obvious explanation is rearrangement at the RNA level. To demonstrate that trans-splicing could account for the chimera, Li *et al.* made extracts from a human endometrial stromal cell line and from a rhesus monkey fibroblast cell line so that they could detect trans-spliced products by a sequence difference between the RNA from the two species. The authors demonstrated in vitro trans-splicing of the rhesus *JAZF1* exons (coding regions of DNA) to human *JJAZ1* exons. Treatment of the rhesus RNA with deoxyribonuclease (to cleave any DNA that might be present) did not prevent formation of the chimeric RNA, confirming that chimeric RNA arose from trans-splicing.

Is it a coincidence that the same RNA occurs in normal cells by trans-splicing and in tumor cells of the same type by DNA rearrangement? The authors suggest the intriguing possibility that whatever leads to the trans-splicing could also lead to the genomic rearrangement. This could occur by at least three general mechanisms. The same sequences could pair at the RNA level to result in trans-splicing and at the DNA level to result in genomic rearrangement. However, sequence analyses of translocation breakpoints in leukemia reveal large deletions and duplications as well as precise nucleotide base-pairing (*7*). Alternatively, genomic rearrangement could be a direct result of the trans-splicing event if genes involved in the rearrangement are brought into close proximity during the RNA trans-splicing process. This idea is consistent with recent reports on the existence of "factories" for transcription and RNA processing (*8*–*10*). Finally, the RNA created by trans-splicing could act as a guide RNA to facilitate the genomic rearrangement, an idea for which there is precedent (*11*). In this case, the trans-spliced RNA would anneal to regions of both of the chromosomes and guide them in a DNA recombination event. Indeed, it is possible that other genomic rearrangements could be guided by cellular RNAs.

If fusion mRNA is widespread, it could explain the conundrum that has long perplexed cancer geneticists: why fusion mRNAs can be detected in apparently normal tissues of healthy people. Such fusions involve common translocations seen in neoplastic hematopoietic cells, but never in solid tumors. If fusion mRNAs are part of normal cell function, then finding fusions of the immunoglobulin heavy chain gene (*IGH*) to the *BCL2* gene in normal spleens, which usually reflects the presence of a t(14;18) translocation in lymphomas, would not be unexpected (*12*). Given that *IGH* and the *MYC* genes frequently colocalize in transcription factories (*9*), this geography could provide a mechanism for having nascent RNAs in juxtaposition; moreover, the genes themselves would be close together. Translocations involving the *IGH* and *IGK/L* genes and the genes encoding T cell receptors (*TCRs*) in lymphoid malignancies are exceptions in that they do not lead to a fusion mRNA, but rather to altered regulation of the apparently normal target protein (*13*). Presumably all translocations are mediated by DNA recombination enzymes, but could this process be guided by RNA produced by trans-splicing? The study by Li *et al.* also raises questions relevant to clinical practice. Potent therapies targeting fusion mRNA and proteins may disrupt critical pathways of normal cell function. Increasingly sensitive methods to determine the presence of a few translocation-bearing cells leads one to question whether translocations or normal cell products are being detected. This is a critical issue because the search for minimal residual disease is in high gear, especially for chronic myeloid leukemia that responds to the drug imatinib, as a "cure" seems within reach (*14*). Many patients suffering from this cancer are translocation-negative on standard cytogenetic analysis, but show a gene fusion (*BCR-ABL*) by reverse transcriptase polymerase chain reaction. For these patients, especially the ones with very low amounts of fusion, it is unclear whether what is being detected is a malignant cell or a trans-splicing event.

As the search for fusions in normal cells will likely be fast-paced for the next few years, two points should be considered, given




**Fusion RNAs.** Either chromosome translocation or RNA trans-splicing can give rise to fusion mRNAs and proteins. Some chromosomal translocations produce two hybrid genes that may produce mRNAs containing the 5′ end of one gene and the 3′ end of the other. Both may encode fusion proteins. Alternatively, normal mRNAs corresponding to both genes can recombine by trans-splicing that may produce equivalent fusion mRNAs and proteins. Only one of the possible fusion mRNAs and proteins is examined by Li *et al.*

CREDIT: ADAPTED BY P. HUEY/*SCIENCE*

the findings of Li *et al*: cell specificity and regulation of the trans-splicing event. So choose the fusion to be investigated, mindful of these constraints.


**References**

1. J. D. Rowley, *Nat. Rev. Cancer* **1**, 245 (2001).
2. F. Mitelman, B. Johansson, F. Mertens, *Nat. Genet.* **36**, 331 (2004).
3. H. Li, J. Wang, G. Mor, J. Sklar, *Science* **321**, 1357 (2008).
4. T. Blumenthal, in *Wormbook*, The *C. elegans* Research Community, Ed.; Wormbook, 10.1895/wormbook.1.5.1, www.wormbook.org.
5. R. Dorn, G. Reuter, A. Loewendorf, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 9724 (2001).
6. T. Takahara *et al.*, *Mol Cell.* **18**, 245 (2005).
7. Y. Zhang, J. D. Rowley, *DNA Repair (Amst).* **5**, 1282 (2006).
8. S. McCracken *et al.*, *Nature* **385**, 357 (1997).
9. C. S. Osborne *et al.*, *PLoS Biol.* **5**, e192 (2007).
10. J. A. Mitchell, P. Fraser, *Genes Dev.* **22**, 20 (2008).
11. M. Nowacki *et al.*, *Nature* **451**, 153 (2008).
12. S. Janz, M. Potter, C. S. Rabkin, *Genes Chromosome Cancer* **36**, 211 (2003).
13. T. W. McKeithan, *Semin Oncol.* **17**, 30 (1990).
14. A. Hochhaus *et al.*, *Blood* **111**, 1039 (2008).


10.1126/science.1163791

BIOCHEMISTRY

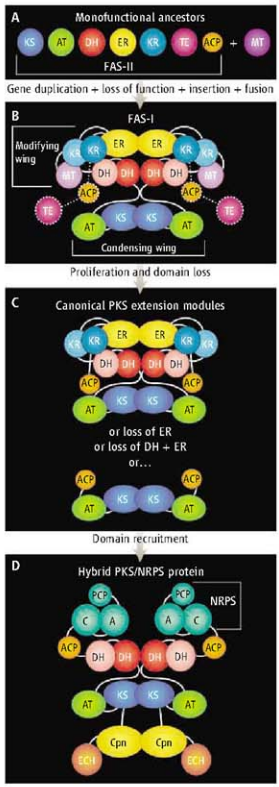
# An Enzyme Assembly Line

Fatty acid synthases and related megaenzymes are highly adaptable to new functions as a result of their modular architecture.

**Janet L. Smith[1,2] and David H. Sherman[1,3]**

The fundamental polymers of biology—proteins, DNA, and RNA—are products of repetitive condensation of simple amino acid or nucleotide building blocks and are comparatively easy to assemble. However, other biomolecules require additional reactions beyond condensation of building blocks. Examples are the fatty acids and the polyketide and nonribosomal peptide secondary metabolites. These molecules are produced by complex enzyme assembly lines that include multiple catalytic domains. Two new crystal structures—one reported recently (*1*), the other by Maier *et al.* on page 1315 of this issue (*2*)—enrich our understanding of how these mega-enzymes function as efficient factories to produce a remarkable range of metabolic products.

Maier *et al.* study the fatty acid synthase (FAS-I) responsible for de novo fatty acid synthesis in the cytosol of animal cells. FAS-I is homologous in sequence and architecture with the very large family of modular polyketide synthases (PKSs), which produce a wide variety of natural products with potential medicinal value. In 2006, the authors reported the structure of FAS-I from a 4.5 Å electron density map, in which most domains could be assigned but no details were visible (*3*). Their new structure provides sufficient detail to understand the fold and the connectivity of six of its eight domains; however, the flexibly tethered acyl carrier protein (ACP) and thioesterase (TE) domains remain invisible. The other structure, reported by Tanovic *et al.* (*1*), is of an intact module of a nonribosomal peptide synthetase (NRPS). Except for the peptidyl carrier protein (PCP), the primary NRPS domains are not related to those of the FAS and PKS systems, but the assembly-line approach is similar.




All data indicate that the ancestor of FAS-I was a set of monofunctional enzymes, presumably resembling the dissociated FAS (type II, FAS-II) that catalyzes fatty acid biosynthesis in modern bacteria and plant plastids (see the figure, panel A). Gene duplication, loss of function, and gene fusion gave rise to the 270-kD polypeptide that functions as the homodimeric FAS-I in mammals (see the figure, panel B). A different mega-enzyme fusion of monofunctional ancestors evolved in fungi (*4*). Of these two assembly-line architectures for fatty acid synthesis, the mammalian FAS-I proved the more adaptable, and it now exists not only for fatty acid biosynthesis but throughout the eubacterial and fungal world for synthesis of polyketides. Indeed, high-resolution structures of PKS components (*5*, *6*) provide critical corroboration for the mammalian FAS structure.

Two key features of the FAS-I architecture explain its remarkable adaptability. First, the structure is segregated into two wings: a selection/condensing wing for addition of new building blocks, and a modifying wing for chemical processing of chain elongation intermediates (see the figure, panel B). The heart of the assembly line is the condensing wing, where an acyltransferase (AT) domain selects

**Assembly-line proliferation.** The dissociated FAS-II **(A)** evolved into the homodimeric FAS-I **(B)** [dotted lines outline disordered parts of the FAS-I structure (*2*); lighter shades indicate inactivated DH, KR, and MT domains]. Duplication of an ancestral FAS-I gene, followed by selective deletion, yielded the canonical extension modules of modern PKS pathways **(C)**. An even richer diversity of polyketides arose by domain recruitment. For example, the CurF protein of the hybrid PKS/NRPS for curacin A (*10*) includes decarboxylase (ECH), cyclopropanase (Cpn), and NRPS domains **(D)**.

[1]Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, USA. [2]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA. [3]Departments of Medicinal Chemistry, Chemistry, and Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109, USA. E-mail: janetsmith@umich.edu; davidhs@umich.edu

a building block and a ketosynthase (KS) domain adds it to the growing chain. The dimeric KS also contributes most of the dimer contacts in the complex.

The second key architectural feature is an open and flexible design that is ideal for insertion or deletion of catalytic domains, especially in the modifying wing. Each two-carbon addition (via malonate) to a fatty acid chain is followed by three reactions—keto reduction (KR), dehydration (DH), and enoyl reduction (ER)—carried out in the modifying wing of the FAS-I. A major source of chemical diversity in polyketides arises from deletion or inactivation of one or more of these modifying domains (see the figure, panel C), providing the chemical variation that is lacking in fatty acids.

In FAS-I and most fungal PKSs, the assembly line is used for iterative synthesis: Each enzyme domain performs the same reaction at each extension step on the growing substrate. In contrast, in most bacterial PKSs, polyketide synthesis is sequential: Each extension step is carried out by an individual FAS-I–like "module," offering the possibility to vary the building block identity and modification chemistry at each step. This scheme greatly expands genetic and protein complexity. Several modules (up to 20 or more) are required to build a complex polyketide, and specific interactions of sequential modules must be faithfully maintained by fusion or by docking domains (*7*, *8*).

A big surprise of the new FAS structure is a vestigial methyltransferase (MT) domain at the periphery of the dimer, following the DH in the polypeptide sequence. Thus, the mega-enzyme ancestor of FAS-I appears to have had a methylation reaction as part of its fatty acid biosynthetic cycle. Was there a prokaryotic methyl branched-chain fatty acid, unknown to us today? The MT domain lost its function in FAS-I, was deleted from most PKS systems, but exists in some PKSs as an active methyltransferase. And herein lies a conundrum; the ubiquity of PKS pathways in bacteria and elsewhere strongly argues that the original FAS-I evolved in a prokaryote. However, other than *Mycobacterium tuberculosis* and related species that generate unusual fatty acids, we know of no modern prokaryote that uses a FAS-I for normal membrane lipid fatty acid biosynthesis (*9*).

In many PKS modules, the open FAS-I architecture has been augmented with a variety of other catalytic domains, such as *S*-acetyltransfer, halogenase, cyclopropanase, decarboxylase, and even entire NRPS modules (see the figure, panel D) (*10*, *11*). The new structure of the terminal module of the surfactin NRPS (*1*) shows how it, too, is highly adaptable. Like FAS-I, the NRPS has a solid platform for condensation, including an adenylation (A) domain to select the amino acid building block and a condensation (C) domain to form a peptide link to the growing chain (*1*). The monomeric C-A didomain (analogous to KS-AT in the FAS-I condensing wing) is fused to a PCP and a terminal TE domain. As in the FAS-I structure, the PCP is flexibly linked to the synthetase by tethers long enough for it to deliver substrate to the active sites of all catalytic domains. Unlike the FAS-I structure, the PCP and TE domains are well ordered in the NRPS module.

The three assembly line types use homologous domains (ACP or PCP) to carry the growing fatty acid, polyketide, or peptide via a pantetheine-linked thioester. The common thioester chemistry and the adaptable architecture have resulted in the proliferation of hybrid PKS-NRPS and even PKS-FAS-I pathways found in phylogenetically diverse bacteria (*9*, *12*). The rich diversity of PKS, NRPS, and hybrid systems demonstrates that nature has not employed a Henry Ford–like assembly line, from which the customer could have any color car so long as it was black. Rather, we see a modular assembly line that is easily copied, modified, and adapted to new function; this is the secret to its success.


**References**

1. A. Tanovic, S. A. Samel, L.-O. Essen, M. A. Marahiel, *Science* **321**, 659 (2008); published online 26 June 2008 (10.1126/science.1159850).
2. T. Maier, M. Leibundgut, N. Ban, *Science* **321**, 1315 (2008).
3. T. Maier, S. Jenni, N. Ban, *Science* **311**, 1258 (2006).
4. S. Jenni, M. Leibundgut, T. Maier, N. Ban, *Science* **311**, 1263 (2006).
5. A. T. Keatinge-Clay, R. M. Stroud, *Structure* **14**, 737 (2006).
6. Y. Tang, C. Y. Kim, I. I. Mathews, D. E. Cane, C. Khosla, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11124 (2006).
7. M. Thattai, Y. Burak, B. I. Shraiman, *PLoS Comput. Biol.* **3**, 1827 (2007).
8. J.-P. Nougayrède *et al.*, *Science* **313**, 848 (2006).
9. R. S. Gokhale, P. Saxena, T. Chopra, D. Mohanty, *Nat. Prod. Rep.* **24**, 267 (2007).
10. Z. Chang *et al.*, *J. Nat. Prod.* **67**, 1356 (2004).
11. L. Gu *et al.*, *Science* **318**, 970 (2007).
12. M. A. Fischbach, C. T. Walsh, J. Clardy, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 4601 (2008).


10.1126/science.1163785

MICROBIOLOGY

# How to Infect a Mimivirus

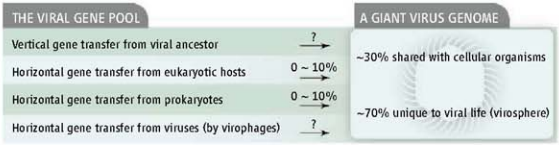
Large DNA viruses such as the giant Mimivirus can be infected by smaller viruses.

**Hiroyuki Ogata and Jean-Michel Claverie**

The giant DNA "Mimivirus" (*Acanthamoeba polyphaga* mimivirus, or APM) was initially mistaken for a bacterium, until La Scola *et al.* classified it as a virus in 2003 (*1*). This highly unusual virus has more genes than many bacteria (*2*), forms the most complex known virus particle (*3*), has a unique DNA delivery system (*4*), and encodes aminoacyl-tRNA synthetases (*5*), normally restricted to cellular organisms. As a possible "missing link" between the cellular and the viral world, APM's discovery revived theories that link DNA viruses to the emergence of the eukaryotic nucleus (*6*). Large viruses closely related to APM are abundant in the sea (*7*) and may play important roles in the geochemical fluxes that regulate Earth's climate. La Scola *et al.* now report in *Nature* (*8*) that the APM family has another unusual property: It is susceptible to infection by another virus, named Sputnik (after "traveling companion" in Russian).

Sputnik—a small icosahedral virus with a DNA genome encoding 21 genes—was isolated with a new strain of APM from a cooling tower in Paris. Attempts to culture Sputnik alone in amoeba cells were not successful. However, when amoebae were inoculated with the two viruses, both Sputnik and APM virions multiplied. La Scola *et al.* (*8*) show that Sputnik reproduces in the "virus factory," the replication and assembly center built by APM in amoeba cells during their lytic infection. The virus factory is a DNA-rich cytoplasmic compartment that appears 4 hours after APM infection and grows to several micrometers in diameter. Sputnik virions reproduce faster than do APM virions; 6 hours after infection, Sputnik virions start to emerge from the virus factory, while the new generation of APM virions only appears after 8 hours. Infection with both viruses decreases the yield of infective APM virions

Structural and Genomic Information Laboratory, CNRS-UPR 2589, IFR-88, Université de la Méditerranée, Parc Scientifique de Luminy, FR-13288 Marseille, France. E-mail: ogata@igs.cnrs-mrs.fr; jean-michel.claverie@univmed.fr




**Origin of genes in large eukaryotic viruses.** The distribution of sequence database matches suggests diverse origins for the genes of large DNA viruses. Horizontal gene transfer may occur through exposure to host or prokaryotic DNA. The many genes unique to viruses are vertically or horizontally transferred between viruses—a process in which the newly discovered virophages may play a key role.

and results in "sick" APM virions with aberrant morphologies. Sputnik thus behaves as a true parasite with a detrimental effect on APM reproduction.

Small viruses requiring other larger viruses for their reproduction have previously been documented. These "satellite viruses" lack essential functions for multiplication, for which they exploit their "helper viruses." La Scola *et al.* (*8*) argue that Sputnik is more than a satellite virus, because it uses its partner's virus factory and impairs its fitness. They therefore call Sputnik a "virophage."

What is the origin of the Sputnik virophage? The authors provide evidence suggesting the existence of related virophages in the oceans (*8*). Marine virologists have reported small viruses occurring with larger ones in marine protist populations (*9*, *10*). During recurrent infection of a cell by the two viruses, one virus may begin to benefit from the other. Like Sputnik, the small marine viruses multiply faster than the larger ones. If the viral genomes can physically interact, genes can be exchanged, and the two viruses may evolve into various states of dependency, from mutualisms to parasitism. In this context, it is worth noting that Sputnik has an integrase (an enzyme that inserts pieces of DNA from one DNA molecule into another). The genome of a marine virus, infecting the planktonic species *Emiliania huxleyi* contains a strange 176-kb central segment (*11*): Genes in this segment lack homologs in other viruses, but harbor a unique promoter. This segment is expressed much earlier than the rest of the viral genome and may be the integrated genome of an unknown virophage.

The genes in giant eukaryotic viruses have multiple origins (see the figure). The APM genome contains eukaryotic- or prokaryotic-like genes. Recent horizontal gene transfers from its eukaryotic hosts or prokaryotic organisms partially account for these genes. However, giant viral genomes also contain genes that are unique to viruses, the origin of which is hotly debated (*6*, *12*, *13*). Do these genes originate in vertical gene transfer from a very old viral common ancestor? The small number of genes shared among modern viruses argues against this possibility. Viral genome mosaicism is also suggested by the occurrence of very similar genes in different viruses (*14*). Furthermore, a substantial amount of horizontal gene transfer may occur between viruses. The Sputnik virophage now provides a new potential vehicle for such horizontal gene transfers. In fact, the Sputnik genome encodes several genes that may originate in vastly different viruses.

Assessing the proportions of vertical gene transfer and virus-virus horizontal gene transfer now appears crucial for understanding the evolution of giant viruses, refining the concept of virus lineage, and elucidating gene flow in the virosphere. The unusual features of the giant Mimivirus revived the popular, yet unresolved question: "Are viruses alive?" The discovery that some of them can get sick adds a new twist to this old debate.

**References**

1. B. La Scola *et al.*, *Science* **299**, 2033 (2003).
2. D. Raoult *et al.*, *Science* **306**, 1344 (2004).
3. P. Renesto *et al.*, *J. Virol.* **80**, 11678 (2006).
4. N. Zauberman *et al.*, *PLoS Biol.* **6**, e114 (2008).
5. C. Abergel *et al.*, *J. Virol.* **81**, 12406 (2007).
6. J.-M. Claverie, *Genome Biol.* **7**, 110 (2006).
7. A. Monier, J.-M. Claverie, H. Ogata, *Genome Biol.* **9**, R106 (2008).
8. B. La Scola *et al.*, *Nature* 10.1038/nature07218 (2008).
9. K. Nagasaki, *J. Microbiol.* **46**, 235 (2008).
10. C. P. Brussaard *et al.*, *Virology* **319**, 280 (2004).
11. M. J. Allen *et al.*, *J. Virol.* **80**, 7699 (2006).
12. H. Ogata, J.-M. Claverie, *Genome Res.* **17**, 1353 (2007).
13. E. V. Koonin, W. Martin, *Trends Genet.* **21**, 647 (2005).
14. K. Nagasaki *et al.*, *Appl. Environ. Microbiol.* **71**, 3599 (2005).

10.1126/science.1164839

PHYSICS

# An End to the Drought of Quantum Spin Liquids

**Patrick A. Lee**

After decades of searching, several promising examples of a new quantum state of matter have now emerged.

Electrons possess magnetic behavior through the quantum mechanical property of spin. The magnetic properties of materials then arise from the collective interaction of electrons on atoms within the crystal. Below a transition temperature, the electron spins of normal magnets "freeze" into an ordered array of magnetic dipoles. Whether the ordering is ferromagnetic (all the dipoles point in the same direction) or antiferromagnetic (the dipoles on adjacent sites point in opposite directions) is determined by the sign and strength of the interaction between the electrons. Early theoretical work has indicated a departure from these ordered states, suggesting that quantum mechanical fluctuations of the spin could be so strong that ordering would be suppressed and the spin ensemble would remain in a liquid-like state, even down to the lowest temperatures. Experimental evidence, which has until recently remained elusive, is emerging in favor of this long-predicted state of quantum matter.

**Ordered spins. (Left)** Néel's picture of antiferromagnet ordering with an alternate spin-up–spin-down pattern across the lattice. **(Right)** Quantum fluctuations lead to mutual spin flips, which Landau argued would disorder Néel's state.

To understand the controversy surrounding this exotic quantum spin liquid state, it is instructive to go back to the description of antiferromagnetism. Soon after the invention of quantum mechanics, Heisenberg pointed out that electron spins on neighboring atoms can have short-range interaction due to quantum mechanical exchange. Louis Néel

Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: palee@mit.edu

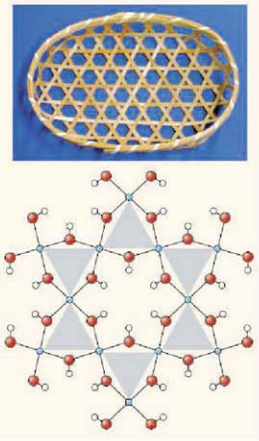
showed that a negative exchange interaction results in an antiparallel arrangement of neighboring spins, and that at low temperatures the spins order in an alternating up-down pattern (see the first figure, left panel) (*1*). This antiferromagnetic, or Néel, state was initially greeted with skepticism—Lev Landau was among the doubters (*2*).

Landau did not publish his objection, but on the basis of Friedel's comments (*2*) we can surmise that his argument was that the correct application of quantum mechanics to such a system would lead to fluctuations (see the first figure, right panel), which may completely randomize the spin order. It was not until 1949, when antiferromagnetic order was directly observed by neutron scattering, that Néel was fully vindicated. In fact, Néel's theory was far more successful than he had the right to expect: For 70 years the Néel state has prevailed.

In 1973, Philip W. Anderson reasoned that the best chance to defeat Néel ordering came in the form of "frustrated" spin systems (*3*), with the special example of spins on a triangular lattice. This lattice is "frustrated" because if an up- and down-spin occupy two corners of a triangle, the spin on the third corner does not know which way to point to obtain the lowest-energy configuration. Instead of a Néel state, Anderson proposed a ground state made up of a quantum mechanical superposition of singlet pairs that cover the lattice. He called this a resonating valence bond (RVB) state, an explicit example of a quantum spin liquid. Unfortunately, it was soon shown that neighboring spins in a triangular lattice manage to order at a 120° angle, and Néel once again won out.

The field of quantum spin liquids languished until 1987, when high-temperature superconductivity was discovered. Anderson pointed out a connection between the RVB spin liquid and the Cooper pairs of a superconductor (*4*). Attempts to justify the RVB theory of superconductivity led to rapid developments of the spin liquid theory. It is now understood that the spin liquid (defined as having an odd number of electron spins on each lattice unit cell) is a new state of matter with properties we have never encountered before. For example, the excited states may be spinons—charge-neutral objects that possess magnetic properties. Depending on the type of spin liquid, the spinon may obey Fermi or Bose statistics and there may or may not be an energy gap. Furthermore, these spinons cannot live by themselves but are generally accompanied by gauge fields, just as electrons are always accompanied by electromagnetic gauge fields (*5*). This is a dramatic example of emergent phenomena, where new particles and fields emerge at low-energy scales but are totally absent in the Hamiltonian that describes the initial system.

Confirmation of the existence of the spin liquid state has been elusive, and only recently have several promising examples surfaced. The first is an organic solid called κ-$(ET)_2Cu_2(CN)_3$ (*6*, *7*) in which the active ingredients are dimers of an organic molecule, ET [bis(ethylenedithio) - tetrathiafulvalene]. A single electron is localized on each dimer, which forms layers of approximately triangular lattices. Despite an exchange energy of ~250 K, no magnetic order was detected down to 30 mK. This material is an insulator but becomes a superconductor (critical temperature $T_c = 3.5$ K) and then a metal under pressure. It is believed that the proximity to an insulator-to-metal transition implies that the spins interact with a more complicated Hamiltonian than the Heisenberg model and allows the spin liquid state to form (*8*, *9*). Remarkably, the spin susceptibility goes to a constant at low temperatures and the specific heat is linear in temperature (*7*). These properties are normally associated with metals, being consequences of the electron Fermi surface. The linear specific heat is particularly unusual for an insulator that is relatively defect-free. Furthermore, the ratio of magnetic susceptibility to the linear temperature coefficient of the specific heat is close to that of free fermions. These observations strongly suggest that the excitations are indeed fermionic spinons that form a Fermi sea, and thus offer strong evidence for a spin liquid ground state.



**Meeting with frustration.** (**Top**) A Kagome basket. (**Bottom**) Structure of $ZnCu_3(OH)_6Cl_2$ (*10*) showing that the Cu ions (blue) occupy a Kagome lattice; O-H is red-white.

Last year an entirely different class of spin liquid was discovered. It has long been suspected that spins on a Kagome lattice support a spin liquid ground state. Kagome is the Japanese name for the weave pattern of a basket (see the second figure, top panel). The structure consists of corner-sharing triangles and is even more frustrated than the triangular lattice considered by Anderson. Last year saw the synthesis of such a solid-state Kagome system: $ZnCu_3(OH)_6Cl_2$, where a single electron spin resides on the Cu (*10*) (see the second figure, bottom panel). Although the exchange energy is ~200 K, this material does not show any magnetic ordering down to millikelvin temperatures. The magnetic excitations are apparently gapless, but unlike the organic compound, the large specific heat at low temperatures is sensitive to magnetic field, which suggests that the low-temperature properties may be dominated by a few percent of local moment defects.

Spin liquids are not limited to two-dimensional systems. A newly synthesized material, $Na_4Ir_3O_8$, has Ir ions that form a three-dimensional network of corner-sharing triangles, termed a hyper-Kagome structure (*11*). Despite an exchange energy of ~300 K, no magnetic order was found down to 1 K and below.

It is an exciting time in the history of antiferromagnetism. After decades of searching, three examples of the defeat of Néel order by quantum fluctuations have been discovered in quick succession. There are good reasons to believe that fermionic spinons will emerge as the low-energy excitations, but more work will be needed to confirm this. An even more intriguing question is whether their partner, the emergent gauge field, can make its presence felt as well. We can be optimistic that even more exciting discoveries lie ahead.


**References**

1. L. Néel, *Ann. Phys.* **5**, 232 (1936).
2. J. Friedel, *Phys. Today* **54**, 88 (October 2001).
3. P. W. Anderson, *Mater. Res. Bull.* **8**, 153 (1973).
4. P. W. Anderson, *Science* **235**, 1196 (1987).
5. P. A. Lee *et al.*, *Nat. Rev. Mod. Phys.* **78**, 17 (2006).
6. Y. Shimizu, K. Miyagawa, K. Kanoda, M. Maesato, G. Saito, *Phys. Rev. Lett.* **91**, 107001 (2003).
7. S. Yamashita *et al.*, *Nat. Phys.* **4**, 459 (2008).
8. O. Motrunich, *Phys. Rev. B* **72**, 045105 (2005).
9. S.-S. Lee, P. A. Lee, *Phys. Rev. Lett.* **95**, 036403 (2005).
10. J. Helton *et al.*, *Phys. Rev. Lett.* **98**, 107204 (2007).
11. Y. Okamoto *et al.*, *Nat. Phys. Rev. Lett.* **99**, 137207 (2007).


10.1126/science.1163196

# Flood or Drought: How Do Aerosols Affect Precipitation?

Daniel Rosenfeld,[1]* Ulrike Lohmann,[2] Graciela B. Raga,[3] Colin D. O'Dowd,[4] Markku Kulmala,[5] Sandro Fuzzi,[6] Anni Reissell,[5] Meinrat O. Andreae[7]

Aerosols serve as cloud condensation nuclei (CCN) and thus have a substantial effect on cloud properties and the initiation of precipitation. Large concentrations of human-made aerosols have been reported to both decrease and increase rainfall as a result of their radiative and CCN activities. At one extreme, pristine tropical clouds with low CCN concentrations rain out too quickly to mature into long-lived clouds. On the other hand, heavily polluted clouds evaporate much of their water before precipitation can occur, if they can form at all given the reduced surface heating resulting from the aerosol haze layer. We propose a conceptual model that explains this apparent dichotomy.

Cloud physicists commonly classify the characteristics of aerosols and clouds into "maritime" and "continental" regimes, where "continental" has become synonymous with "aerosol-laden and polluted." Indeed, aerosol concentrations in polluted air masses are typically one to two orders of magnitude greater than in pristine oceanic air (Fig. 1) (*1*). However, before humankind started to change the environment, aerosol concentrations were not much greater (up to double) over land than over the oceans (*1*, *2*). Anthropogenic aerosols alter Earth's energy budget by scattering and absorbing the solar radiation that energizes the formation of clouds (*3*–*5*). Because all cloud droplets must form on preexisting aerosol particles that act as cloud condensation nuclei (CCN), increased aerosols also change the composition of clouds (i.e., the size distribution of cloud droplets). This, in turn, determines to a large extent the precipitation-forming processes.

Precipitation plays a key role in the climate system. About 37% of the energy input to the atmosphere occurs by release of latent heat from vapor that condenses into cloud drops and ice crystals (*6*). Reevaporation of clouds consumes back the released heat. When water is precipitated to the surface, this heat is left in the atmosphere and becomes available to energize convection and larger-scale atmospheric circulation systems.

The dominance of anthropogenic aerosols over much of the land area means that cloud composition, precipitation, the hydrological cycle, and the atmospheric circulation systems are all affected by both radiative and microphysical impacts of aerosols, and are likely to be in a different state relative to the pre-industrial era.

## The Opposing Effects of Aerosols on Clouds and Precipitation

The radiative effects of aerosols on clouds mostly act to suppress precipitation, because they decrease the amount of solar radiation that reaches the land surface, and therefore cause less heat to be available for evaporating water and energizing convective rain clouds (*7*). The fraction of radiation that is not reflected back to space by the aerosols is absorbed into the atmosphere, mainly by carbonaceous aerosols, leading to heating of the air above the surface. This stabilizes the low atmosphere and suppresses the generation of convective clouds (*5*). The warmer and drier air thus produces circulation systems that redistribute the remaining precipitation (*8*, *9*). For example, elevated dry convection was observed to develop from the top of heavy smoke palls from burning oil wells (*10*). Warming of the lower troposphere by absorbing aerosols can also strengthen the Asian summer monsoon circulation and cause a local increase in precipitation, despite the global reduction of evaporation that compensates for greater radiative heating by aerosols (*11*). In the case of bright aerosols that mainly scatter the radiation back to space, the consequent surface cooling also can alter atmospheric circulation systems. It has been suggested that this mechanism has cooled the North Atlantic and hence pushed the Intertropical Convergence Zone southward, thereby contributing to the drying in the Sahel (*12*, *13*).

Aerosols also have important microphysical effects (*14*). Added CCN slow the conversion of cloud drops into raindrops by nucleating larger number concentrations of smaller drops, which are slower to coalesce into raindrops or rime onto ice hydrometeors (*15*, *16*). This effect was shown to shut off precipitation from very shallow and short-lived clouds, as in the case of

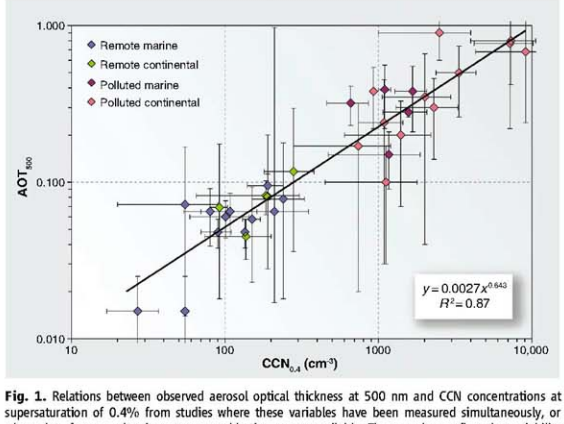



**Fig. 1.** Relations between observed aerosol optical thickness at 500 nm and CCN concentrations at supersaturation of 0.4% from studies where these variables have been measured simultaneously, or where data from nearby sites at comparable times were available. The error bars reflect the variability of measurements within each study (standard deviations or quartiles). The equation of the regression line between aerosol optical thickness ($y$) and $CCN_{0.4}$ ($x$) is given by the inset expression; $R$ is the correlation coefficient. The aerosols exclude desert dust. [Adapted from (*1*)]

[1]Institute of Earth Sciences, Hebrew University of Jerusalem, Jerusalem 91904, Israel. [2]Institute for Atmospheric and Climate Science, ETH Zürich, 8092 Zürich, Switzerland. [3]Universidad Nacional Autónoma de México, Mexico City 04510, Mexico. [4]School of Physics and Centre for Climate and Air Pollution Studies, Environmental Change Institute, National University of Ireland, Galway, Ireland. [5]Department of Physics, University of Helsinki, Post Office Box 64, Helsinki 00014, Finland. [6]Istituto di Scienze dell'Atmosfera e del Clima–CNR, Bologna 40129, Italy. [7]Biogeochemistry Department, Max Planck Institute for Chemistry, Post Office Box 3060, D-55020 Mainz, Germany.

*To whom correspondence should be addressed. E-mail: daniel.rosenfeld@huji.ac.il

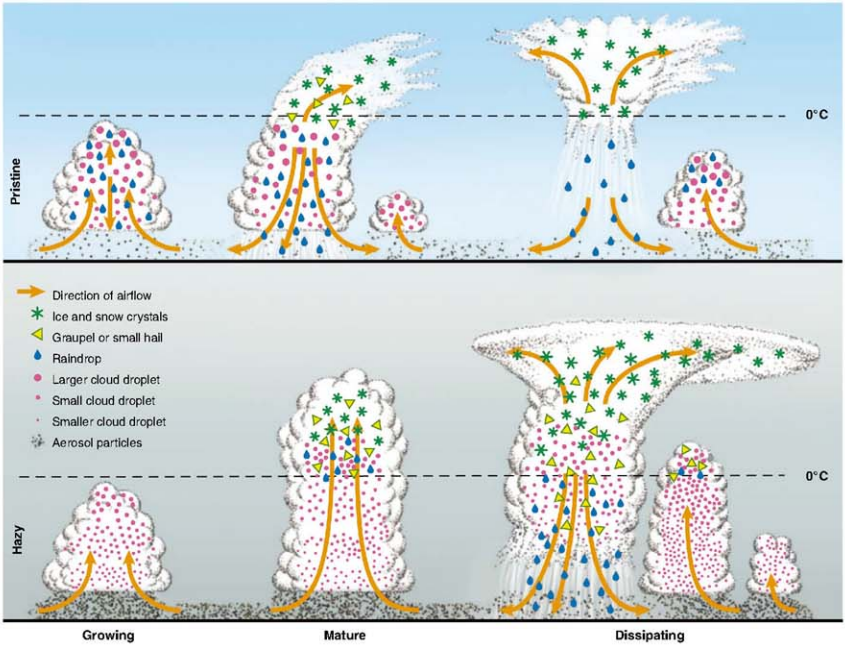



**Fig. 2.** Evolution of deep convective clouds developing in the pristine (top) and polluted (bottom) atmosphere. Cloud droplets coalesce into raindrops that rain out from the pristine clouds. The smaller drops in the polluted air do not precipitate before reaching the supercooled levels, where they freeze onto ice precipitation that falls and melts at lower levels. The additional release of latent heat of freezing aloft and reabsorbed heat at lower levels by the melting ice implies greater upward heat transport for the same amount of surface precipitation in the more polluted atmosphere. This means consumption of more instability for the same amount of rainfall. The inevitable result is invigoration of the convective clouds and additional rainfall, despite the slower conversion of cloud droplets to raindrops (*43*).

smoke from ship smokestacks in otherwise pristine clouds over the ocean (*17*). This created the expectation that polluted areas would suffer from reduced rainfall. On the other hand, it was expected that accelerating the conversion of cloud water to precipitation (i.e., increasing the autoconversion rate) by cloud seeding would enhance rainfall amounts. It turns out, however, that polluted areas are not generally drier, and rain enhancement by cloud seeding remains inconclusive (*18*, *19*).

With the advent of satellite measurements, it became possible to observe the larger picture of aerosol effects on clouds and precipitation. (We exclude the impacts of ice nuclei aerosols, which are much less understood than the effects of CCN aerosols.) Urban and industrial air pollution plumes were observed to completely suppress precipitation from 2.5-km-deep clouds over Australia (*20*). Heavy smoke from forest fires was observed to suppress rainfall from 5-km-deep tropical clouds (*21*, *22*). The clouds appeared to regain their precipitation capability when ingesting giant (>1 μm diameter) CCN salt particles from sea spray (*23*) and salt playas (*24*). These observations were the impetus for the World Meteorological Organization and the International Union of Geodesy and Geophysics to mandate an assessment of aerosol impact on precipitation (*19*). This report concluded that "it is difficult to establish clear causal relationships between aerosols and precipitation and to determine the sign of the precipitation change in a climatological sense. Based on many observations and model simulations the effects of aerosols on clouds are more clearly understood (particularly in ice-free clouds); the effects on precipitation are less clear."

A recent National Research Council report that reviewed "radiative forcing of climate change" (*25*) concluded that the concept of radiative forcing "needs to be extended to account for (1) the vertical structure of radiative forcing, (2) regional variability in radiative forcing, and (3) nonradiative forcing." It recommended "to move beyond simple climate models based entirely on global mean top of the atmosphere radiative forcing and incorporate new global and regional radiative and nonradiative forcing metrics as they become available." We propose such a new metric below.

### How Can Slowing the Conversion of Cloud Droplets to Raindrops Enhance Rainfall?

A growing body of observations shows that submicrometer CCN aerosols decrease precipitation

from shallow clouds (*17*, *20*, *21*, *26*–*28*) and invigorate deep convective rain clouds with warm (> ~15°C) cloud base (*29*–*33*), although the impact on the overall rainfall amount is not easily detectable (*34*, *35*). These observations are supported by a large number of cloud-resolving model studies (*36*–*43*). The simulations also show that adding giant CCN to polluted clouds accelerates the autoconversion, mainly through nucleating large drops that rapidly grow into precipitation particles by collecting the other smaller cloud droplets (*44*). However, the autoconversion rate is not restored to that of pristine clouds (*42*).

Fundamentally, the amount of precipitation must balance the amount of evaporation at a global scale. Therefore, the consequence of aerosols suppressing precipitation from shallow clouds must be an increase in precipitation from deeper clouds. Such compensation can be accomplished not only at the global scale (*45*) but also at the cloud scale; that is, the clouds can grow to heights where aerosols no longer impede precipitation (*46*). All of this is consistent with the conceptual model shown in Fig. 2. This model suggests that slowing the rate of cloud droplet coalescence into raindrops (i.e., autoconversion) delays the precipitation of the cloud water, so that more water can ascend to altitudes where the temperature is colder than 0°C. Even if the total rainfall amount is not decreased by the increase in aerosols, delaying the formation of rain is sufficient to cause invigoration of cloud dynamics. By not raining early, the condensed water can form ice precipitation particles that release the latent heat of freezing aloft (*6*, *29*, *30*) and reabsorb heat at lower levels where they melt after falling.

The role of ice melting below the 0°C isotherm level in invigoration has been successfully modeled (*47*), although models also predict invigoration through increased aerosol loads even without ice processes (*43*). These model simulations suggest that the delay of early rain causes greater amounts of cloud water and rain intensities later in the life cycle of the cloud. The enhanced evaporative cooling of the added cloud water, mainly in the downdrafts, provides part of the invigoration by the mechanism of enhanced cold pools near the surface that push upward the ambient air. The greater cooling below and heating above lead to enhanced upward heat transport, both in absolute terms and normalized for the same amount of surface precipitation. The consumption of more convective available potential energy (CAPE) for the same rainfall amount would then be converted to an equally greater amount of released kinetic energy that could invigorate convection and lead to a greater convective overturning, more precipitation, and deeper depletion of the static instability (*6*). Simulations have shown that greater heating higher in the troposphere enhances the atmospheric circulation systems (*48*).

In clouds with bases near or above the 0°C isotherm, almost all the condensate freezes, even if it forms initially as supercooled raindrops in a low-CCN environment. Moreover, the slowing of the autoconversion rate by large concentrations of CCN can leave much of the cloud droplets airborne when strong updrafts thrust them above the homogeneous ice nucleation level of ~ –38°C, where they freeze into small ice particles that have no effective mechanism to coagulate and fall as precipitation. This phenomenon was observed by aircraft (*49*) and simulated for convective storms in west Texas (*50*) and the U.S. high plains (*51*). When the same simulation (*50*) was repeated with reduced CCN concentrations, the calculated rainfall amount increased substantially. The same model showed that adding small CCN aerosols in warm-base clouds has the opposite effect to that of cold-base clouds: increasing the precipitation amount by invigorating the convective overturning, while keeping the precipitation efficiency (i.e., surface precipitation divided by total cloud condensates) lower (*52*).

The invigoration due to aerosols slowing the autoconversion can be explained according to fundamental theoretical considerations of the pseudo-adiabatic parcel theory (Fig. 3). The CAPE measures the amount of moist static energy that is available to drive the convection. Its value is normally calculated with reference to a pseudo-adiabatic cloud parcel that rises while precipitating all its condensate in the form of rain, even at subfreezing temperatures.

Consider the case of a tropical air parcel that ascends from sea level with initial conditions of cloud base pressure of 960 hPa and temperature of 22°C. When not allowing precipitation, all the condensed water remains in the parcel and requires 415 J $kg^{-1}$ to rise to the height of the –4°C isotherm (point $d_1$ in Fig. 3), which is the highest temperature at which freezing can practically occur in the atmosphere. Freezing all the cloud water would warm the air and add thermal buoyancy by an amount that would almost exactly balance the condensate load ($d_2$). When the ice hydrometeors precipitate from a parcel, it becomes more positively buoyant because of its reduced weight ($d_3$), so that the released convective energy at the top of the cloud ($d_4$) is the largest. Specifically, it is greater by ~1000 J $kg^{-1}$ relative to the case where cloud water is precipitated as rain below the –4°C isotherm and as ice above that level ($c_1$). However, further delaying the conversion of cloud water into precipitation to greater heights above the 0°C level weakens the convection. In the extreme case of extending the suppression from the –4°C to the –36°C isotherm level ($a_1$), additional energy of 727 J $kg^{-1}$ is invested in lifting the condensates. There is no effective mechanism for precipitating cloud water that glaciated homogeneously into small ice particles. This would prevent the unloading of the parcel, taking up even more convective energy and further suppressing the convection and the precipitation. In reality, cloud parcels always mix with the environment, but this applies equally to all the scenarios in Fig. 3, so that qualitatively the contrasting aerosol effects remain the same. Although the idealized calculations here are useful to establish the concepts, the exact calculations require running three-dimensional models on the full life cycle of convective cloud systems, followed by validation with detailed observations.

The importance of the aerosol control of the released convective energy by adding as much as 1000 J $kg^{-1}$ can be appreciated by considering that CAPE averages ~1000 to 1500 J $kg^{-1}$ in the Amazon (*30*). Simulations of aerosols invigorating peak updrafts by 20% (*37*, *52*) are

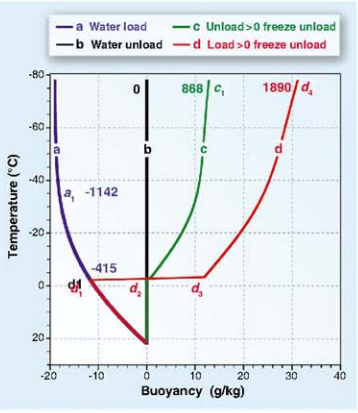



**Fig. 3.** The buoyancy of an unmixed adiabatically raising air parcel. The zero-buoyancy reference is the standard parcel: liquid water saturation, immediately precipitating all condensates without freezing (vertical line *b*). Cloud base is at 22°C and 960 hPa. The buoyancy of the following scenarios is shown: (*a*) suppressing rainfall and keeping all condensed water load, without freezing; (*b*) precipitating all condensed water, without freezing; (*c*) precipitating all condensates, with freezing at $T < -4$°C; (*d*) Suppressing precipitation until $T = -4$°C, and then freezing and precipitating all condensed water above that temperature. The released static energy (J $kg^{-1}$) with respect to reference line *b* is denoted by the numbers.

consistent with an increase of released convective energy by nearly 50%.

### Role of Radiative Versus Microphysical Aerosol Effects

Until now, the radiative and microphysical impacts of aerosols on the climate system have been considered separately and independently; their various, often conflicting, influences have not been amenable to quantitative weighting on the same scale. Given the opposing microphysical and radiative effects on the vigor and rainfall amounts of deep warm-base convective clouds, there is a need to assess the combined effects of these two factors (*25*).

A quantitative comparison between the strengths of the radiative and microphysical effects of the aerosols is presented in Fig. 4. Because optically active aerosols are larger than 0.05 μm in radius, and because mature pollution aerosols of this or larger size can act as CCN (*53*), CCN concentrations generally increase with aerosol optical thickness (AOT) (Fig. 1). The empirical relationship between AOT and CCN is shown in Fig. 4 by $AOT = 0.0027 \times (CCN_{0.4})^{0.64}$ (*1*), where $CCN_{0.4}$ is the concentration of CCN active at a supersaturation of 0.4%. The cloud droplet concentration $N_c$ is proportional to $(CCN_{0.4})^k$, where $k$ is typically smaller than 1. Using $k = 0.825$ relates 2000 cloud drops $cm^{-3}$ to $10^4$ $CCN_{0.4}$ $cm^{-3}$, which corresponds to AOT = 1. The value of $k$ was inferred from Ramanathan *et al.* (*7*), although Freud *et al.* (*54*) imply that $k$ is closer to 1. In turn, $N_c$ was shown to be related to the depth above cloud base ($D$) required for onset of rain (*54*). This depth determines the thermodynamic track of the rising parcel (Fig. 3) and hence the vigor of the convection and the extent of convective overturning, which determines the rainfall amount produced by the cloud system throughout its life cycle. The cloudy parcel ascends along curve *a* in Fig. 3 as long as the cloud top has not reached $D$, and shifts to a track between curves *c* and *d* according to the amount of condensed water at that height.

The dependence of $D$ on CCN is obtained by a compilation of aircraft measurements (*27*, *54*, *55*) that provides an approximate relation of $D = 80 + (4 \times CCN_{0.4})$. According to this relation, $CCN_{0.4}$ should reach ~1200 $cm^{-3}$ for preventing rainout from typical tropical clouds before reaching the practical freezing temperature of –4°C, which is at $D \approx 5$ km. At this point the invigoration effect is at its maximum, where the cloud parcel follows curve *d* in Fig. 3. Adding CCN beyond this point suppresses the vigor of the convection by shifting the cloud parcel gradually from curve *d* to curve *a* in Fig. 3. This means that the microphysical effect on invigorating the convection has a maximum at moderate CCN concentrations. This maximum becomes smaller for cooler-base clouds, where the distance to the freezing level is shorter, so that fewer CCN are sufficient to suppress the onset of rain up to that level.

At the point of strongest microphysical invigoration, AOT is still at the modest value of ~0.25. Added aerosols increase the AOT and reduce the flux of solar energy to the surface, which energizes convection. As a result, with increasing aerosol loads beyond the optimum, the weakening of the microphysical invigoration is reinforced by the suppressive effect of reduced surface heating.

The interplay between the microphysical and radiative effects of the aerosols may explain the observations of Bell *et al.* (*33*), who showed that the weekly cycle of air pollution aerosols in the southeastern United States is associated with a weekday maximum and weekend minimum in the intensity of afternoon convective rainfall during summer. This was mirrored by a minimum in the midweek rainfall over the adjacent sea areas, reflecting an aerosol-induced modulation of the monsoonal convergence of air and its rising over land with return flow aloft to the ocean. This is a remarkable finding, as it suggests that the microphysical impacts of aerosols on invigorating warm-base deep clouds are not necessarily at the expense of other clouds in the same region, but can lead to changes in regional circulation that lead to greater moisture convergence and regional precipitation.

This weekly cycle emerged in the late 1980s and strengthened through the 1990s, along with the contemporary reversal of the dimming trend of solar radiation reaching the surface, which took place until the 1980s (*56*). This was likely caused by the reversal in the emissions trends of sulfates and black carbon (*57*). It is possible that the weekly cycle emerged when the overall aerosol levels decreased to the range where the microphysical impacts are dominant, as shown in Fig. 4.

### Measuring Radiative and Microphysical Aerosol Effects with the Same Metric

The precipitation and the radiative effects of the aerosols (both direct and cloud-mediated) can be integrally measured when considering the combined changes in the energy of the atmosphere and the surface. The commonly used metrics are the radiative forcing at the top of the atmosphere (TOA) and at the BOA (bottom of the atmosphere, i.e., Earth's surface), measured in W $m^{-2}$. The atmospheric radiative forcing is the difference between TOA and BOA forcing (*7*). Here we propose a new metric, the aerosol thermodynamic forcing (TF) (*58*), representing the aerosol-induced change in the atmospheric energy budget that is not radiative in nature. In contrast to TOA radiative forcing, TF does not change the net Earth energy budget, but rather redistributes it internally; hence, TF can affect temperature gradients and atmospheric circulation. The main source of TF is the change in the amount of latent heat released by aerosol-induced changes in clouds and precipitation. It can be expressed as a change in latent heat flux (in units of W $m^{-2}$) in the atmospheric column.

The vertical distribution of the atmospheric heating is critically important because it determines the vertical lapse rate and hence the CAPE, which quantifies the ability to produce convective clouds and precipitation. Atmospheric radiative heating due to absorbing aerosols tends to reduce CAPE and thereby suppress the development of convective clouds, whereas the microphysical effects of aerosols allow a deeper exploitation of CAPE and hence invigoration of convection and associated precipitation.

All the components of the aerosol radiative (direct and cloud-

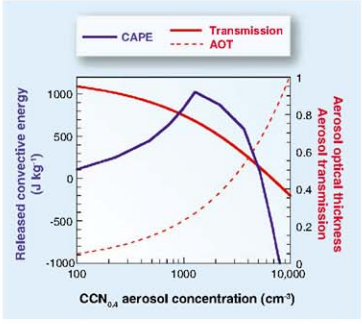



**Fig. 4.** Illustration of the relations between the aerosol microphysical and radiative effects. The aerosol optical thickness (AOT) is assumed to reach 1 at $CCN_{0.4} = 10^4$ $cm^{-3}$ (dashed red line), which corresponds to nucleation of 2000 cloud drops $cm^{-3}$. The related transmission of radiation reaching the surface is shown by the solid red line. The vigor of the convection is shown by the blue line, which provides the released convective available potential energy (CAPE) of a cloud parcel that ascends to the cloud top near the tropopause. The calculation is based on the scheme in Fig. 3, with respect to curve *c* as the zero reference. Note that a maximum in CAPE occurs at $CCN_{0.4} \approx 1200$ $cm^{-3}$, which corresponds to the maximum cloud invigoration according to curve *d* of the scheme in Fig. 3. The AOT corresponding to the $CCN_{0.4}$ at the microphysical optimum is only 0.25. Adding aerosols beyond this point substantially decreases the vigor of the cloud because both microphysical and radiative effects work in the same direction: smaller release of convective energy aloft and less radiative heating at the surface.

mediated) and thermodynamic forcing and the resulting changes in CAPE can now be quantified as energy flux perturbations in units of W $m^{-2}$. Consider the example of smoke changing tropical convection from thermodynamic path *c* to path *d* in Fig. 3. At the end of the convective cycle, an additional 1000 J $kg^{-1}$ are depleted from CAPE relative to convection under pristine conditions. The resultant increased convective overturning is likely to produce more rainfall and increase the temperature by converting more latent heat into sensible heat, at a rate of 29 W $m^{-2}$ for each added millimeter of rainfall during 24 hours. This can be considered as a cloud-mediated TF of aerosols, which works to enhance rainfall and accelerate the hydrological cycle, resulting in a positive sign for TF. On the other hand, if the smoke becomes very thick, its radiative impact would be to reduce surface latent and sensible heating and warm the mid-troposphere. For example, an AOT of 1 induces a BOA forcing of –45 W $m^{-2}$ in the Amazon (*5*). This stabilization of the atmosphere would cause less convection and depletion of CAPE, less rainfall, and a resulting deceleration of the hydrological cycle (*7*). Furthermore, too much aerosol can suppress the precipitation-forming processes to the extent of changing from thermodynamic path *d* to path *a* in Fig. 3 (see also Fig. 4), hence reversing the cloud-mediated TF of aerosols from positive to negative, adding to the negative radiative forcing.

Thermodynamic forcing can occur even without changing the surface rainfall: The energy change when polluted clouds develop along track *d* in Fig. 3, with respect to the pristine reference state shown in track *c*, would be defined as TF. In this case, the TF solely due to added release of latent heat of freezing is 2.44 W $m^{-2}$ $mm^{-1}$ $day^{-1}$ of heating above the freezing level and the same amount of cooling due to melting below the melting level. This is a net vertical redistribution of latent heat. For an area-average rainfall of 20 mm $day^{-1}$, the TF scales to 48.8 W $m^{-1}$. In addition, we should consider the thermodynamic consequences of the aerosol-induced added rainfall due to increased convective overturning. This would convert latent heat to sensible heat at a rate of 29 W $m^{-2}$ $mm^{-1}$ $day^{-1}$. Such deeper consumption of CAPE would require a longer time for the atmosphere to recover for the next convective cycle, representing a temporal redistribution of heating and precipitation.

## Concluding Thoughts

The next challenge will be to map the radiative and cloud-mediated thermodynamic forcing of the aerosols in the parameter space of AOT versus CCN. The good correlation between AOT and CCN means that, at least at large scales, the radiative and microphysical effects of aerosols on cloud physics are not free to vary independently (*1*), and hence mainly the diagonal of the parameter space is populated.

According to Fig. 4, there should be an optimum aerosol load in the tropical atmosphere that should lead to the most positive aerosol thermodynamic forcing, manifested as the most vigorous convection. This optimum probably occurs at AOT $\approx$ 0.25 and $CCN_{0.4} \approx 1200\ cm^{-3}$. Remarkably, these fundamental considerations for AOT $\approx$ 0.25 for optimal cloud development were matched recently by observations in the Amazon (*59*).

This hypothesis reconciles the apparent contradictory reports that were reviewed in two major assessments (*18*, *19*) as impeding our overall understanding of cloud-aerosol impacts on precipitation and the climate system. The main cause for the previous uncertainties was the nonmonotonic character of competing effects, which is inevitable in a system that has an optimum. The new conceptual model outlined here improves our understanding and ability to simulate present and future climates. It also has implications for intentional weather and climate modification, which are being considered in the context of cloud seeding for precipitation enhancement and geoengineering. Testing this hypothesis is planned within the Aerosol Cloud Precipitation Climate (ACPC) initiative (*60*, *61*).

### References and Notes


1. M. O. Andreae, *Atmos. Chem. Phys. Discuss.* **8**, 11293 (2008).
2. M. O. Andreae, *Science* **315**, 50 (2007).
3. U. Lohmann, J. Feichter, *Atmos. Chem. Phys.* **5**, 715 (2005).
4. V. Ramanathan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 5326 (2005).
5. I. Koren, Y. J. Kaufman, L. A. Remer, J. V. Martins, *Science* **303**, 1342 (2004).
6. D. Rosenfeld, *Space Sci. Rev.* **125**, 149 (2006).
7. V. Ramanathan, P. J. Crutzen, J. T. Kiehl, D. Rosenfeld, *Science* **294**, 2119 (2001).
8. S. Menon, J. Hansen, L. Nazarenko, Y. F. Luo, *Science* **297**, 2250 (2002).
9. C. Wang, *J. Geophys. Res.* **109**, D03106 (2004).
10. Y. Rudich, A. Sagi, D. Rosenfeld, *J. Geophys. Res.* **108**, 10.1029/2003JD003472 (2003).
11. R. L. Miller, I. Tegen, J. Perlwitz, *J. Geophys. Res.* **109**, D04203 (2004).
12. L. D. Rotstayn, U. Lohmann, *J. Geophys. Res.* **107**, 10.1029/2002JD002128 (2002).
13. I. M. Held, T. L. Delworth, J. Lu, K. L. Findell, T. R. Knutson, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 17891 (2005).
14. W. Cotton, R. Pielke, *Human Impacts on Weather and Climate* (Cambridge Univ. Press, Cambridge, 2007).
15. R. Gunn, B. B. Phillips, *J. Meteorol.* **14**, 272 (1957).
16. P. Squires, *Tellus* **10**, 256 (1958).
17. L. F. Radke, J. A. Coakley Jr., M. D. King, *Science* **246**, 1146 (1989).
18. National Research Council, *Critical Issues in Weather Modification Research* (National Academies Press, Washington, DC, 2003).
19. Z. Levin, W. Cotton, *Aerosol Pollution Impact on Precipitation: A Scientific Review. Report from the WMO/IUGG International Aerosol Precipitation Science Assessment Group (IAPSAG)* (World Meteorological Organization, Geneva, Switzerland, 2007).
20. D. Rosenfeld, *Science* **287**, 1793 (2000).
21. D. Rosenfeld, *Geophys. Res. Lett.* **26**, 3105 (1999).
22. D. Rosenfeld, W. L. Woodley, in *Cloud Systems, Hurricanes, and the Tropical Rainfall Measuring Mission (TRMM)*, W.-K. Tao, R. Adler, Eds. (American Meteorological Society, Boston, 2003), pp. 59–80.
23. D. Rosenfeld, R. Lahav, A. Khain, M. Pinsky, *Science* **297**, 1667 (2002); published online 15 August 2002 (10.1126/science.1073869).
24. Y. Rudich, O. Khersonsky, D. Rosenfeld, *Geophys. Res. Lett.* **29**, 10.1029/2002GL016055 (2002).
25. National Research Council, *Radiative Forcing of Climate Change: Expanding the Concept and Addressing Uncertainties* (National Academies Press, Washington, DC, 2005).
26. D. Rosenfeld, Y. J. Kaufman, I. Koren, *Atmos. Chem. Phys.* **6**, 2503 (2006).
27. D. Rosenfeld *et al.*, *J. Geophys. Res.* **113**, D15203 (2008).
28. M. O. Andreae *et al.*, *Science* **303**, 1337 (2004).
29. J. Molinié, C. A. Pontikis, *Geophys. Res. Lett.* **22**, 1085 (1995).
30. E. Williams *et al.*, *J. Geophys. Res.* **107**, 10.1029/2001JD000380 (2002).
31. I. Koren, Y. J. Kaufman, D. Rosenfeld, L. A. Remer, Y. Rudich, *Geophys. Res. Lett.* **32**, L14828 (2005).
32. J. C. Lin, T. Matsui, R. A. Pielke Sr., C. Kummerow, *J. Geophys. Res.* **111**, D19204 (2006).
33. T. L. Bell *et al.*, *J. Geophys. Res.* **113**, D00Z09 (2008).
34. D. M. Schultz, S. Mikkonen, A. Laaksonen, M. B. Richman, *Geophys. Res. Lett.* **34**, L22815 (2007).
35. T. Bell, D. Rosenfeld, *Geophys. Res. Lett.* **35**, L09803 (2008).
36. A. Khain, A. Pokrovsky, M. Pinsky, A. Seifert, V. Phillips, *J. Atmos. Sci.* **61**, 2963 (2004).
37. A. Khain, D. Rosenfeld, A. Pokrovsky, *Q. J. R. Meteorol. Soc.* **131**, 2639 (2005).
38. B. H. Lynn *et al.*, *Mon. Weather Rev.* **133**, 59 (2005).
39. C. Wang, *J. Geophys. Res.* **110**, D21211 (2005).
40. S. C. van den Heever, G. G. Carrió, W. R. Cotton, P. J. DeMott, A. J. Prenni, *J. Atmos. Sci.* **63**, 1752 (2006).
41. A. Seifert, K. D. Beheng, *Meteorol. Atmos. Phys.* **92**, 67 (2006).
42. A. Teller, Z. Levin, *Atmos. Chem. Phys.* **6**, 67 (2006).
43. W. K. Tao *et al.*, *J. Geophys. Res.* **112**, D24S18 (2007).
44. D. B. Johnson, *J. Atmos. Sci.* **39**, 448 (1982).
45. U. Lohmann, *Atmos. Chem. Phys.* **8**, 2115 (2008).
46. E. R. Graber, Y. Rudich, *Atmos. Chem. Phys.* **6**, 729 (2006).
47. V. T. J. Phillips, A. Pokrovsky, A. Khain, *J. Atmos. Sci.* **64**, 338 (2007).
48. M. W. DeMaria, *J. Atmos. Sci.* **42**, 1944 (1985).
49. D. Rosenfeld, W. L. Woodley, *Nature* **405**, 440 (2000).
50. A. P. Khain, D. Rosenfeld, A. Pokrovsky, *Geophys. Res. Lett.* **28**, 3887 (2001).
51. Z. Cui, K. S. Carslaw, Y. Yin, S. Davies, *J. Geophys. Res.* **111**, D05201 (2006).
52. A. P. Khain, N. BenMoshe, A. Pokrovsky, *J. Atmos. Sci.* **65**, 1721 (2008).
53. U. Dusek *et al.*, *Science* **312**, 1375 (2006).
54. E. Freud, D. Rosenfeld, M. O. Andreae, A. A. Costa, P. Artaxo, *Atmos. Chem. Phys.* **8**, 1661 (2008).
55. M. C. VanZanten, B. Stevens, G. Vali, D. H. Lenschow, *J. Atmos. Sci.* **62**, 88 (2005).
56. M. Wild *et al.*, *Science* **308**, 847 (2005).
57. D. G. Streets, Y. Wu, M. Chin, *Geophys. Res. Lett.* **33**, L15806 (2006).
58. The term "thermodynamic aerosol effect" was first mentioned in (*25*), but in a more restrictive context.
59. I. Koren, J. V. Martins, L. A. Remer, H. Afargan, *Science* **321**, 946 (2008).
60. B. Stevens, *iLEAPS Newsletter* **5**, 10 (2008).
61. The Aerosol Cloud Precipitation Climate (ACPC) initiative is a joint initiative by the International Geosphere/Biosphere Programme (IGBP) core projects Integrated Land Ecosystem/Atmosphere Process Study (iLEAPS) and International Global Atmospheric Chemistry (IGAC) and the World Climate Research Programme (WCRP) project Global Energy and Water Cycle Experiment (GEWEX).


62. This paper resulted from discussions held during an ACPC workshop hosted and supported by the International Space Science Institute, Bern, Switzerland, through its International Teams Program.

10.1126/science.1160606

# Shadow Enhancers as a Source of Evolutionary Novelty

Joung-Woo Hong, David A. Hendrix, Michael S. Levine*

The dorsal-ventral patterning of the early *Drosophila* embryo is controlled by a sequence-specific transcription factor, Dorsal, which is related to mammalian NF-κB (*1*). Dorsal works in concert with two additional transcription factors, Twist and Snail, to regulate gene expression in the early embryo. Chromatin immunoprecipitation (ChIP)–chip assays identified a few hundred binding clusters for Dorsal, Twist, and Snail scattered throughout the *Drosophila* genome (*2*). Over 35 of these clusters function as authentic enhancers when tested in transgenic embryos.

ChIP-chip assays predicted that many of the Dorsal target genes contain two separate enhancers for the same or similar expression pattern. This prediction was experimentally confirmed for *vnd* and *miR-1* (*2*). *vnd* contains two enhancers that mediate expression in the presumptive neurogenic ectoderm, whereas *miR-1* contains at least two enhancers for expression in the ventral mesoderm. In both cases, the secondary enhancers map within 5 kb of the transcription start site.

However, some of the potential secondary enhancers identified by the ChIP-chip assays are predicted to map quite far from Dorsal target genes. For example, *brinker* (*brk*) is regulated by a known enhancer located in the 5′ flanking region (*3*). A potential secondary enhancer maps within the intron of a neighboring gene, *Atg5*, located ~13 kb downstream of the *brk* transcription start site. A ~1-kb genomic DNA fragment encompassing the *Atg5* intron was tested for enhancer activity in transgenic embryos (Fig. 1A). It directs broad lateral stripes of lacZ reporter gene expression, similar to the endogenous *brk* expression pattern that is recapitulated by the previously identified 5′ enhancer.

A similar situation is seen for the Dorsal target gene, *sog*. Bioinformatics methods identified an intronic enhancer that recapitulates the normal *sog* expression pattern in the presumptive neurogenic ectoderm (*3*). ChIP-chip assays identified this enhancer, as well as a second cluster of Dorsal, Twist, and Snail binding sites located 20 kb 5′ of the *sog* transcription start site, downstream of a neighboring gene (Fig. 1B). The newly identified binding cluster generates broad lateral stripes of gene expression in transgenic embryos, similar to those produced by the intronic enhancer. The secondary enhancers identified in this study are almost certainly dedicated to the regulation of *brk* and *sog* transcription units because the associated genes, *Atg5* and *CG8117*, respectively, are not significantly expressed in the early embryo (Fig. 1 and fig. S1).

ChIP-chip assays suggest that as many as one-third or even one-half of all Dorsal target genes might be regulated by secondary enhancers (*2*). We propose the term "shadow enhancer" for remote secondary enhancers mapping far from the target gene and mediating activities overlapping the primary enhancer. Phylogenetic comparisons suggest that the *brk* and *sog* shadow enhancers are evolving more rapidly than the primary enhancers mapping within or near the two genes (figs. S2 and S3). Despite these different rates of divergence, the overall structures of the shadow enhancers are clearly related to their respective primary enhancers (fig. S4). Given the conservation of the shadow enhancers in all 12 sequenced drosophilids, it is likely that they are essential for fitness.

Why are Dorsal target genes regulated by shadow enhancers? They might help ensure precise and reproducible patterns of gene expression during embryogenesis. It is possible that shadow enhancers are pervasively used in animal development. For example, the mouse *sonic hedgehog* gene is regulated in the floorplate of the embryonic neural tube by two separate enhancers with slightly distinct activities (*4*). Shadow enhancers can explain why deletions of well-defined enhancers sometimes produce no apparent mutant phenotypes [e.g., (*5*)]. We suggest that shadow enhancers might arise from duplication, comparable to the duplication and divergence of protein-coding sequences.

The evolution of cis-regulatory DNAs is a major mechanism of animal diversity [e.g., (*6*)]. However, there is the potential problem that such change could compromise essential genetic activities. Shadow enhancers have the potential to evolve novel binding sites and achieve new regulatory activities without disrupting the core patterning functions of critical developmental control genes.

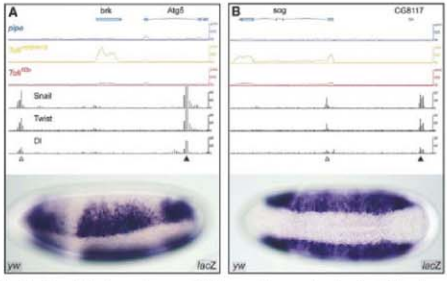



**Fig. 1.** Identification of shadow enhancers. **(A)** Genome browser showing the *Brinker* locus (*brk*) and neighboring gene *Atg5* (http://flybuzz.berkeley.edu/cgi-bin/gbrowse/fly4_3/). The first three lines from the top—blue, yellow, and red—show the levels of steady-state RNAs in the *pipe*, *Toll*$^{rm9/rm10}$, and *Toll*$^{10b}$ mutants, respectively (*7*). *brk* transcripts are absent in *pipe* and *Toll*$^{10b}$ mutants but present in *Toll*$^{rm9/rm10}$ mutants. *Atg5* is inactive in all mutants, suggesting that the intronic enhancer is dedicated to *brk* regulation. The last three lines show the distributions of Snail, Twist, and Dorsal (Dl) based on whole-genome ChIP-chip assays (*2*). The leftmost cluster (open arrowhead) coincides with the known, primary enhancer in the 5′ flanking region. A second cluster (solid arrowhead) is detected 13 kb downstream of the *brk* transcription start site within *Atg5*. An ~1-kb genomic DNA fragment encompassing the 3′ binding cluster (shadow enhancer) was tested in transgenic embryos (solid arrowhead). The 5′ enhancer was tested previously (*3*). Both fragments function as authentic enhancers to generate lateral stripes of gene expression in the neurogenic ectoderm. **(B)** Same as (A) except that the *sog* locus is shown. *sog* transcripts are predominantly detected in *Toll*$^{rm9/rm10}$ mutants where the gene is fully active (see the blue, yellow, and red lines, which show the results of the whole-genome tiling arrays). ChIP-chip assays identify two clusters of Dorsal, Twist, and Snail binding sites within intron 1 and 3′ of the neighboring gene, *CG8117*. Both genomic DNA fragments function as authentic enhancers to direct lateral stripes of gene expression. Gene prediction models are displayed above each graphical presentation. Each 3′ end is indicated by a triangle.

### References and Notes

1. A. Stathopoulos *et al.*, *Cell* **111**, 687 (2002).
2. J. Zeitlinger *et al.*, *Genes Dev.* **21**, 385 (2007).
3. M. Markstein *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 763 (2002).
4. Y. Jeong *et al.*, *Development* **133**, 761 (2006).
5. N. Xiong, C. Kang, D. H. Raulet, *Immunity* **16**, 453 (2002).
6. S. Jeong *et al.*, *Cell* **132**, 783 (2008).
7. F. Biemar *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 12763 (2006).
8. This study was funded by the NIH (GM46638) and the Moore Foundation.

### Supporting Online Material

www.sciencemag.org/cgi/content/full/321/5894/1314/DC1
Materials and Methods
Figs. S1 to S4
Tables S1 to S3
References and Notes

19 May 2008; accepted 2 July 2008
10.1126/science.1160631

Department of Molecular and Cell Biology, Division of Genetics, Genomics, and Development, Center for Integrative Genomics, University of California, Berkeley, CA 94720, USA.

*To whom correspondence should be addressed. E-mail: mlevine@berkeley.edu

# The Crystal Structure of a Mammalian Fatty Acid Synthase

Timm Maier, Marc Leibundgut, Nenad Ban*

Mammalian fatty acid synthase is a large multienzyme that catalyzes all steps of fatty acid synthesis. We have determined its crystal structure at 3.2 angstrom resolution covering five catalytic domains, whereas the flexibly tethered terminal acyl carrier protein and thioesterase domains remain unresolved. The structure reveals a complex architecture of alternating linkers and enzymatic domains. Substrate shuttling is facilitated by flexible tethering of the acyl carrier protein domain and by the limited contact between the condensing and modifying portions of the multienzyme, which are mainly connected by linkers rather than direct interaction. The structure identifies two additional nonenzymatic domains: (i) a pseudo-ketoreductase and (ii) a peripheral pseudo-methyltransferase that is probably a remnant of an ancestral methyltransferase domain maintained in some related polyketide synthases. The structural comparison of mammalian fatty acid synthase with modular polyketide synthases shows how their segmental construction allows the variation of domain composition to achieve diverse product synthesis.

Fatty acids fulfill a variety of vital functions: They are central constituents of biological membranes, serve as energy storage compounds, and act as second messengers or as covalent modifiers governing the localization of proteins. In bacteria and plants, fatty acid biosynthesis is accomplished by a series of monofunctional proteins in a dissociated type II fatty acid synthase (FAS) system (*1*). In contrast, the type I FASs of fungi and animals are huge multifunctional polypeptides that integrate all steps of fatty acid synthesis into large macromolecular assemblies. Fungal FAS is a 2.6-MD $\alpha_6\beta_6$-heterododecamer with the catalytic domains distributed over two polypeptides (*2*–*4*), whereas mammalian FAS (mFAS) consists of a 270-kD polypeptide chain (comprising all seven required domains) that assembles into homodimers for enzymatic activity (*5*, *6*).

Despite this variation in structural organization, all organisms employ a conserved set of chemical reactions for fatty acid biosynthesis (*1*, *6*–*8*). Stepwise elongation of precursors is achieved by cyclic decarboxylative condensation of acyl-coenzyme A (CoA) with the elongation substrate malonyl-CoA, initiated by the starter substrate acetyl-CoA. In the priming step, the acetyl transferase loads acetyl-CoA onto the terminal thiol of the phosphopantheteine cofactor of the acyl carrier protein (ACP), which passes the acetyl moiety over to the active site cysteine of the β-ketoacyl synthase (KS). Malonyl transferase (MT) transfers the malonyl group of malonyl-CoA to ACP, and the KS catalyzes the decarboxylative condensation of the acetyl and malonylmoieties to an ACP-bound β-ketoacyl intermediate. The β-carbon position is then modified by sequential action of the NADPH (the reduced form of nicotinamide adenine dinucleotide, $NADP^+$)–dependent β-ketoreductase (KR), a dehydratase (DH), and the NADPH-dependent enoyl reductase (ER) to yield a saturated acyl product elongated by two carbon units. This acyl group functions as a starter substrate for the next round of elongation, until the growing fatty acid chain reaches a length of 16 to 18 carbon atoms and is released from ACP. In mFAS, the malonyl and acetyl transferase reactions are catalyzed by a single bifunctional protein domain, the malonyl-acetyl transferase (MAT), and the products are released from ACP as free fatty acids by a thioesterase (TE) domain (*6*).

Humans eating a typical Western diet take in a surplus of fatty acids. Consequently, de novo fatty acid biosynthesis and FAS activity are low in most body tissues. However, FAS is overexpressed in many cancer cells, and its expression level is correlated with tumor malignancy (*9*). FAS inhibitors have demonstrated anti-tumor activity in vivo and in vitro, and in recent years FAS has emerged as an important drug target for the treatment of human cancer (*10*, *11*). The medical use of FAS inhibitors has been hampered by off-target activities. Recently, more specific inhibitors of type I FAS have been described (*12*) and remain to be tested.

Currently, high-resolution structures are known for all components of bacterial (*1*) and fungal FAS (*2*, *4*, *13*), whereas the structural information for mFAS is limited to high-resolution structures for the isolated MAT [Protein Data Bank (PDB) entry 2jfd], ACP (*14*, *15*), and TE domains (*16*, *17*) and a domain architecture model based on a 4.5 Å resolution x-ray crystallographic map (*5*). Structure determination of KS–acyl transferase didomain fragments and KR domains of polyketide synthases (PKS) (*18*–*21*)—large modular megasynthases involved in the microbial synthesis of a number of bioactive compounds and drugs—has confirmed the anticipated close structural relation between mFAS and PKS modules (*6*, *22*). Here, we present the crystal structure of mFAS in its free and $NADP^+$-bound states, in which the flexibly tethered C-terminal ACP/TE domains (*23*) remain unresolved.

Institute of Molecular Biology and Biophysics, ETH Zurich, 8092 Zurich, Switzerland.

*To whom correspondence should be addressed. E-mail: ban@mol.biol.ethz.ch

## Overall Structure and Topology

The crystal structures of natively purified mFAS from pigs, free and in complex with the cofactor $NADP^+$, have been determined at 3.2 and 3.3 Å resolution and refined to $R/R_{free}$ values of 0.22/0.26 and 0.19/0.24, respectively (where $R/R_{free}$ are $\Sigma|F_{obs}(h) - F_{calc}(h)|\Sigma F_{obs}(h)$ calculated for the working/test set of reflections). Diffraction data were affected by anisotropy with one weaker direction of reciprocal space (*24*). mFAS assembles into an intertwined dimer approximating an "X" shape (Fig. 1A). This structure agrees well with our previous architectural model at intermediate resolution (*5*) and additionally provides the connectivities of domains, the detailed features of active sites, and the nature of linking sequences outside the conserved core domains. mFAS is segregated into a lower condensing portion, containing the condensing KS and the MAT domains, and an upper portion including the DH, ER, and KR domains responsible for β-carbon modification (Fig. 1, A and B). Two additional nonenzymatic domains are located at the periphery of the modifying part. The first of these domains is homologous to the methyltransferase family and is thus named "pseudo-methyltransferase" (ΨME). The second represents a truncated KR fold dimerizing with the catalytic KR domain and is referred to as "pseudo-ketoreductase" (ΨKR). The condensing and modifying parts of mFAS are loosely connected and form only tangential contacts. The structural organization of domains deviates dramatically from their linear arrangement in sequence (Fig. 1, A and C).

The two polypeptides dimerize through an extended contact area of 5400 $Å^2$, which involves more than 150 residues per chain (table S2). The main contributions to this interface arise from homophilic interactions of the KS and ER domains, with areas of ~2600 and 1600 $Å^2$, which resemble the dimer organization of monofunctional homologs (*25*, *26*). Additional dimer contacts (800 $Å^2$) are provided by the DH domain through homophilic interactions between the double "hot dog" folds via a loop around residue 941. The remaining interactions are formed by the C-terminal part of the linker region between the MAT and DH domains (residues 846 to 860) with the KS domain of the other chain (400 $Å^2$).

## Interdomain Linking and Interaction

The characteristics of multienzyme complexes are to a great degree determined by the nature of

the interactions and linking of functional subunits. Notably, animal FAS invests only ~9% of its total sequence for linkers (Fig. 2, A to E) and an additional 16% for the lateral noncatalytic ΨME and ΨKR domains (Fig. 1A). No scaffolding insertions are found in the catalytic cores. This is in contrast to the fungal FAS, the other megasynthase for which a high-resolution structure is available (*2*, *4*, *13*). In that multienzyme, almost 50% of the total sequence forms a complex structural matrix of numerous inter- and intradomain insertions, which define the spatial organization of the catalytic domains. The only structured linker domain in mFAS connects the KS and MAT domains [KS-MAT linker domain (Fig. 2A)] and is composed of amino acids 420 to 490 between KS and MAT and of residues 809 to 837 joining MAT and DH. It includes two short α-helices facing the KS and a three-stranded antiparallel β sheet on the MAT side and acts as an adapter, preventing any direct interaction between the KS and MAT domains. Similar linker domains were recently found in the KS–acyl transferase didomain structures of two PKS modules (*18*, *21*). Although an additional helix is inserted in the PKS linker domains (fig. S1), the relative positions of the transferase and KS domains remain essentially the same in mFAS and PKS (fig. S2).

The connection between the condensing and modifying part of mFAS is provided by residues 838 to 858 between the KS-MAT linker domain and the DH domain (Fig. 2, D and E). Again, the conformation and position of this linker closely resembles those observed in KS-acyl transferase didomain structures from PKS modules, even though these didomains are derived from modules with a considerably different sequence context, containing only KR domains for β-carbon modification (*18*, *21*). Besides the linker itself, only very limited contacts are formed between the condensing and modifying parts of mFAS (with an interaction area of 230 Å$^2$). It is even possible that some percentage of molecules in the crystal have an alternative connectivity between the two parts (equivalent to a rotation of the upper portion of the molecule), which would escape detection by crystallographic methods (figs. S3 and S4).

The KR domain acts as a central connector for the modifying part of mFAS and interacts with the DH, ER, and noncatalytic ΨME and ΨKR domains (Fig. 3, A and B, and table S3). In contrast, neither the DH nor the ER domain interacts with either of the noncatalytic domains, and the contact between the DH and ER domains is very weak. The KR domain interacts with the second hot dog subdomain of DH, forming an 800 Å$^2$ interface. The contact between KR and ER is less intricate and extends over an area of 400 Å$^2$. About 10% (or 1100 Å$^2$) of the KR surface is involved in a contact with the ΨKR domain, mimicking one of the two major dimerization interfaces observed in the tetrameric KR of bacteria (fig. S5). The ΨME, which has the highest mobility based on atomic displacement parameter analysis (fig. S6), protrudes from the mostly planar body of mFAS. It is docked via interactions with the KR and the ΨKR domain, the former providing 20% (200 Å$^2$) and the latter 80% (800 Å$^2$) of the docking area.

Most of the linker regions in the modifying domains are solvent exposed (Fig. 2). An important exception is β strand–forming regions at the N terminus of the ΨKR-ER linker (residues 1513 to 1518) and the C terminus of the DH2-ΨME linker (residues 1117 to 1123). These are buried between the KR, ΨKR, and ΨME domains and have an important structural role, as discussed in the next paragraph (Fig. 2C).

### The Nonenzymatic Domains

The KR character of the mFAS ΨKR domain, which has approximately half the size of the active KR domain, is maintained only in the

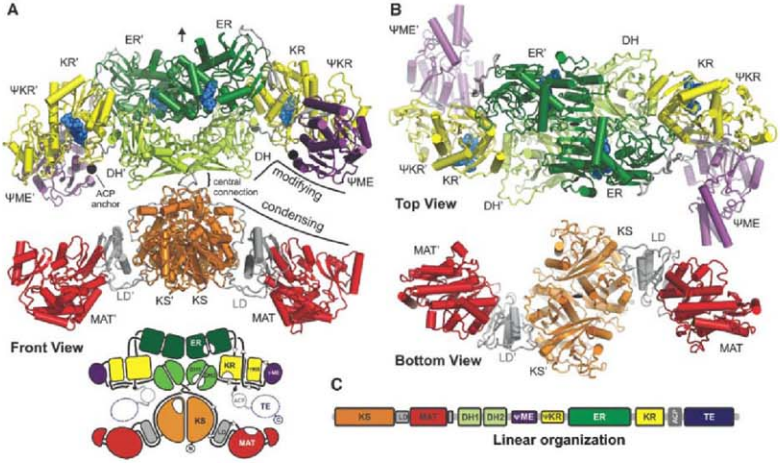



**Fig. 1.** Structural overview. **(A)** Cartoon representation of mFAS, colored by domains as indicated. Linkers and linker domains are depicted in gray. Bound NADP$^+$ cofactors and the attachment sites for the disordered C-terminal ACP/TE domains are shown as blue and black spheres, respectively. The position of the pseudo-twofold dimer axis is depicted by an arrow; domains of the second chain are indicated by an appended prime. The lower panel (front view) shows a corresponding schematic diagram. **(B)** Top (upper panel) and bottom (lower panel) views, demonstrating the "S" shape of the modifying (upper) and condensing (lower) parts of mFAS. The pseudo-twofold axis is indicated by an ellipsoid. **(C)** Linear sequence organization of mFAS, at approximate sequence scale.

conserved dimerization interface. Because of extensive truncation of its core, it has lost the ability to bind NADPH. Consequently, the ΨKR domain functions mainly to support the integrity of the active site of the catalytic KR domain. The KR/ΨKR arrangement closely resembles the structure of a KR-ΨKR domain derived from 6-deoxyerythronolide B synthase (DEBS) PKS module 1 (*19*) (Fig. 3A). The mFAS ΨKR domain lacks the two N-terminal sheet-helix windings of the DEBS1 ΨKR, which itself is already shorter than the catalytically active KR fold (Fig. 3A). Because of the insertion of the ER and ΨME domains into the KR-ΨKR fold, two β strands originally formed by the linkers flanking the KR and ΨKR domains are no longer directly adjacent to these two domains in the mFAS sequence. Rather, they are provided by amino acid stretches leading from the DH2 into the ΨME domain, 300 amino acids upstream of the ΨKR, and the linker between ΨKR and ER, separated by 360 residues from the KR domain (Fig. 3C).

The ΨME domain is structurally closely related to *S*-adenosyl-methionine (SAM)–dependent methyltransferases, in spite of low sequence homology (fig. S7A and table S4). The core of these enzymes consists of a seven-stranded β sheet with three helices on each side and the C-terminal strand in anti-parallel orientation (*27*). The methyltransferase fold of mFAS carries an additional short β strand and three helices at its N terminus (residues 1125 to 1224). At its C terminus, a short linker (residues 1407 to 1413) leads directly into the adjacent ΨKR fold. Such a topology is characteristic of small-molecule (including lipid) methyltransferases (*27*). Nevertheless, the D/ExGxGxG motif involved in SAM cofactor binding (*27*) is not conserved in any of the metazoan FAS sequenced so far (fig. S8). In FAS of mammals, this motif is changed to ExLxGxG, which probably prevents cofactor binding, in agreement with the absence of methyltransferase activity and methylated products in mFAS systems. Notably, this motif is strictly conserved in several iterative and modular PKSs found in fungi and bacteria that share a related overall domain organization with FAS but are able to methylate their polyketide substrate with an intrinsic C-methyltransferase activity (fig. S7B) (*28*–*30*). Thus, the ΨME domain of FAS most likely represents an inactive version of a previously functional enzyme in a common precursor of mFAS and PKSs.

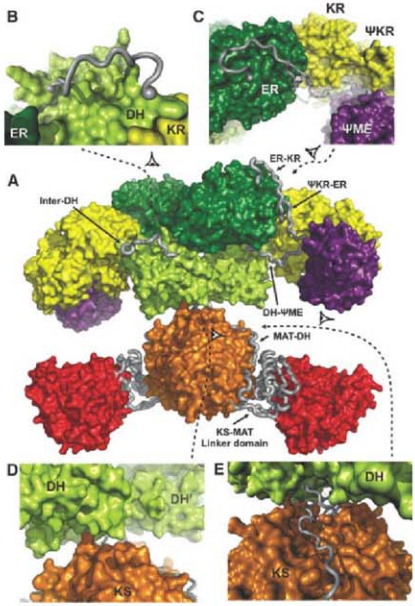



**Fig. 2.** Interdomain linkers. **(A)** Surface representation of individual mFAS domains (front view), colored as in Fig. 1. Linking regions are shown as tubes. **(B to E)** Close-up views of individual linkers. The direction of view is indicated by arrowheads in (A). (B) Linker connecting the two subdomains of the DH domain only loosely interacts with the main body of the double hot dog fold. (C) Linkers in the KR/ER region are wrapped around the domains with close interactions to the domain surfaces and pronounced linker-linker contacts; they mediate interactions between the KR, ΨKR, and ΨME domains. (D) Modifying upper and condensing lower parts of FAS are only in tangential contact in the region of the central connection. Few residues besides the connecting linkers mediate the sparse interactions via a small interface area. (E) MAT-DH linker meanders through a groove on the surface of the KS domain.

### Catalytic Domains and Cofactor Binding

*Ketoacyl synthase.* The KS enzymes of all systems for fatty acid or modular polyketide synthesis share a common fold and chemical mechanism, but their substrate specificities differ considerably (*1*, *6*). In the bacterial systems, which lack an acetyltransferase, KASIII (FabH) directly accepts acetyl-CoA as starter substrate. Further acyl chain extensions from C4 to C14 and from C14 to C16 are carried out by KASI (FabB) and KASII (FabF), respectively. Modular PKS contain specialized KS with some specificity for the β-carbon status but accept a wide range of substrate lengths (*6*). In the fungal type I FAS, only a single KS (which accepts C2 to C16 primers) is required for fatty acid synthesis. Likewise, mFAS has a single KS domain for all steps of fatty acid elongation. In contrast to PKS KS, mFAS KS is highly specific for saturated acyl chains and does not accept β-ketoacyl, β-enoyl- or β-hydroxyacyl substrates (*6*, *31*). On the basis of structural alignments, mFAS KS is closely related to KS domains from the DEBS I PKS system [1.3 Å root mean square deviation (RMSD)] but is structurally also very similar to fungal KS (RMSD 1.8 Å). It is more closely related to bacterial KASI and KASII (1.8 Å/1.6 Å RMSD) than to KASIII (2.9 Å RMSD), reflecting its ability to elongate long ACP-bound acyl chains.

Despite the pronounced structural similarity between the KS domains of mFAS and PKS, the selectivity of FAS for saturated acyl chains can be explained by a considerable constriction at the base of the active site phosphopantetheine binding pocket leading into the large acyl chain substrate binding tunnel, which connects both active sites of the KS dimer (Fig. 4, A and B). A number of residues lining this narrow tunnel are highly conserved in mFAS but are substituted with smaller residues in all DEBS PKS modules (Fig. 4B), resulting in a wider, more permissive tunnel in the KS of DEBS.

*Malonyl-acetyl transferase.* The acyl transferases of FAS and PKS systems are composed of an α/β-hydrolase core fold and a ferredoxin-like subdomain, which together create the active site cleft (*32*). In mFAS, the MAT domains of the two monomers have slightly different relative orientations of the two subdomains, probably selected by crystal-packing interactions. The structurally closest relatives are acyl transferases from DEBS didomain structures (2.0 Å/2.2 Å RMSD), the human mitochondrial MT (2.2 Å), and the *Escherichia coli* FAS MT (FabD) (2.3 Å RMSD) (table S4). All members of this family share a conserved active site with a catalytic Ser-His dyad. MTs are distinguished from acetyl transferases by the presence of a conserved active site arginine (*2*), which forms a bidentate salt bridge with the malonyl carboxylate (*33*). The mFAS MAT and related MTs display a rather

broad specificity for malonyl-CoA derivatives (e.g., propionyl-CoA and methylmalonyl-CoA). However, only mFAS MAT uses both acetyl-CoA and malonyl-CoA with equal efficiency (*6*). A structural comparison with bacterial and PKS MTs reveals three candidate residues for this specificity (Fig. 4C). Phe$^{682}$ replaces a serine in most PKS and bacterial homologs, whereas Phe$^{553}$ and Met$^{499}$ substitute for conserved glutamine residues. Together, these three substitutions create a considerably more hydrophobic active site. The two phenylalanines form a hydrophobic cavity, which may allow Met$^{499}$ to flip onto the methyl group of an acetyl substrate. Thus, the dual specificity of mFAS MAT appears to result from the combined presence of the conserved arginine for salt-bridging malonyl substrates and the more hydrophobic nature of the active site. The double specificity can be changed by mutating the arginine to alanine, which then transforms the MAT into an acetyl transferase (*34*).

*Ketoreductase.* The NADPH-dependent KR domain belongs to the family of short-chain dehydrogenases/reductases (SDRs) (*35*)—single-domain proteins that have a characteristic Rossmann fold and a substrate binding extension inserted before the last helix. mFAS KR is structurally closely related to both the tetrameric bacterial KR (FabG, 2.0 Å RMSD) and ER (FabI, 2.2 Å RMSD) and to the fungal KR domain (2.6 Å RMSD) (table S4). As for the KS and MAT domains, the closest structural homology is observed with the KR domains from modular PKS (RMSDs of 1.5 Å for tylosin PKS KR and 1.6 Å for DEBS KR). The arrangement of residues in the active site of the mFAS KR domain is consistent with a proton-relay mechanism described for bacterial FabG (*36*). However, two residues of the proton-wire, Asn$^{2038}$ and Lys$^{1995}$, have swapped positions (Fig. 4D), as previously observed in PKS KR domains (*19*, *20*). Loops in the vicinity of the active site cleft are disordered in the apo form of mFAS and become stabilized upon cofactor binding. This includes residues 1975 to 1990, corresponding to the β4/α4 loop in FabG, that are presumably stabilized by interactions of Met$^{1973}$ with the active site Lys$^{1995}$, and part of the substrate binding extension (residues 2072 to 2075). The direction of substrate entry into the active site can be inferred from the stereospecificity of mFAS KR, which produces an *R*-hydroxyl group (*6*, *20*). The substrate approaches the NADPH cofactor from above the nicotinamide ribose, as observed for the structurally related mycobacterial ER InhA (*37*, *38*).

*Dehydratase.* The mFAS DH domain adopts a pseudodimeric double hot dog fold (Fig. 5A). The subdomain arrangement is more similar to

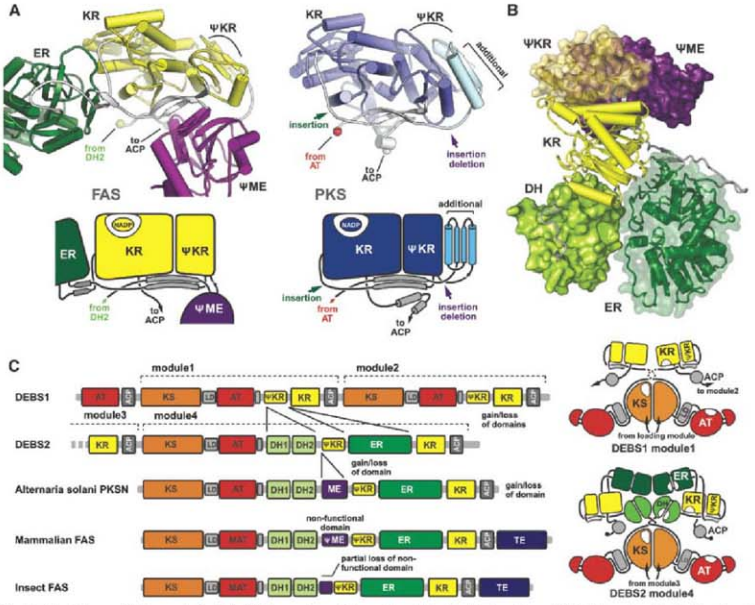



**Fig. 3.** Modularity of the modifying part of mFAS. **(A)** Comparison of the KR/ΨKR arrangement in mFAS (at left) and a related polyketide synthase, DEBS1 (*19*) (at right). The lower panels provide a schematic overview. The only modifying domain present in DEBS1 PKS includes KR-ΨKR, which is N- and C-terminally extended compared with mFAS KR/ΨKR. The additional ER and DH domains are integrated into mFAS without disturbing the path of the KR/ΨKR linker compared with PKS, leaving the N and C termini in identical positions. **(B)** KR interacts with all other domains in the modifying part of mFAS, whereas no direct interactions occur between either of the nonenzymatic domains and DH or ER. **(C)** Schematic sequence diagram depicting the integration and removal of additional domains in the modifying parts of PKS, mFAS, and insect FAS, on the basis of structural and sequence alignments (at left). The putative domain topologies of DEBS modules 1 and 4 are shown schematically on the right.

the functional DH dimers of bacterial FabA/Z than to the pseudodimeric fungal DH (fig. S9 and table S4) (*39*, *40*). Each mFAS pseudodimer contains a single composite active site formed by residues $His^{878}$ from the N-terminal hot dog fold and $Asp^{1033}$ and $His^{1037}$ from the C-terminal fold (Fig. 5B). The catalytic importance of these amino acids in mFAS has been demonstrated by mutagenesis and is further corroborated by a topologically similar arrangement in the active centers of bacterial DHs (Fig. 5B) (*39*–*42*). This also suggests a similar two-base reaction mechanism, as proposed for *E. coli* FabA, with $His^{878}$ and $Asp^{1033}$ participating in substrate protonation and deprotonation (*39*). The histidine at position 1037 is only present in the active site of chicken and pig FAS, whereas in all other sequenced mFAS the corresponding amino acid is a glutamine (fig. S10). The equivalence of a histidine and a glutamine at this position has been verified by mutagenesis (*42*). In the mFAS structure, $His^{1037}$ is positioned toward $Asp^{1033}$ at hydrogen-bonding distance, indicating a stabilizing function similar to those of the glutamine in other mFAS and bacterial DHs (*40*).

A hydrophobic substrate binding tunnel starts at the pseudodimer interface, stretches through the C-terminal hot dog domain, and has an open end that points toward the top of the FAS assembly (Fig. 5A). In contrast to type II DHs, which harbor two equivalent active sites in each homodimer, the second catalytic site is inactive in mFAS: The loop harboring the second catalytic histidine in bacteria is reduced to a short turn, and the corresponding accessory catalytic residues located in the central helix of the fold are replaced by tryptophane and lysine (Fig. 5A). The hydrophobic tunnel is entirely absent, and the domain is truncated by 30 residues at the N terminus.

*Enoylreductase.* In contrast to all other functional domains of the fatty acid elongation cycle, the mFAS ER has a different fold from its functional analogs in the bacterial type II FAS system, where the ERs are either SDR [FabI, FabL, FabV (*43*–*45*)] or TIM barrel proteins (FabK)—the latter also found as an ER domain in fungal type I FAS (*2*, *46*). Instead, the mammalian ER establishes a subfamily of medium-chain dehydrogenases/reductases (MDRs) (*47*) that is structurally related to bacterial quinone oxidoreductase (table S4). The mFAS ER contains two subdomains, a nucleotide binding Rossmann-fold (residues 1651 to 1794) and a substrate binding portion (residues 1530 to 1650 and 1795 to 1858). It binds the $NADP^+$ cofactor in an open extended conformation between the two subdomains (Fig. 5C). Our structure identifies $Lys^{1771}$ and $Asp^{1797}$ as candidate donor residues for substrate protonation after hydride transfer from NADPH (Fig. 5D). These two residues are in close proximity to the hydride-donating nicotinamide C4, in a similar position as the suggested active site tyrosines in other MDR subfamilies (e.g., the mitochondrial ER) (*48*–*50*). The two residues are strictly conserved in mFAS (fig. S11), and a corresponding lysine/aspartate pair is observed in the apo-form structure of the nucleotide binding subdomain of a related type I PKS ER domain (PDB entry 1pqw). The active site of ER is located in a narrow crevice created in part by the bound nucleotide cofactor, very different from the substrate binding groove in the related quinone reductase (*50*). Substrate entry probably occurs through a tunnel along the cofactor toward the nicotinamide ring. The tunnel continues through a constriction toward the back of the ER domain, where an opening would allow exit of long acyl chains (Fig. 5C).

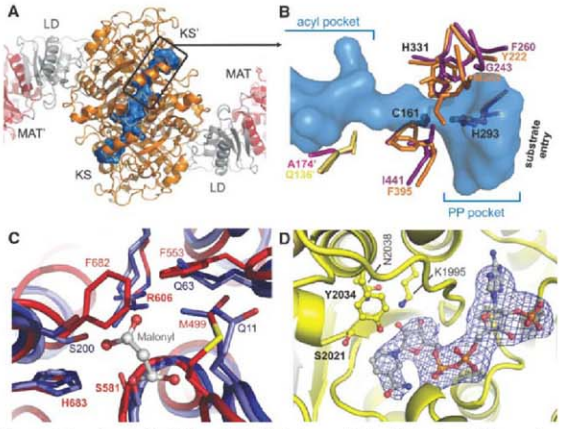



**Fig. 4.** Active sites of KS, MAT, and KR. **(A)** A large substrate binding tunnel (blue surface representation) traverses the dimeric KS domain. The substrate entry site of the KS domains is oriented toward the MAT domain of the other chain. **(B)** A narrow constriction in the substrate binding tunnel of the KS domain adjacent to the conserved active site residues (labeled in black) prevents the entry of larger modified substrates. Four residues involved in the formation of the constriction (orange for one subunit and yellow for the second one) are conserved in mFAS and replaced with smaller residues in the more permissive KS domains of PKS, as exemplified by KS of DEBS module 5 (purple for one subunit and pink for the second) (*18*). **(C)** MAT active site of mFAS (red) compared to bacterial MT, with and without bound substrate (light and dark blue, respectively) (*32*, *33*). Conserved active site residues in mFAS are indicated in bold. Two phenylalanines and a methionine characteristic for the acetyl-CoA/malonyl-CoA double-specific mFAS create a more hydrophobic binding groove and may close onto the methyl group of an acetyl moiety to promote efficient binding. **(D)** Active site of KR with bound $NADP^+$ and a 3.3 Å unbiased simulated annealed omit electron density map for the cofactor, contoured at 2.7σ. The proton-donating tyrosine is in equivalent position to bacterial homologs, but the asparagine ($N^{2038}$) and lysine ($K^{1995}$) involved in proton replenishment are swapped.

### Structural Relation to PKS

The structural information presented here provides extensive evidence for the evolutionary relationship between mFAS and bacterial and fungal PKSs. (i) All catalytic mFAS domains are most closely related to PKS domains at the sequence level. Notably, the domains of mFAS are more similar to PKS domains than to the bacterial FAS counterparts, despite the differences in their substrate specificities. (ii) The structure of mFAS demonstrates that the similarity to PKSs extends to the linkers and the overall architecture—e.g., those in the KS/MAT or the KR/ΨKR regions—despite very low sequence conservation. (iii) Finally, an additional piece of evidence is provided by the existence of the nonfunctional ΨME domain, which can be considered a remnant of a catalytically active domain present in a common evolutionary ancestor of mFAS and PKS that is still preserved in several PKSs (*28*–*30*).

Functionally, modular PKSs differ from FAS by their non-iterative mode of action, where each module carries out a single precursor elongation step equivalent to one round of chain elongation by mFAS (*51*). The modules are concatenated into large polypeptides, several of which may assemble into production lines with more than 10 modules. This allows the synthesis of a variety of structurally diverse compounds. Based on the

structural similarity discussed above, mFAS can be considered as a single PKS module specialized for iterative fatty acid synthesis.

Individual PKS modules contain substrate loading and condensing domains and variations of domains involved in β-carbon processing that control the chemical structure of the produced polyketide. Overall comparison of the PKS DEBS module 1 KR-ΨKR, which is the only β-carbon–modifying domain in this module, with the corresponding parts of mFAS reveals the structural basis for the notable modularity of modifying domains in megasynthases: Mediated by short linking sequences emanating from conserved secondary structure elements, the full ER domain is inserted between the ΨKR and KR domain, whereas the ΨME domain is integrated into the linker leading into the N terminus of the ΨKR domain (Fig. 3, A and C). As a consequence of this architectural solution, the insertions do not affect the core folds of the ΨKR and KR domains. Furthermore, because of a very flexible mode of interaction between KS and DH and weak contacts between KR and either DH or ER (table S3) of the mFAS, it is relatively easy to envision the architecture of some representative PKS modules (Fig. 3C). For example, the minimal PKS module, such as module 1 of DEBS1, which includes only the KR of the possible β-carbon–processing domains, would have this domain linked to KS with a short 10–amino acid linker, similar to the linker connecting KS to DH in mFAS. Other truncated variants of the mFAS architecture are also detectable at the sequence level. In the case of the closely related insect FAS, the ΨME domain lacks the N-terminal extension (Fig. 3C and fig. S8), whereas there is no methyltransferase in DEBS module 4 (Fig. 3C) (*19*, *52*). Short extensions may substitute for missing domain interactions, as indicated by a C-terminal addition of two helices to the KR-ΨKR of DEBS 1 (*19*) (Fig. 3A), which covers the region of the ΨKR surface occupied by the ΨME in mFAS.

These results also imply that the iterative mode of elongation (in which ACP shuttles substrates within one module) and a noniterative elongation (where substrates are passed between modules) can be accomplished with a similar overall architecture of the molecules. Notably, compared to the mFAS structure, no supplementary elements with a potential role in oligomerization are observed in the PKS structures available so far, except for a single helix in the KS-MAT linker domain (fig. S1) and a small C-terminal extension of the KR domain (Fig. 3A). Apparently, the N- and C-terminal docking domains (*53*, *54*) are sufficient to determine the higher-order assembly of PKS modules and polypeptides.

The architecture and the fold of mFAS and of the related PKSs are extremely versatile. This is in contrast to fungal FAS, which forms a barrel-shaped 2.6 MD $\alpha_6\beta_6$-heterododecameric assembly with three full sets of active sites enclosed in each of the two reaction chambers (*2*, *13*). The scaffold of the cage-forming fungal FAS appears less tolerant toward product modifying domain insertions and excisions because of symmetry constraints and the tight embedding of catalytic domains. Consistently, no naturally occurring fungal type I FAS with an altered domain composition has been detected, and all fungal FAS and most of their homologs produce only saturated fatty acid products (*8*).

### Substrate Shuttling by the ACP Domain

The entry sites to the active centers of the mFAS enzymatic domains are grouped around the two lateral clefts. In each cleft, the entry sites of MAT, DH, and ER are oriented toward one face of mFAS, and those of KS and KR toward the other face (Fig. 6). The flexibly tethered ACP and the following TE domains are not visualized in the structure. However, the structure defines the anchor point of ACP at residue 2113 in the center of the upper portion of the lateral clefts of mFAS. Together with recent structures of closely related rat and human ACP (*14*, *15*), which define the first ordered ACP residue at positions 2125 to 2127 (porcine FAS numbering), the flexible KR-ACP linker is composed of 12 to 14 amino acids, corresponding to a maximum length of ~40 Å. With a length of 23 to 26 residues, the ACP-TE linker is substantially longer and could span up to an 80 Å distance, as deduced from the corresponding isolated domain structures (*14*–*17*). In contrast to the fungal ACP linkers, which have a high Pro/Ala content that can increase their stiffness (*55*), the tethers flanking mFAS ACP have no unusual amino acid composition (Fig. 6). Whereas the motion of fungal ACP is constrained by double-tethering, no second anchor point is apparent for mFAS ACP, because the subsequent TE domain is not located at a defined position relative to the body of mFAS. Still, the TE domain may influence the motion of the tethered ACP either by transiently interacting with other domains or by steric effects.

Confining the path of the ACP may be one role of the protruding noncatalytic ΨME domain. This domain is docked to the body of mFAS in the vicinity of the ACP anchor point (Fig. 1, A

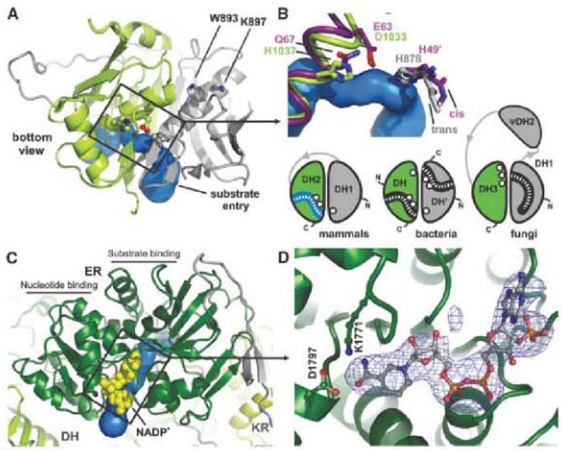



**Fig. 5.** Active sites of DH and ER. **(A)** Pseudo-dimeric DH domain only harbors a single active site and substrate binding tunnel with an open end. Active site residues are shown in ball-and-stick representation. These residues are not conserved in the corresponding position of the second subdomain ($Trp^{893}$ and $Lys^{897}$). **(B)** Close-up view of the DH active site topology (top) and schematic comparison to bacterial and fungal FAS (bottom). The mFAS active site residues (green) have their functional groups in similar positions as their bacterial FabZ counterparts (*40*) (purple), despite the exchange of two amino acids and the loss of an unusual non-proline cis peptide bond at $H^{878}$. **(C)** Open-ended substrate binding tunnel (blue) of the ER domain of mFAS shown in the presence of the bound $NADP^+$ cofactor (yellow spheres). **(D)** Two amino acids, $Asp^{1797}$ and $Lys^{1771}$, are candidate proton donor residues for enoyl reduction based on the positioning of their functional groups at ~4.2 Å distance to the C4 of the $NADP^+$ nicotinamide ring. A 3.3 Å unbiased simulated annealed omit electron density map for the bound $NADP^+$ cofactor, contoured at 3.3σ, is shown.

and B) and narrows the accessible region along the MAT, KS, DH, and ER substrate entry sites to a rim that is just slightly wider than the ACP (Fig. 6 and fig. S7C). The discrete ACPs of bacterial-type FAS (*56*–*58*) sequester fully saturated fatty acyl chains within their hydrophobic core. In contrast, mFAS ACP does not bury the acyl chain inside its core (*14*), raising the question whether other parts of the molecule may have taken over this function (for instance, by providing hydrophobic rims for a sliding motion of ACP-tethered acyl chains on the surface of mFAS). However, mapping of conservation or electrostatic potential on the mFAS surface did not reveal such regions.

The linker length, together with the steric constraints of the mFAS structure, allows ACP to reach a full set of active sites in one cleft but no active sites from the other cleft. During the elongation cycle (Fig. 6), the ACP is first loaded with substrates at the lateral MAT domain. ACP then has to deliver the substrates to the KS entry pocket on the opposite face of mFAS. The shortest route would lead directly through the cleft (path 1 in Fig. 6), which is just sufficiently open to allow the passage of ACP. After condensation at the KS, the ACP must reach the KR on the same face as the KS active site (path 2) before crossing the cleft again to approach the DH domain (path 3). From here, ACP proceeds toward the nearby ER (path 4) and finally delivers the fully saturated substrate to the KS active center before being reloaded at the MAT for the next cycle. During this cycle, the ACP interacts with the MAT and the β-carbon–processing domains of one chain, but it also interacts with the KS of the second polypeptide chain in the FAS dimer. Notably, the partially preserved active site cleft of the catalytically inactive ΨME domain could easily be accessed by ACP from the reaction chamber, as required in PKSs that display methyltransferase activity (fig. S7C) (*28*, *30*).

The requirement for ACP to shuttle back and forth through the cleft does not appear to be the most favorable solution for efficient substrate transfer and catalysis by mFAS. An alternative is suggested by considering the properties of the junction between the lower condensing and the upper modifying part of mFAS: They are joined only via the MAT-DH linkers, which are expected to mediate a flexible junction between the two halves of ~200 kD molecular mass each (*59*). Moreover, the pseudosymmetry-related DH/KS contacts on either side of the joint are not identical as would be expected for a stable interaction in solution. Consequently, the flexible connection of the mFAS halves may allow rotational motion around the dimer axis or a certain degree of tilting. Such motion would drag the ACP between the two faces of FAS and may contribute considerably to productive substrate shuttling. Mutant complementation and cross-linking studies have demonstrated that the vast majority of sub-

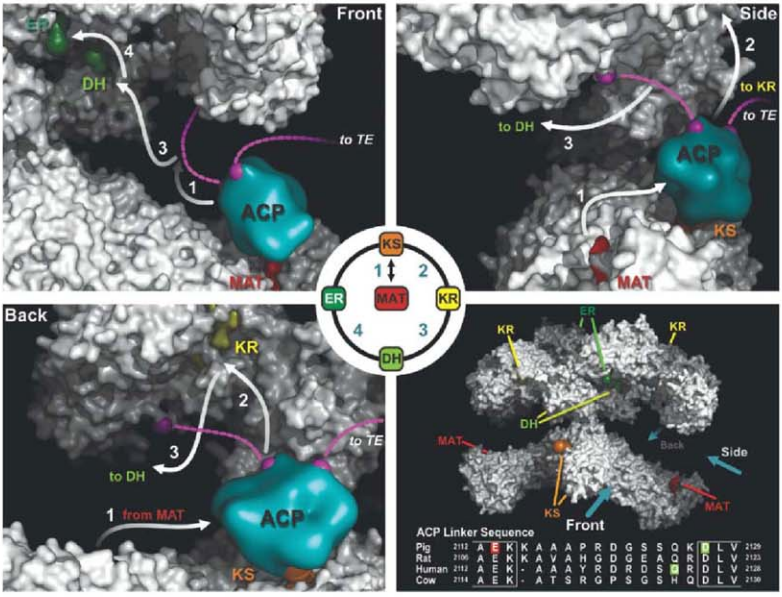



**Fig. 6.** Substrate shuttling by the ACP in mFAS. After substrate loading at the MAT on one side of the reaction chamber, the flexibly linked ACP has to shuttle the substrates to the other side for condensation at the KS and reduction at the KR (paths 1 and 2). To reach the DH and ER domain, ACP has to cross the cleft a second time (paths 3 and 4) before the saturated acyl chain can be back-transferred to KS to serve as primer for the next elongation cycle. The flexible linkers of ACP are depicted by dashed lines (pink). The precise length of the KR-ACP linker is defined by the KR and ACP domain borders in the current and previously solved structures (*14*, *16*, *17*) (bottom right, red and green). A schematic representation of the mFAS ACP domain based on experimentally determined structures (*14*, *15*) was positioned to the active site clefts by superposition with fungal ACP bound to the KS domain (*13*) and by orienting it according to residual electron density observed in the active site cleft of MAT (fig. S12).

strates are processed in mFAS by individual full sets of active sites, according to the path of ACP described above. However, these studies have also shown that a minority of substrates can be shuttled between the two sets of active sites, either by ACP serving both MAT domains or by direct interaction of ACP with both KS domains (*6*, *60*–*62*). In light of the large 135 Å distance between the ACP anchor point located in one catalytic cleft and the MAT in the other, the most plausible explanation for the minor mode-of-domain interaction is a large-scale rotation of the upper portion of mFAS, relative to the lower portion (fig. S4).

The molecular description of active sites in mFAS should stimulate the development of improved inhibitors as anticancer drug candidates. As demonstrated by structural homology, this structure is also a good template for the organization of PKS modules; it agrees with and extends present theoretical models of PKS architecture (*19*, *22*). Furthermore, the structure of mFAS paves the way for structure-based experiments to answer remaining questions on the dynamics and substrate shuttling mechanism in megasynthases.

### References and Notes


1. S. W. White, J. Zheng, Y. M. Zhang, C. O. Rock, *Annu. Rev. Biochem.* **74**, 791 (2005).
2. S. Jenni *et al.*, *Science* **316**, 254 (2007).
3. S. Jenni, M. Leibundgut, T. Maier, N. Ban, *Science* **311**, 1263 (2006).
4. I. B. Lomakin, Y. Xiong, T. A. Steitz, *Cell* **129**, 319 (2007).
5. T. Maier, S. Jenni, N. Ban, *Science* **311**, 1258 (2006).
6. S. Smith, S. C. Tsai, *Nat. Prod. Rep.* **24**, 1041 (2007).
7. F. Lynen, *Eur. J. Biochem.* **112**, 431 (1980).
8. E. Schweizer, J. Hofmann, *Microbiol. Mol. Biol. Rev.* **68**, 501 (2004).
9. F. P. Kuhajda *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 6379 (1994).
10. J. A. Menendez, R. Lupu, *Nat. Rev. Cancer* **7**, 763 (2007).
11. F. P. Kuhajda, *Cancer Res.* **66**, 5977 (2006).
12. H. Orita *et al.*, *Clin. Cancer Res.* **13**, 7139 (2007).
13. M. Leibundgut, S. Jenni, C. Frick, N. Ban, *Science* **316**, 288 (2007).
14. E. Piskon *et al.*, *J. Biol. Chem.* **283**, 518 (2008).
15. G. Bunkoczi *et al.*, *Chem. Biol.* **14**, 1243 (2007).
16. B. Chakravarty, Z. Gu, S. S. Chirala, S. J. Wakil, F. A. Quiocho, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15567 (2004).
17. C. W. Pemble IV, L. C. Johnson, S. J. Kridel, W. T. Lowther, *Nat. Struct. Mol. Biol.* **14**, 704 (2007).
18. Y. Tang, C.-Y. Kim, I. I. Mathews, D. E. Cane, C. Khosla, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11124 (2006).
19. A. T. Keatinge-Clay, R. M. Stroud, *Structure* **14**, 737 (2006).
20. A. T. Keatinge-Clay, *Chem. Biol.* **14**, 898 (2007).
21. Y. Tang, A. Y. Chen, C.-Y. Kim, D. E. Cane, C. Khosla, *Chem. Biol.* **14**, 931 (2007).
22. C. Khosla, Y. Tang, A. Y. Chen, N. A. Schnarr, D. E. Cane, *Annu. Rev. Biochem.* **76**, 195 (2007).
23. A. K. Joshi, A. Witkowski, H. A. Berman, L. Zhang, S. Smith, *Biochemistry* **44**, 4100 (2005).
24. Materials and methods are available as supporting material on *Science* Online.
25. J. G. Olsen, A. Kadziola, P. von Wettstein-Knowles, M. Siggaard-Andersen, S. Larsen, *Structure* **9**, 233 (2001).
26. J. M. Thorn, J. D. Barton, N. E. Dixon, D. L. Ollis, K. J. Edwards, *J. Mol. Biol.* **249**, 785 (1995).
27. J. L. Martin, F. M. McMillan, *Curr. Opin. Struct. Biol.* **12**, 783 (2002).
28. I. Fujii, N. Yoshida, S. Shimomaki, H. Oikawa, Y. Ebizuka, *Chem. Biol.* **12**, 1301 (2005).
29. I. Molnar *et al.*, *Chem. Biol.* **7**, 97 (2000).
30. D. J. Edwards *et al.*, *Chem. Biol.* **11**, 817 (2004).
31. A. Witkowski, A. K. Joshi, S. Smith, *Biochemistry* **36**, 16338 (1997).
32. L. Serre, E. C. Verbree, Z. Dauter, A. R. Stuitje, Z. S. Derewenda, *J. Biol. Chem.* **270**, 12961 (1995).
33. C. Oefner, H. Schulz, A. D'Arcy, G. E. Dale, *Acta Crystallogr. D Biol. Crystallogr.* **62**, 613 (2006).
34. V. S. Rangan, S. Smith, *J. Biol. Chem.* **272**, 11975 (1997).
35. B. Persson, Y. Kallberg, U. Oppermann, H. Jornvall, *Chem. Biol. Interact.* **143-144**, 271 (2003).
36. A. C. Price, Y. M. Zhang, C. O. Rock, S. W. White, *Structure* **12**, 417 (2004).
37. D. A. Rozwarski, C. Vilcheze, M. Sugantino, R. Bittman, J. C. Sacchettini, *J. Biol. Chem.* **274**, 15582 (1999).
38. T. J. Sullivan *et al.*, *Am. Chem. Soc. Chem. Biol.* **1**, 43 (2006).
39. M. Leesong, B. S. Henderson, J. R. Gillig, J. M. Schwab, J. L. Smith, *Structure* **4**, 253 (1996).
40. M. S. Kimber *et al.*, *J. Biol. Chem.* **279**, 52593 (2004).
41. A. K. Joshi, S. Smith, *J. Biol. Chem.* **268**, 22508 (1993).
42. S. Pasta, A. Witkowski, A. K. Joshi, S. Smith, *Chem. Biol.* **14**, 1377 (2007).
43. C. Baldock, J. B. Rafferty, A. R. Stuitje, A. R. Slabas, D. W. Rice, *J. Mol. Biol.* **284**, 1529 (1998).
44. R. J. Heath, N. Su, C. K. Murphy, C. O. Rock, *J. Biol. Chem.* **275**, 40128 (2000).
45. R. P. Massengo-Tiasse, J. E. Cronan, *J. Biol. Chem.* **283**, 1308 (2008).
46. J. Saito *et al.*, *Protein Sci.* **17**, 691 (2008).
47. E. Nordling, H. Jornvall, B. Persson, *Eur. J. Biochem.* **269**, 4267 (2002).
48. T. T. Airenne *et al.*, *J. Mol. Biol.* **327**, 47 (2003).
49. T. Hori *et al.*, *J. Biol. Chem.* **279**, 22615 (2004).
50. Y. Shimomura, Y. Kakuta, K. Fukuyama, *J. Bacteriol.* **185**, 4211 (2003).
51. D. A. Hopwood, D. H. Sherman, *Annu. Rev. Genet.* **24**, 37 (1990).
52. C. D. Richter *et al.*, *FEBS J.* **274**, 2196 (2007).
53. C. D. Richter, D. Nietlispach, R. W. Broadhurst, K. J. Weissman, *Nat. Chem. Biol.* **4**, 75 (2008).
54. K. J. Weissman, *ChemBioChem* **7**, 485 (2006).
55. R. N. Perham, *Biochemistry* **30**, 8501 (1991).
56. A. Roujeinikova *et al.*, *Structure* **10**, 825 (2002).
57. A. Roujeinikova *et al.*, *J. Mol. Biol.* **365**, 135 (2007).
58. G. A. Zornetzer, B. G. Fox, J. L. Markley, *Biochemistry* **45**, 5217 (2006).
59. S. Jones, J. M. Thornton, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13 (1996).
60. S. Smith, A. Witkowski, A. K. Joshi, *Prog. Lipid Res.* **42**, 289 (2003).
61. A. K. Joshi, V. S. Rangan, A. Witkowski, S. Smith, *Chem. Biol.* **10**, 169 (2003).
62. V. S. Rangan, A. K. Joshi, S. Smith, *Biochemistry* **40**, 10792 (2001).


63. All data were collected at the Swiss Light Source (SLS, Paul Scherrer Institute, Villigen). We thank C. Schulze-Briese, S. Gutmann, R. Bingel-Erlenmeyer, S. Russo, A. Pauluhn, and T. Tomizaki for their outstanding support at the SLS; S. Jenni and M. Sutter for critically reading the manuscript and all members of the Ban laboratory for suggestions and discussions; R. Grosse-Kunstleve, P. Afonine, and P. Adams for support with the PHENIX software; and A. Jones for support with the program O. This work was supported by the Swiss National Science Foundation (SNSF) and the National Center of Excellence in Research Structural Biology program of the SNSF. Structure factors and atomic coordinates of the porcine FAS in the apo- and NADP$^+$-bound form have been deposited in the Protein Data Bank with accession codes *2vz8* and *2vz9*.


### Supporting Online Material

www.sciencemag.org/cgi/content/full/321/5894/1315/DC1
Materials and Methods
Figs. S1 to S15
Tables S1 to S4
References

3 June 2008; accepted 31 July 2008
10.1126/science.1161269

# Internally Generated Cell Assembly Sequences in the Rat Hippocampus

**Eva Pastalkova, Vladimir Itskov,* Asohan Amarasingham, György Buzsáki†**

**A long-standing conjecture in neuroscience is that aspects of cognition depend on the brain's ability to self-generate sequential neuronal activity. We found that reliably and continually changing cell assemblies in the rat hippocampus appeared not only during spatial navigation but also in the absence of changing environmental or body-derived inputs. During the delay period of a memory task, each moment in time was characterized by the activity of a particular assembly of neurons. Identical initial conditions triggered a similar assembly sequence, whereas different conditions gave rise to different sequences, thereby predicting behavioral choices, including errors. Such sequences were not formed in control (nonmemory) tasks. We hypothesize that neuronal representations, evolved for encoding distance in spatial navigation, also support episodic recall and the planning of action sequences.**

A prominent theory states that the hippocampal system primarily serves spatial navigation (*1*, *2*); a component of this theory is that the place-dependent activity of neurons [place cells (*1*, *2*)] in the hippocampus arises from external serially ordered environmental stimuli (*3*–*7*). Place cells are thought to embody the representation of a cognitive map, enabling flexible navigation. However, neural theories of other cognitive processes that may depend on the hippocampus, such as episodic memory and action planning, draw on the activity of hypothetical internally organized cell assemblies (*8*–*13*).

Center for Molecular and Behavioral Neuroscience, Rutgers, The State University of New Jersey, 197 University Avenue, Newark, NJ 07102, USA.

*Present address: Center for Neurobiology and Behavior, Columbia University, 1051 Riverside Drive, New York, NY 10032, USA.

†To whom correspondence should be addressed. E-mail: buzsaki@axon.rutgers.edu

Several observations have refined the navigation theory. Hippocampal neurons can predict where the animal is coming from, or its destination (*14*–*17*); the sequential activity of place cells during locomotion is replicated within single cycles of the theta oscillation (8 to 12 Hz) (*18*–*20*); furthermore, the temporal recruitment of active neurons in the population bursts of rest and sleep also reflects, again on a faster time scale, their sequential activity as place cells during locomotion (*21*–*23*). Thus, the sequential activation of hippocampal neurons can be disengaged from external landmarks (*24*, *25*). However, internally generated assembly sequences operating at the time scale of behavior have not yet been reported.

The frameworks of environment-controlled versus internally generated assembly sequences give rise to distinct predictions. Imagine that a rat is frozen in position during its travel (and yet the theta oscillation is maintained). According to the navigation theory, a subset of landmark-controlled place cells should then display sustained activity, and other neurons would remain suppressed (*2*–*6*). In contrast, if assembly sequences were generated by internal mechanisms, neurons might rather display continually changing activity. We tested these predictions by examining the activity of hippocampal neurons while a rat was running in a wheel at a relatively constant speed (*26*, *27*) during the delay of a hippocampus-dependent alternation memory task.

**Internally generated cell-assembly sequences.** Rats were trained to alternate between the left and right arms of a figure-eight maze [Fig. 1A and supporting online material (SOM) text]. During the delay period between maze runs (10 s for rat 1; 20 s each for rats 2 and 3), the animals were trained to run steadily in the same direction in a wheel (Fig. 1A). To confront the predictions of the navigation theory with those of the internal sequence-generation hypothesis, we compared

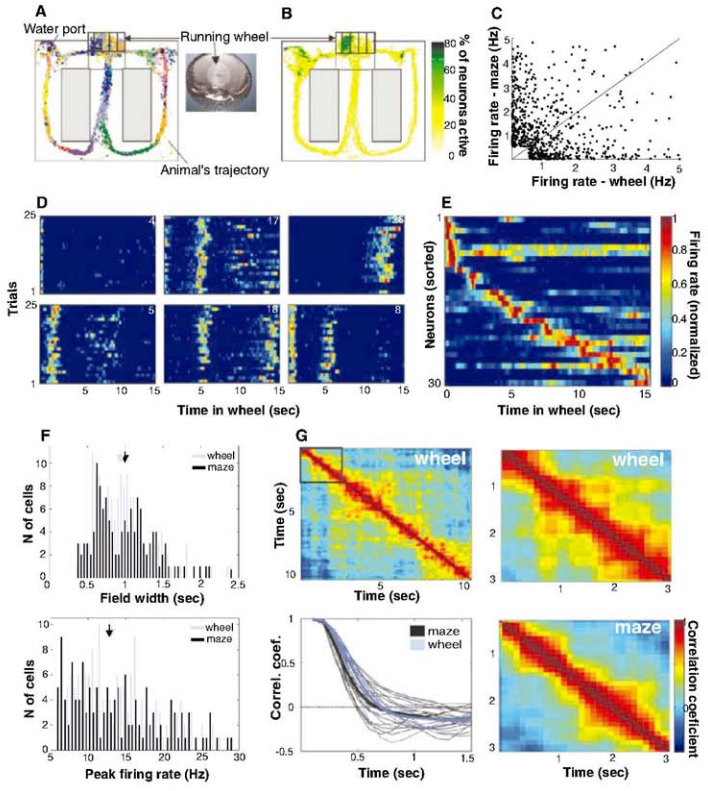



**Fig. 1.** Episode fields in the wheel and place fields in the maze are similar. **(A)** Color-coded spikes (dots) of simultaneously recorded hippocampal CA1 pyramidal neurons. The rat was required to run in the wheel facing to the left during the delay between the runs in the maze. **(B)** Percent of neurons firing >0.2 Hz within each pixel. The highest percentage of neurons was active when rats were running in the wheel. **(C)** Relationship between firing rate of neurons active in rats running the wheel and the maze ($r_s = -0.3$, $P < 0.0001$, 681 neurons, three rats, 17 sessions). **(D)** Normalized firing rate of six simultaneously recorded neurons during wheel running (each line shows the color-coded activity on single trials turning to the left arm). The episode fields occurred at specific segments of the run. **(E)** Normalized firing rate of 30 simultaneously recorded neurons during wheel running, ordered by the latency of their peak firing rate. **(F)** Width (top) and peak firing rate (bottom) of episode and place fields (wheel, $n = 135$ neurons; maze, $n = 162$ neurons). Arrows indicate medians. **(G)** Population vector cross-correlation matrix (SOM text). The width of the diagonal stripe indicates the rate at which neuronal assemblies transition. (Lower left) The decay of the population vector correlation during wheel running and maze traversal. Thin lines, individual sessions; thick lines, group means.

the firing patterns of CA1 hippocampal neurons in rats running the wheel and the maze.

We analyzed the activity of ~500 pyramidal cells recorded in the wheel and ~600 neurons in the maze (mean firing rate >0.5 Hz) (Fig. 1A). Pyramidal neurons were transiently active in rats running both the maze [place cells (*1*)] and the wheel. Although the position and direction of the rat's head were stationary during wheel running (fig. S1), the percentage of neurons active in the pixels occupied by the head during wheel running was three to four times greater than in any area of comparable size in the maze (Wilcoxon rank sum test, $P < 0.0001$) (Fig. 1B). Thus, if pyramidal neurons were solely activated by environmental cues (*2*–*6*), this finding would reflect several-fold–stronger neuronal representation of the animal's position within the wheel. Many individual pyramidal cells were active both in rats running the wheel and rats running the maze, but the sequential order of their activation in rats in the wheel was unrelated to that of rats in the maze, and their firing rates in these two areas were inversely correlated [Spearman correlation coefficient ($r_s$) = –0.3, $P < 0.0001$, $n = 681$ neurons (Figs. 1C and 4B); contrast this with the population of interneurons, $r_s = 0.85$, $P < 0.0001$, $n = 125$ interneurons (fig. S2)]. The average proportion of pyramidal neurons simultaneously active [firing at least a single spike in 100-ms windows (averaging over 100-ms windows)] was similar in the wheel (10.75 ± 3.97%) and the maze (12.56 ± 4.32%) (fig. S3).

Pyramidal neurons typically fired transiently, and reliably in successive trials, at specific times of wheel running (episode fields), and most cells had multiple peaks of varying sizes (Fig. 1D). Typically, and reminiscent of a synfire chain (*11*), at least one episode cell was active at every moment of a wheel run (Fig. 1E).

Were episode cells in rats in the wheel generated by the same mechanism as place cells in rats in the maze? We looked for evidence of differing mechanisms by comparing several measures of the firing of episode and place cells. First, we calculated the duration of activity (field width) (Fig. 1F) of single cells [including only fields with a peak firing rate of ≥6.0 Hz and ≥4.5 SD above the mean firing rate (SOM text)]. The temporal and spatial extent of the field was determined as those times and positions at which firing rates were at least 10% that of the peak firing rate (in the wheel or maze) (*19*, *28*). By these criteria, 32% of the neurons recorded in the wheel and 22% in the maze had at least one field. Neither the distribution of field widths (medians were 0.94 and 1.0 s, respectively; Wilcoxon test, $P = 0.44$) nor peak firing rates (medians were 13.08 and 12.8 Hz, respectively; $P = 0.61$) differed significantly between the episode and place fields (Fig. 1F). Second, to measure the average lifetime of assembly activity for a population, we determined the maximal time lag at which the autocorrelation of the population's activity was above 0.5 (*29*) and again found no significant difference, with respect to the median, between the populations of episode and place cells (medians were 0.83 and 0.75 s, respectively; $P = 0.32$) (Fig. 1G). Third, we compared the relationship between spikes and the local field potential in episode and place cells. On linear tracks, sequentially generated spikes of a place cell gradually shift to earlier and earlier phases of the theta oscillation as the rat passes through the place field (phase precession), and there is a systematic relationship between the phase of spikes and the animal's position (*3*, *18*–*20*, *28*, *30*, *31*). The navigation theory predicts that the phase of spikes will remain fixed if environmental inputs do not change (*3*, *26*, *27*). In contrast, episode cells displayed phase precession during wheel running (Fig. 2A). Similarly to place cells, the theta frequency oscillation of episode cells was higher than that of the field theta rhythm (Fig. 2B), and the slope of the phase precession was inversely related to the length of the episode field (Fig. 2, A and D) (*3*, *19*, *20*, *28*, *30*, *31*).

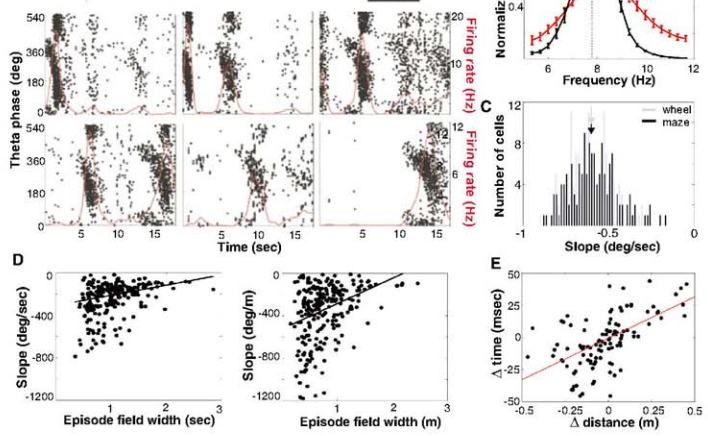



**Fig. 2.** Episode neurons in the wheel display theta phase precession and temporal compression. **(A)** (Top) Unfiltered (light gray) and filtered (4 to 10 Hz) (dark gray) traces of LFP and phase advancement of action potentials (dots). (Bottom) Activity of six example neurons from the same session. Each dot is an action potential, displayed as a function of theta phase and time from the beginning of wheel running from all trials. One and a half theta cycles are shown (*y* axis). Red line, smoothed firing rate. **(B)** Power spectra of spike trains generated during wheel running ($n = 283$ pyramidal neurons) and the simultaneously recorded LFP. Faster oscillation of neurons occurs relative to LFP. **(C)** Slope of theta phase precession within episode fields in the wheel and within place fields in the maze. **(D)** Relationship between phase precession slope and episode length (left, $r_s = 0.46$, $P < 0.0001$) and episode field width (right, $r_s = 0.52$, $P < 0.0001$), respectively. **(E)** Temporal compression of spikes sequences. Correlation of the distance between the peaks of episode fields of neuron pairs in the wheel with the temporal offset of the pair's cross-correlogram peaks is shown. Each dot represents a neuron pair ($n = 105$ eligible pairs; three rats; $r_s = 0.59$; $P < 0.0001$).

Furthermore, the slopes correlated more strongly with the length of the episode field ($r_s = 0.52$, $P < 0.0001$) than with the time it took the rat to run through the same field ($r_s = 0.46$, $P < 0.0001$) (Figs. 2D and 3) because of the variability of the rat's running speed (*28*). The distributions of phase precession slopes for the episode and place fields were also similar (medians were –0.6°/s and –0.6°/s, respectively; $P = 0.6$) (Fig. 2C). Finally, we compared the spike timing relationships among neurons. During maze traversals, the distance between the place-field peaks of a neuronal pair was correlated with the temporal offset between its spikes within the theta cycle, a phenomenon known as distance-time compression (SOM text) (*18*, *19*). Analogously, the distance between peaks of the episode fields of neuron pairs (episode fields with peak firing rate >5 Hz and >3 SD above the mean firing rate were included in this analysis; $n = 105$ pairs) was correlated with the temporal offsets between the spikes at the theta time scale ($r_s = 0.59$, $P < 0.0001$) (Fig. 2E). These findings indicate that the mechanisms generating place and episode fields are similar.

**Body cues are not sufficient to generate assembly sequences.** It has been suggested that in addition to generating a cognitive map of the environment (*2*), the hippocampus and its associated structures integrate self-motion–induced information (*7*, *32*, *33*). Were the episode cell sequences generated by idiothetic self-motion cues? We examined population firing patterns in two control (nonmemory) tasks. In the first task (control 1), the animals (rats 3 and 4) were required to run in the wheel for a water reward available in an adjacent box (*26*). In the second task (control 2), the animals (rats 2 and 3) had continuous access to a wheel adjacent to their home cage, and recordings were made during spontaneous wheel-running episodes. Transient firing patterns, consistent across trials, were rarely observed during the control tasks. Rather, the majority of active neurons exhibited relatively sustained firing throughout the wheel-running

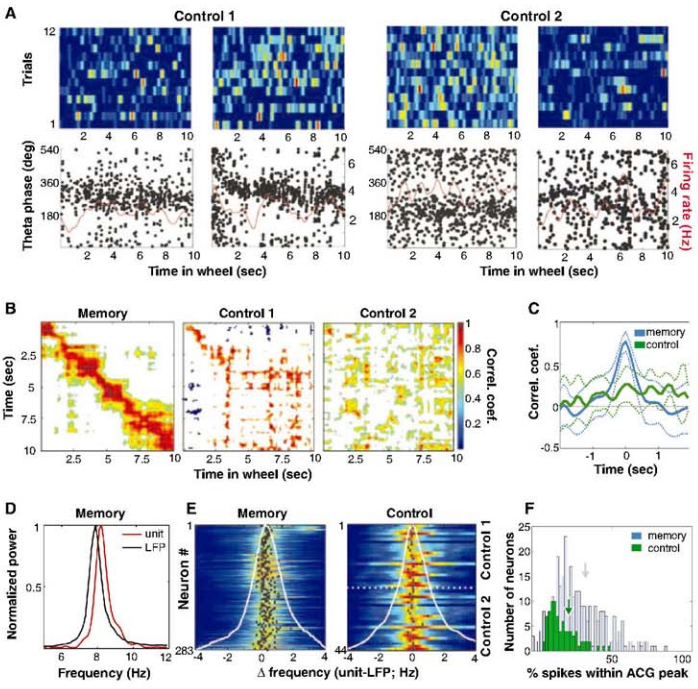



**Fig. 3.** Firing patterns during wheel running depend on the context of the task. **(A)** (Top) Activity of representative single neurons (color-coded) during wheel running in control tasks 1 and 2 (compare with Fig. 1D). (Bottom) Unit discharges (dots) from all trials within a session as a function of theta phase, plotted against time from the beginning of a wheel run. Red line, smoothed mean firing rate. Relatively steady firing rates and a steady theta phase occur in both control tasks. **(B)** Cross-correlation matrices in three different tasks (memory and control 2 are from the same rat). In the memory task, trials with the same future choices [left (L)–trials$_n$ versus L-trials$_{n+1}$ and right (R)–trials$_n$ versus R-trials$_{n+1}$] were cross-correlated, whereas in control tasks trials$_n$ and trials$_{n+1}$ were cross-correlated. Only pixel values significantly different from chance are shown (Spearman rank correlation, $P < 0.01$). **(C)** Population-vector correlation coefficient values in the memory task ($n = 17$ sessions) and control tasks ($n = 8$ sessions) (mean ± SD). **(D)** Power spectrum of spike trains of an episode neuron (unit) and simultaneously recorded LFP during wheel running in the memory task (*30*). The frequency of unit firing oscillation is higher than the frequency of LFP. **(E)** Difference between unit and LFP oscillation frequency in the memory (left) and control (right) tasks. Each line is a color-coded normalized cross-correlogram between power spectra of a pyramidal neuron and simultaneously recorded LFP. A shift of the maximal correlation values to the right indicates that unit theta oscillation is faster than LFP theta oscillation (black dots, maxima of the cross-correlograms; white line, sum of all neurons). There is a significant frequency shift in the memory task (0.44 ± 0.6 Hz) and a lack of frequency shift in control tasks (combined control 1 and 2, 0.07 ± 0.3 Hz). **(F)** Ratio of spikes in the center and tail of temporal autocorrelograms (SOM text). High values indicate compact episode fields; low values indicate spikes scattered throughout the time of wheel running (memory task, $n = 287$ neurons; control tasks, $n = 85$ neurons; rank sum test, $P < 0.0001$). Arrows indicate medians.

periods (Fig. 3A and fig. S4) (*5*, *26*–*27*). During runs of opposite direction in the wheel, different populations of neurons were active (fig. S5) (*26*), arguing for the importance of distant cues (*2*, *20*) and against a critical role of idiothetic inputs (*26*). In addition, the temporal organization of cell assemblies in control tasks was less precise, as reflected by much weaker correlations between temporally adjacent populations during the control tasks than during the memory task (Fig. 3, B and C), despite the similarity in firing rates during all tasks (fig. S6). As another contrast to the memory task, neurons recorded during the control tasks fired throughout the trial, with spikes locked to a similar phase of the theta cycle (Fig. 3A). Consistent with these observations, neurons in the rats performing the memory task oscillated faster than the local field potential (LFP) [difference ($\Delta$) = 0.44 ± 0.6 Hz] (Figs. 2B and 3, D and E), an indication of phase precession (*19*, *20*, *29*, *30*), whereas during the control tasks, the power spectra of the units and LFP were similar ($\Delta$ = 0.07 ± 0.3 Hz) (Fig. 3E). Finally, to quantify differences in temporal clustering of spikes, we examined an autocorrelogram of each neuron. We applied (after filtering, 0.2 to 2 Hz) the same definition for the peak region boundaries that we used for the episode field detection boundary of the episode field (the 10% boundary) and then compared, for each neuron, the ratio of the number of spikes that fell within the peak region boundary to those that fell outside. These ratios were significantly larger during the memory task and reflected the temporal compactness of firing during the memory task as opposed to the control tasks (Fig. 3F). Thus, the indicators of temporally precise sequential activity in neuronal populations were absent during the control tasks, despite indistinguishable motor characteristics across all tasks.

**Assembly sequences depend on memory load.** What is the behavioral function of internally generated cell-assembly sequences? Temporarily inactivating neuronal circuits in the dorsal hippocampus, we found that performance in the delayed alternation task depends on the integrity of the hippocampus (fig. S7) (*17*). Thus, we hypothesized that information about choice behavior is reflected in assembly sequences (*34*). All correctly performed trials were sorted according to the rat's future choice of arm (left or right), and choice-specific firing effects were identified by comparing the firing patterns of single neurons with those of surrogate spike trains created by shuffling the left and right labels (Fig. 4, A and B, and SOM text) (*34*). Some neurons were active exclusively before either the left or right choice, whereas others showed differential firing rates and/or fired at different times after the beginning of wheel running (Fig. 4A, figs. S8 to S10, and movie S1). The largest proportion of neurons exhibiting choice-predictive activity was at the beginning of the run; this proportion decreased as a function of time during the delay and

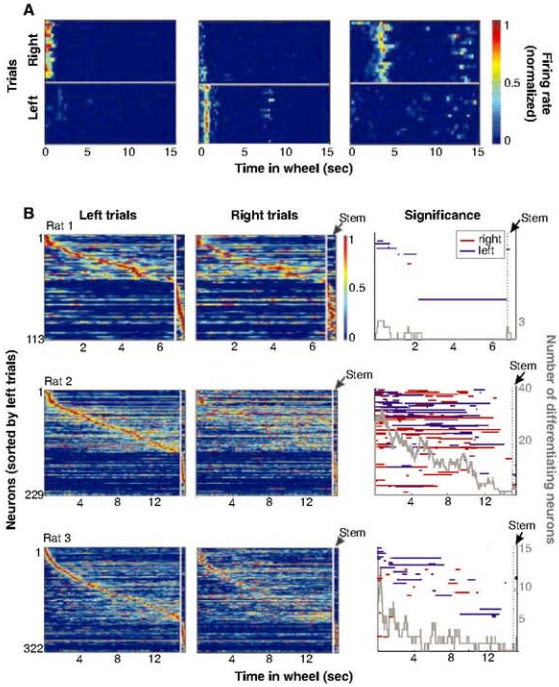



**Fig. 4.** Cell-assembly activity in the wheel predicts the future choice of the rat in the maze. **(A)** Examples of three neurons that strongly differentiated between wheel-running trials preceding right and left choices (fig. S7 and movie S1). **(B)** Normalized firing rate profiles of neurons during wheel running and in the stem of the maze, ordered by the latency of their peak firing rates during left trials (each line is a single cell; cells are combined from all sessions). White line, time gap between the end of wheel running and the initiation of maze stem traversal. (Middle) Normalized firing rates of the same neurons during right trials. (Right) Time periods of significant differences ($P < 0.05$) in firing rates between left and right trials for respective neurons (red line, R > L; blue line, L > R). Gray line, number of neurons discriminating between left and right trials as a function of wheel-running time.

in the stem of the maze (Fig. 4B), suggesting a critical role for initial conditions in specifying the sequences (fig. S11). In addition, we designed a probabilistic model of the relationship between neuronal firing patterns and the animal's choices (SOM text). Using this model, the accuracy of single-trial prediction, under cross-validation, varied from low (near 50%) and not significant to 100% and significant across many sessions (fig. S9).

Because the rat was performing an alternation task, past and future choices were deterministically related on correctly performed trials, and it was not possible to disambiguate their influence on neuronal activity. To distinguish such retrospective and prospective factors (*14*–*17*), we examined cell-assembly sequences during error trials. Neurons that reliably predicted the behavioral choice of the rat on correct trials continued to predict the choice behavior on error trials (Fig. 5A, fig. S12, and movie S1) (*15*, *24*). Similarly, population sequences that differentiated correct behavioral choices continued to predict behavioral choice errors (Fig. 5, B and C, and fig. S13). Although there were only a few error trials, a majority of them could be predicted from the