

DIVERSE DISCIPLINES, ONE COMMUNITY

# Biomedical Computation

Published by the Mobilize Center, an NIH Big Data to Knowledge Center of Excellence

REVIEW

## Taking on the **EXPOSOME**

*Bringing Bioinformatics Tools  
to the Environmental Side  
of the Health Equation*

**ALSO:**

*Learning  
from Patients'  
Health Records*

**And:**

*ZIKA!  
Computational  
Biology to  
the Rescue*

FALL 2016



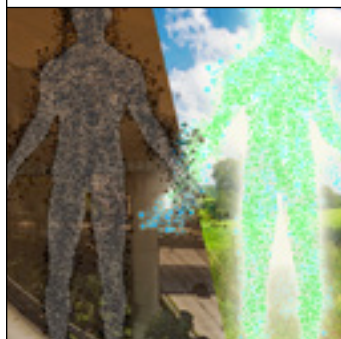


## 8

### Learning from Patients' Health Records:

*Bringing Machine Learning to the Clinic*

BY KATHARINE MILLER

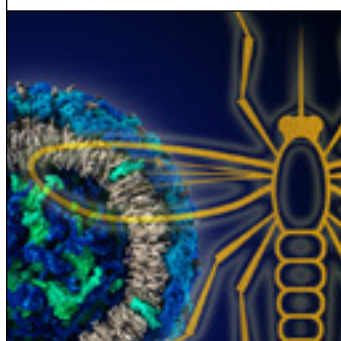


## 14

### Taking on the Exposome:

*Bringing Bioinformatics Tools to the Environmental Side of the Health Equation*

BY KRISTIN SAINANI, PhD



## 22

### Zika!

*Computational Biology to the Rescue*

BY ALEXANDER GELFAND

## DEPARTMENTS

### 1 GUEST EDITORIAL

BIG (DATA) CHANGES:

Improving the Evaluation of Biomedical Academic Software Development Projects

BY AVI MA'AYAN, PhD

### 2 MOBILIZE NEWS

BRINGING LIGHT TO DARK DATA:

Using Snorkel to Label Training Data

BY KATHARINE MILLER

### 3 BIG DATA HIGHLIGHT

THE HARMONIZOME:

A Prototype for Integrated Datasets

BY KATHARINE MILLER

## UNDERCURRENTS

4

### TACKLING TUMORS: TURNING IMMUNE CELLS INTO CANCER KILLERS

BY ESTHER LANDHUIS

### 30 SEEING SCIENCE

ANIMATING HYPOTHESES

BY KATHARINE MILLER

**Cover and Exposome Art:** Pages 14-15: City skyline and freeways © Oneinchpunch; country road © Pietus; human silhouette © Sebastian Kaulitzki; food images © Skypixel; cleaning spray bottle © Jakub Jirsák; icons © Wong Yien Keat, Vectomart, and Leremy. All of Dreamstime.com.

**Health Records Art:** Page 8: © Dicraftsman; p. 9: © PureSolution; p. 11: © Bakhtiar Zein; p. 12: created by Rachel Jones of Wink Design Studio using other illustrations; p. 13: upper left image © Idey, key by Rachel Jones, magnifying glass © PureSolution. Unless noted, illustrations are Dreamstime.com

**Zika Art:** Page 22: Created by Rachel Jones of Wink Design Studio using zika model (credit in caption on page) and mosquito icon © Ivan Kotliar on Dreamstime.com.

#### Fall 2016

Volume 12, Issue 2  
ISSN 1557-3192

#### Co-Executive Editors

Scott Delp, PhD  
Russ Altman, MD, PhD

**Associate Editor** Joy Ku, PhD

**Managing Editor** Katharine Miller

#### Science Writers

Alexander Gelfand, Esther Landhuis,  
Katharine Miller, Kristin Sainani, PhD

#### Community Contributors

Avi Ma'ayan, PhD

#### Layout and Design

Wink Design Studio

#### Printing

AMP Printing

#### Editorial Advisory Board

Ivet Bahar, PhD,  
Jeremy Berg, PhD,  
Gregory F. Cooper, MD, PhD,  
Mark W. Craven, PhD,  
Jiawei Han, PhD,  
Isaac S. Kohane, MD, PhD,  
Santosh Kumar, PhD,  
Merry Lindsey, PhD,  
Avi Ma'ayan, PhD,  
Mark A. Musen, MD, PhD,  
Saurabh Sinha, PhD,  
Jun Song, PhD,  
Andrew Su, PhD,  
Paul M. Thompson, PhD,  
Arthur W. Toga, PhD,  
Karol Watson, MD

For general inquiries, subscriptions,  
or letters to the editor, visit our  
website at [www.bcr.org](http://www.bcr.org)

#### Office

Biomedical Computation Review  
Stanford University  
318 Campus Drive  
Clark Center Room W352  
Stanford, CA 94305-5444

Publication is supported by NIH  
Big Data to Knowledge (BD2K)  
Research Grant U54EB020405.

Information on the BD2K program can be  
found at <http://datascience.nih.gov/bd2k>.

#### The NIH program and science officers for the Mobilize Center are:

Grace Peng, National Institute of  
Biomedical Imaging and Bioengineering,  
Theresa Cruz, National Institute of  
Child Health and Human Development,  
Daofen Chen, National Institute of  
Neurological Disorders and Stroke

#### Biomedical Computation Review is published by:

The Mobilize Center,  
an NIH Big Data to Knowledge (BD2K)  
Center of Excellence  
[mobilize.stanford.edu](http://mobilize.stanford.edu)



# BIG (DATA) CHANGES

## *Improving the Evaluation of Biomedical Academic Software Development Projects*



**W**ith the increasing diversity of assays that produce massive quantities of data, experimental labs are becoming more and more dependent on software tools and online databases, as well as collaborations with bioinformaticians. There is a clear need to attract, train, and support more biomedical data analysts as well as fund more computational “dry-lab” projects.

But throwing more money at software development projects is not enough. The National Institutes of Health (NIH) needs to change how it evaluates such projects. NIH study sections have been optimized for decades to fairly evaluate experimental wet-lab projects. Applying the same approach to computational projects doesn't yield what is really needed: having the most useful software tools and databases continually maintained and enhanced for the long term.

The current NIH review process lacks standards for fair and objective ways to assess the usage, performance, and impact of software tools and databases. Commonly used metrics such as download volume, unique users, and query submissions are currently not required or verifiable. One possible solution is to develop a system similar to Google Analytics where software developers would insert an NIH-certified JavaScript code into their tool and database hosting websites. This code would collect and send user information to a centralized public repository. An alternative would be to develop a global authentication mechanism where users would need to sign in before using NIH supported tools and databases—though this solution may deter users who wish to stay anonymous.

The popularity of tools and databases is not always the best measure of their quality. NIH should seek out more objective benchmarks to assess the quality of algorithms, tools, and databases so the best—those that maximally extract knowledge from the raw data—are selected and recommended.

NIH also needs to find ways to better incentivize the maintenance and enhancement of software tools and databases past their funding term. Currently, useful and popular tools and databases may abruptly disappear upon cessation of funding. Such sudden disappearance can leave wet-bench investigators hanging, without the ability to continue their projects or reproduce their results. Hence, there is a need to develop resources for hosting web-based software applications and databases so that they can remain online and available even after the conclusion of an NIH-supported project.

This can be solved by requiring NIH grant-supported biomedical software developers to provide their tools and databases in self-contained executable environments, such as Docker containers, so that they can be redeployed and hosted in the cloud. In this way, the NIH could cover the low monthly bill of keeping these software services active and available for many years after the funded project has expired. The source code for such projects could also be mandated to be open and placed in codebase repositories for the community to potentially continue to enhance it. Metadata and versioning of tools should also be required for better indexing and provenance.

A fair review of software projects would also consider yet other differences between wet- and dry-lab projects. The life cycle of software projects is shorter than that of typical experimental projects. In addition, to complete software projects, academics often need to hire professional software developers who are presently in high demand and require higher salaries than most NIH-funded researchers can afford. It is also difficult to retain these employees because they are often attracted to work in industry. Big Data science is gradually engulfing biomedical research where computational analysis is becoming the central pillar. Rapid adaptation to these changes is essential, including better management of academic software research development projects. □

---

**The current NIH review process lacks standards for fair and objective ways to assess the usage, performance, and impact of software tools and databases.**

---

# BRINGING LIGHT TO DARK DATA:

## *Using SNORKEL to Label Training Data*

Unstructured data—sometimes called dark data—abounds in many domains, including biomedicine. It includes text, such as published scientific literature or physician notes, as well as tables, figures and images. Using computers and advanced machine-learning approaches, researchers are becoming adept at extracting valuable knowledge from dark data. But machine-learning algorithms often require large sets of labeled training data, which in many areas requires the efforts of domain experts. This is a serious bottleneck: “Making training data is so expensive that there are a lot of domains that could never afford to do it,” says **Jason Fries, PhD**, a postdoctoral research fellow who is part of the Mobilize Center at Stanford University. Moreover, researchers who want to use machine learning find that creating labeled training data takes up a significant portion of their time.

To address that problem, Fries, **Alex Ratner, Steven Bach, PhD**, and others in **Chris Re’s** lab at Stanford are developing an application framework called Snorkel that can automatically generate labeled training data using sources of “weak supervision”—i.e., sources of rules that were not directly intended for the labeling purpose and for which there’s no expectation that the labels will be perfect. For many tasks, Snorkel’s results are surprisingly good. In recent work extracting mentions of diseases and chemical names from PubMed abstracts, for example, “Snorkel can train a model that performs as well as one trained on human-labeled data,” Fries says.

Snorkel starts with a bunch of noisy

rules—heuristics—for finding mentions of some concept, such as a disease. In biomedicine, these are often derived from ontologies, but they can come from other sources as well. Snorkel then automatically learns the accuracy of these heuristics as they generate labeled training data, and then uses that accuracy information to de-noise those labels. Under the hood, Snorkel is training what’s called a generative model, and can be intuitively understood as having parallels to crowdsourcing algorithms, where the goal is to figure out which people do a better job than others at a particular task, and to take that accuracy into account in de-noising the data. Similarly, in Snorkel, if rules tend to agree with each other and cover a lot of data, Snorkel will trust them more than it will contrarian rules. But Snorkel has a significant advantage over crowdsourcing: “Instead of one person or a few people labeling a small subset, you have labeling functions that can scale to millions of samples,” Fries says.

In the labeling of diseases, Snorkel actually has another advantage: It captures some of the inherent disagreement that exists around disease labels. In small human-labeled datasets, gold standard disease definitions are often imperfectly negotiated by a small group of people, Fries says. Snorkel provides a natural mechanism for learning in the presence of disagreement without resorting to manual adjudication.

In addition to extracting mentions of diseases and chemicals from PubMed abstracts, Snorkel can successfully extract relationships from the scientific literature, such as causal relations between genetic mutations and phenotypes.

For biomedicine, where ontologies are prevalent, Snorkel should prove particularly valuable, Fries notes. In their initial experiments, there is only a small gap between the quality of labels generated using Snorkel with weak supervision by ontologies and the quality of ordinary labeled data—and in some areas there is no gap. “That’s a nice finding of the work,” Fries says. □



### DETAILS

For more information about Snorkel or to download this open-source application framework, visit <http://hazyresearch.github.io/snorkel/>

*The Mobilize Center for Mobility Data Integration and Insight is an NIH Big Data to Knowledge (BD2K) Center of Excellence at Stanford University.*

# THE HARMONIZOME:

## *A Prototype for Integrated Datasets*

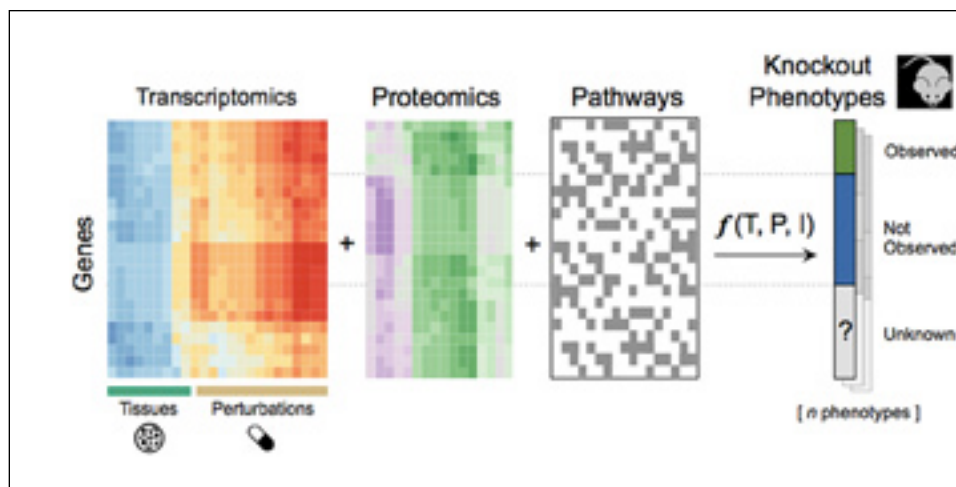
For years now, biocomputational scientists have been talking about the need for better data integration. “People talk a lot about how data are in silos and not connected,” says **Avi Ma’ayan, PhD**, professor of pharmacological sciences at the Icahn School of Medicine at Mount Sinai and principal investigator of the BD2K-LINCS Data Coordination and Integration Center. After all that talk, Ma’ayan and his colleagues figured it was time to take action. So they created the Harmonizome, a collection of all the hottest and most exciting databases that everyone is using. “It allows you to find knowledge about genes and proteins that was buried in those silos but now is accessible.”

To create the Harmonizome, Ma’ayan’s team gathered together the major omics databases as well as databases on mouse and human phenotypes and processed them into a relatively simple format. That processing involved taking either raw data or formatted data from existing databases and mapping it onto common IDs for genes. They also processed the data into simplified formats such as relational tables, making it ready for machine learning. “It makes it very easy for someone to do predictions of functions for genes,” Ma’ayan says.

That’s what makes Ma’ayan most excited—the potential for using the Harmonizome to impute knowledge across data resources. His favorite example thus far, which was included among other examples in a paper about the Harmonizome published in *Database* in 2016, involves the prediction of mouse phenotypes. Using the Harmonizome, his team was able to create tables that describe functions and attributes of various genes and then use those to predict mouse phenotypes associated with specific knockouts. For example, from mouse knockout experiments, the researchers first flagged gene knockouts that increase

mouse lifespan. Using the Harmonizome, Ma’ayan and his colleagues predicted the probability of genes, not yet knocked out in mice, for likelihood of increasing lifespan. “You can do this—predict other genes that should be relevant to aging—using machine learning,” he says. “And those could be future drug targets for potentially increasing our lifespan or improving our healthspan.”

Ma’ayan thinks of the Harmonizome as a proto-



*By applying machine learning to data from diverse resources integrated to create the Harmonizome, Ma’ayan’s team was able to predict likely knockout mouse phenotypes that have not yet been observed. Image Courtesy of Avi Ma’ayan.*

type that is leading the way by showing what can be done. Some other data integration efforts allow search at the metadata level only. “The nice thing about the Harmonizome is that it enables search at the data level,” he says. But, he acknowledges, making it scalable could be challenging.

Still, the Harmonizome has proven popular. During its first year, the site had 60,000 unique users visit and 250,000 page views. “We get about 400 users per day now,” Ma’ayan says, with about 40 percent sticking around for a while because they are finding it useful. He’d like to learn more about how others are using the resource. “I’m sure people can think of creative ways to use it that we haven’t thought of,” Ma’ayan says. “That will be the coolest thing.” □



# TACKLING TUMORS

## *Turning Immune Cells into Cancer Killers*

**T**umors often contain a hodgepodge of cells. Some cells have genetic glitches, others don't; some obey normal growth rules, others divide out of

chasing answers, and many think the insights gained could revolutionize how we fight cancer.

The revolution has already begun.

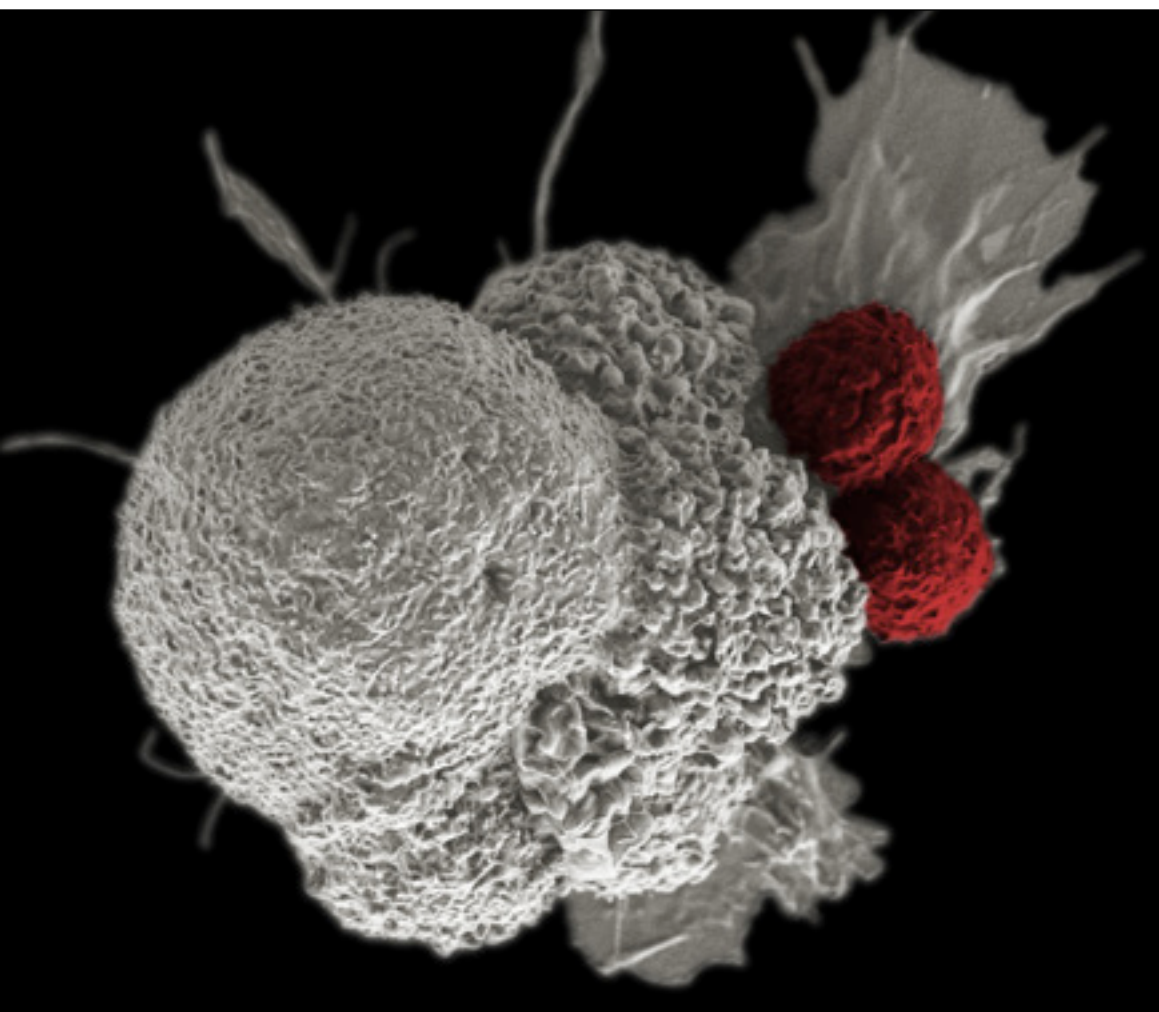
complex molecular profiling data can show how to activate key immune cells to fight cancer, why immunological cancer treatments work for some patients and not others, and which additional molecules could serve as potential targets for customized therapies to train people's immune cells to fight their own cancer.

### Reprogramming Tumor Macrophages

Several types of immune cells, including macrophages and T-cells, turn traitorous in the face of cancer. Macrophages are scavengers. Typically they roam the body and chew up unwanted debris and dying cells, even cancerous ones. But macrophages also help with wound healing, and tumors can corrupt the scavenging cells to adopt this role inappropriately. With the switcheroo, macrophages now "see the tumor like a wound," says **Michele De Palma, PhD**, of École Polytechnique Fédérale de Lausanne (EPFL) School of Life Sciences in Lausanne, Switzerland. "They go there and help the tissue grow."

But in cancerous tissue, "healing" responses are harmful. Could macrophages be reprogrammed to act more like tumor killers?

A clue came several years ago when De Palma and coworkers discovered that macrophages at tumor sites accumulate miR-511-3p, a specific microRNA molecule (miRNA), and turn down a number of genes that are typically active in macrophages. Unlike typical RNA, miRNAs themselves don't get made into



*In this pseudo-colored scanning electron micrograph, cytotoxic T cells (red) attack an oral squamous cell cancer (white) as part of a natural immune response. Source: National Cancer Institute \ Duncan Comprehensive Cancer Center at Baylor College of Medicine. Creator: Rita Elena Serda.*

control. Immune cells enter the mix as well, initially swooping in to reject the tumor as they would other foreign substances. Later, however, these same cellular sentinels may inexplicably let down their guard, allowing the cancer to gain a foothold.

How do tumors outwit the body's defense system? Scientists are

Cancer treatments that harness the immune system are now a reality, and more are on the way. But with its many players and varied activities, the immune system's response to tumors, which are themselves evolving, often stymies understanding. In this setting, computational biologists are playing an important role. Plumbing

protein; instead, they dial back the activity of other genes by binding and preventing translation of their messenger RNAs (mRNAs). Discovered in the early 1990s, miRNAs play key roles in a range of developmental processes and in some human diseases—including cancer, where studies have linked changes in a cell's miRNA expression to its road toward malignancy. De Palma's finding suggests miR-511-3p regulates genes to render tumor-associated macrophages ineffective at fighting cancer.

The researchers wondered if shutting down miRNAs could shift macrophage behavior so they would fight tumors instead of ignore them. In a more recent study, they designed experiments to find out.

These studies used mice engineered for two traits: the tendency to grow cancerous tumors and an absence of DICER, an enzyme that is necessary for miRNAs to mature. Essentially, most miRNA activity is blocked in the macrophages of these mice. The result: a radical change in the macrophages' gene expression profiles and behavior. DICER-deficient tumor macrophages became "very nasty to tumors," DePalma says. They behaved like macrophages fighting a bacterial infection. And they didn't just battle the tumor alone—they recruited cytotoxic T cells to attack and eliminate the tumor, De Palma's team reported in June 2016 in *Nature Cell Biology*.

The team didn't stop there. Knowing macrophages have several hundred miRNAs, and having established that shutting them off turns bystander macrophages into tumor killers, De Palma wanted to find out which specific miRNAs were responsible for this turnaround—and that's where bioinformatics came in.

They used two approaches, one with mouse data and another with human data. In the first, the team isolated tumor-associated mouse macrophages with and without DICER and compared their

transcriptomes to identify differentially expressed genes. This allowed them to determine which mouse mRNAs are targeted by specific miRNAs. The second strategy, which was developed in collaboration with co-author **Chia-Huey Ooi, PhD**, a bioinformatician at Roche in Basel, Switzerland, involved analyzing 171 publicly available blood samples from patients with acute myeloid leukemia (AML). For each sample, the researchers determined mRNA/miRNA signatures—whenever gene A gets expressed, miRNA B goes up—and used those signatures to predict which miRNAs potentially regulate the gene signature of macrophages in the DICER-deficient mouse tumor models.

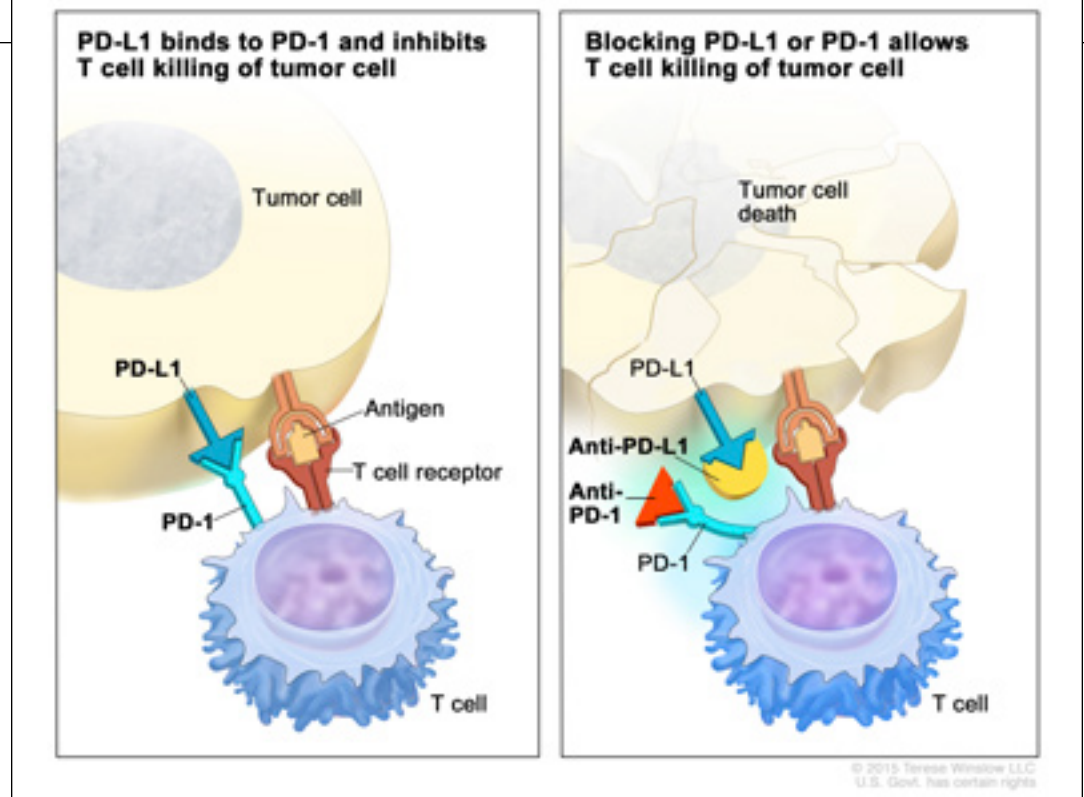
Both approaches identified Let-7 as one of the miRNAs responsible for reprogramming macrophages into tumor tolerators. Wet-lab experiments confirmed the finding: When the researchers restored Let-7 miRNA activity in DICER-deficient tumor macrophages—in which the absence of DICER causes a broad miRNA shutdown—the macrophages reverted to ignoring the

tumor. Spurred by the new findings, the researchers are now working on using nanoparticles to block DICER or Let-7 activity in tumor-associated macrophages.

## Finding Neoantigens

Whereas De Palma's experiments showed macrophages could be reinvigorated to lure killer T cells to tumor sites, existing cancer immunotherapy drugs spur T cells into action by targeting immune checkpoint molecules found on their surface. These therapies have brought lasting relief to former U.S. president Jimmy Carter and other patients with previously incurable cancers.

Key to developing these drugs was the discovery that a T-cell checkpoint molecule called PD-1 recognizes PD-L1 proteins on the surface of some cancer cells and tumor-infiltrating immune cells, particularly macrophages. Interaction between these molecules creates a "stealth shield" for the tumor, preventing the immune system from seeing it, explains **Richard Chen, MS, MD**, chief scientific officer at Personalis, a Silicon Valley genomics company. Checkpoint



*In cancerous cells, checkpoint proteins, such as PD-L1 on tumor cells and PD-1 on T cells, help keep immune responses in check. The binding of PD-L1 to PD-1 prevents T cells from "seeing" the tumor cell, allowing the cancer to grow unchecked (left panel). Blocking the PD-1/PD-L1 interaction with an immune checkpoint inhibitor (anti-PD-L1 or anti-PD-1) unshields the tumor cell so T cells can kill it (right panel). Source: National Cancer Institute. Printed with permission © 2015 Terese Winslow LLC, U.S. Govt. has certain rights.*

inhibitors prevent those molecular interactions and break the stealth shield, making the tumor visible to immune cells.

But there's a vexing problem: Checkpoint blockade doesn't work for many eligible cancer patients. Even when a drug succeeds in "unshielding" a tumor, there's no guarantee a cell's particular collection of tumor peptides will actually trigger an immune response. "Each tumor is unique and complex," Chen says, and determining a tumor's immunogenicity poses a difficult problem.

To gauge which patients will respond to immunotherapy, Chen's team is using bioinformatics and machine-learning approaches to probe their tumor's genetic mutations and determine how the tumor is evading the immune system.

Tumor-specific peptides are a key determinant of immunogenicity. As tumor cells accumulate mutations, some give rise to tumor-specific mutant peptides, or neoantigens, that the immune system considers foreign. Studies have shown that tumors with more neoantigens stimulate stronger immune responses.

Personalis has developed a platform called ACE ImmunoID™, which combines DNA and RNA profiling to gauge a tumor's mutations and neoantigen load. While existing assays can screen for certain proteins known to be important for immunogenicity, such as PD-1 and PD-L1, Personalis' method achieves higher specificity and sensitivity. The patented technology optimizes chemistry and probes to fill in common sequencing gaps in DNA and RNA. In addition, it uses computational algorithms to predict, from the sequencing information, which mutations could result in neoantigens that are likely to bind major histocompatibility (MHC) proteins—the set of cell surface proteins that help the immune system recognize foreign molecules. Interaction with MHC is a key prerequisite for a neoantigen to be immunogenic, Chen says, so people with MHC-binding neoantigens are more likely to benefit from checkpoint blockade therapies.

The prediction algorithm uses a

machine-learning approach that includes neural networks trained using experimental data on thousands of peptides that do or don't bind well to MHC molecules. With enough input data, the machine-learning algorithm can learn to predict whether a new peptide is likely to bind to MHC. Strong neoantigens identified by this method could also help researchers design tumor-specific vaccines. "You'd have a completely personalized therapeutic," Chen says.

Similar tools are under development elsewhere. Researchers at the La Jolla Institute for Allergy and Immunology in California have created the Immune Epitope Database Analysis Resource ([www.iedg.org](http://www.iedg.org)), a collection of tools for predicting and analyzing immune epitopes in the context of T- and B-cell responses. Another group led by scientists at the Technical University of Denmark has developed NetMHC (<http://www.cbs.dtu.dk/services/NetMHC/>), which uses artificial neural networks to estimate the affinity of user-submitted peptide sequences to specific MHC alleles.

### Probing Transcriptomes

Other researchers are taking a different approach to understanding why checkpoint blockades work in some patients and not others. Perhaps, they say, it has to do with the specific nature of a tumor's heterogeneity, including differences in the T cells that are present. Gene expression profiles of tumors as a whole can't address that. "If you take tumor tissue and grind it up for sequencing, all you detect is a mixture of signals," says **Benjamin Izar, MD**, an oncology fellow working with **Levi Garraway, MD, PhD**, at the Dana-Farber Cancer Institute, Boston, and at the Broad Institute of MIT and Harvard. "It's hard to say where the signal came from and what it means in the context of the tumor."

That's why Izar and Garraway teamed up with colleagues **Aviv Regev** and **Itay Tirosh** to perform single-cell RNA sequencing not just on cancerous cells but non-malignant types including immune

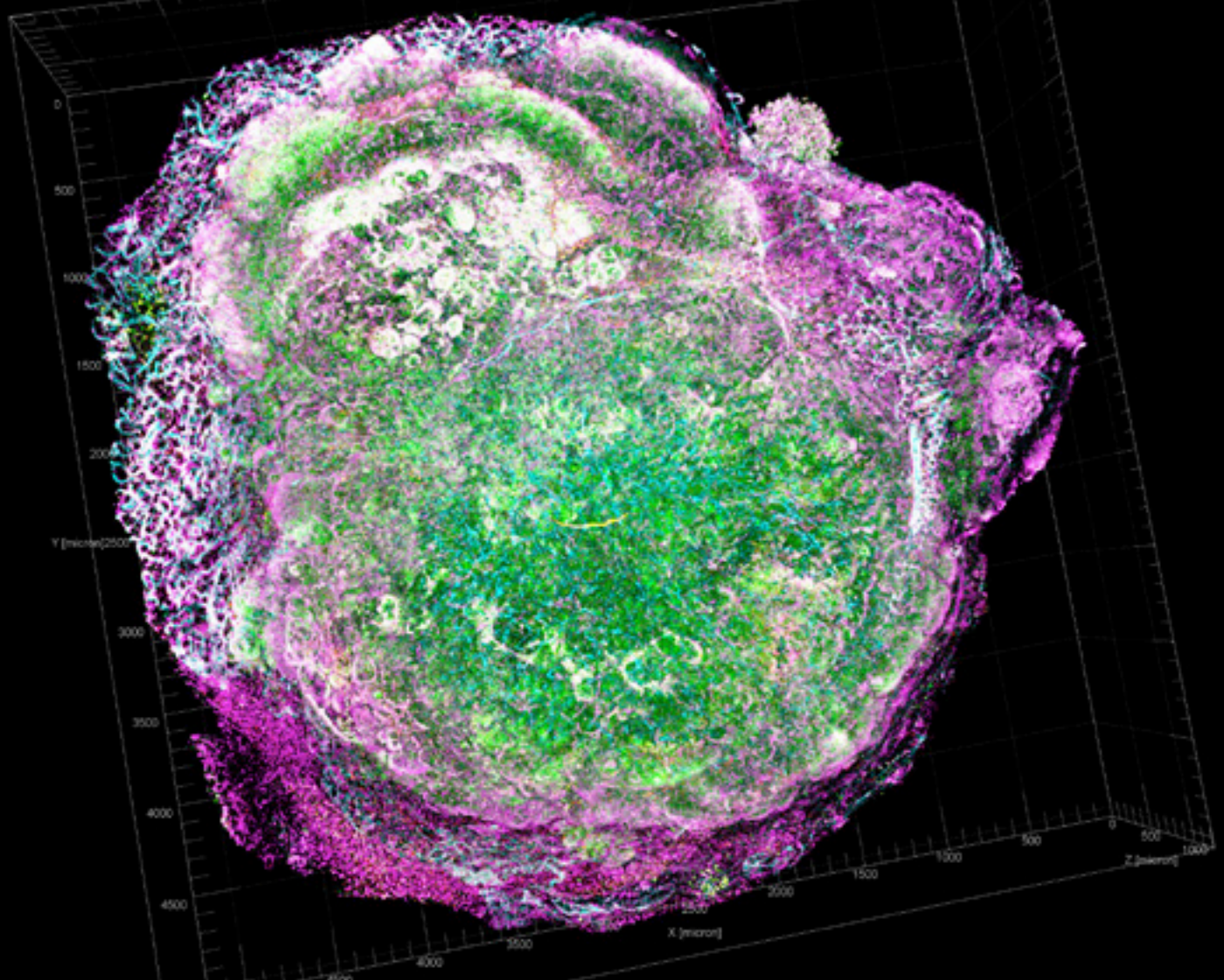
cells and connective tissue from patient tumors. "We wanted a broad, unbiased reflection of what is actually in the tumor," Izar says of their study published April 2016 in *Science*. In total the team analyzed 4,645 cells in tumors collected from 19 people with melanoma skin cancer. Some patients had never been treated for their cancer. Others had taken a drug designed to target melanoma cells with a specific mutation. Still others had received immune checkpoint inhibitors. The sequencing process yielded thousands of transcriptomes. The data were complex and noisy, says Izar. So they used various statistical and machine-learning methods to visualize and interpret the data.

Several interesting features jumped out when the researchers analyzed the transcriptomes of T cells, whose presence and function at tumor sites has been shown to predict responses to immune checkpoint therapies. Many of the tumor T cells expressed markers of "exhaustion," says Izar. Unlike normal cytotoxic T cells that help control cancer by recognizing key molecules on the surface of tumor cells, some T cells lose their fighting power, and this exhaustion is marked by transcriptional changes in specific genes.

Still, it can be hard to distinguish activated T cells from exhausted ones. Analyzing gene expression at the single-cell level could help identify better markers for truly exhausted T cells because those might be the patients who will respond to immunotherapies, says Izar.

In a paper published in August 2016 in *Genome Biology*, another Boston team, led by **X. Shirley Liu, PhD**, at the Dana-Farber Cancer Institute, also used RNA-sequencing data to evaluate the clinical impact of immune cells in various types of cancer. However, instead of directly measuring transcriptomes, the team analyzed published data from over 10,000 RNA-sequencing samples across 23 cancer types from The Cancer Genome Atlas. They developed a computational algorithm called TIMER (Tumor IMMune Estimation Resource) that estimates the





*The heterogeneity of the tumor microenvironment plays a crucial role in allowing cancer to grow and evade destruction. This image of a mouse model for HER2-positive breast cancer uses a novel imaging technique called transparent tumor tomography that three-dimensionally illuminates the tumor microenvironment at a single-cell resolution. HER2 (green), Ki-67 (red), PD-L1 (purple), immune cells (yellow), and endothelial cells (cyan). Source: National Cancer Institute \ Univ. of Chicago Comprehensive Cancer Center. Creator: Steve Seung-Young Lee.*

tumors' immune-cell composition and correlates the immune cells' presence and gene expression with clinical outcomes. The analysis found, unexpectedly, that the abundance of CD8 cytotoxic T cells does not always correlate with expression of CTLA4, an immune checkpoint protein sometimes targeted in checkpoint blockade immunotherapies. The researchers think this might explain why some patients don't respond to CTLA4-blocking treatments despite expressing high levels of CTLA4.

Though T cells and macrophages have been a big focus, other immune cells can also determine how well a patient responds to cancer immunotherapy

drugs. Using a computational approach called CIBERSORT, researchers led by Stanford oncologist **Ash Alizadeh** characterized the cell composition of around 18,000 human tumors by surveying their gene expression profiles. Their analysis, reported in a 2015 *Nature Medicine* paper, found complex relationships between 22 immune subset signatures and overall survival across 25 cancer histologies. For example, they found that people whose tumors contained high numbers of plasma cells (a type of immune cell) had a better prognosis, while those with a high concentration of neutrophils (another type of immune cell) tended to have a worse

outlook. The findings could be used to find new targets for cancer therapies—or to help predict patients' chances of responding to some existing treatments.

Using these and other diverse approaches, scientists hope to refine and identify additional molecular signatures in patient tumors to help predict responses to immunotherapies. But it won't be easy. "It's different from a cholesterol test where you measure one entity and you're done," Chen says. "When you're talking about genomics-based diagnostics there's significant complexity in the informatics and sequencing. There are multiple dimensions to the problem." □

Physicians are forever recording information about their patients. They take vital signs, order lab tests and imaging, prescribe medications, check boxes to define patients' diagnoses for billing purposes, and write or dictate narrative descriptions of each patient's status. For the most part, all of

this information goes into the patient's electronic health record (EHR) where it remains untouched for any purpose other than billing or the patient's next visit.

These EHRs represent a vast untapped gold mine for improving patient care. "There is no other industry that doesn't learn from its prior customers," says

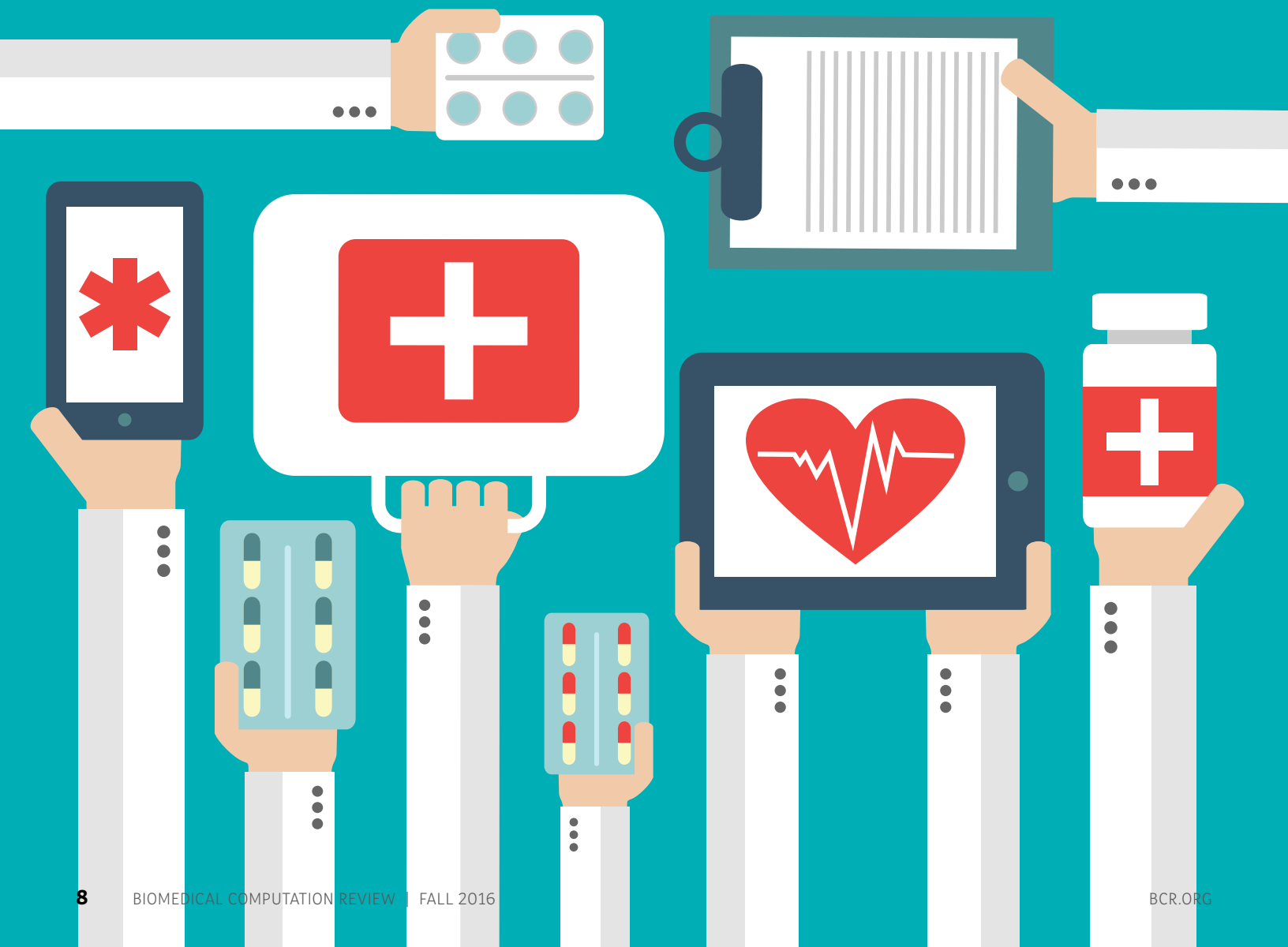
**Nigam Shah, MBBS, PhD**, associate professor of medicine at Stanford University.

In clinical settings, EHRs can be mined to identify patients at high, medium, and low risk for various outcomes, allowing healthcare providers to intervene proactively. For example: Who is likely to be admitted to the ICU or ER?

# LEARNING FROM Patients' Health Records:

## *Bringing Machine Learning to the Clinic*

BY KATHARINE MILLER



EHRs can also be used to personalize risk assessment. For example, someday, a clinician might be able to query a warehouse of EHR data to find how other patients highly similar to one of theirs fared when given various treatments.

EHRs can also be used to predict differences in how diseases progress. For example: When will pre-diabetes progress to full-onset diabetes? Or when will an aplastic mole progress to full-blown melanoma?

Applying machine learning to EHRs for the benefit of patients has its challenges. Medical record systems vary among institutions, are not standardized, and are constantly evolving; diagnostic codes used for billing purposes are often unreliable; and narrative descriptions in natural language are hard for computers to interpret. Moreover, privacy concerns limit access to EHRs; datasets from some institutions may be too small to be useful, especially for rare diseases; and when datasets are larger, the statistical challenges exceed an individual clinician's grasp.

There are also methodological hurdles to cross. "There are probably a dozen widely used machine-learning algorithms and thousands of variations," says **David Page, PhD**, professor of biostatistics and medical informatics at the University of Wisconsin's School of Medicine and Public Health. "We try to be very open-minded about what method would work best."

And then there are the economics of it. Institutions like Stanford University, Shah's employer, may be willing to foot the bill for a data warehouse full of EHRs without concern for the financial return, but the larger healthcare industry would have to pay for EHR work using patient-care dollars—and would need to show benefit to specific patients to collect those funds. "We haven't figured that out yet," Shah says. "How do we demonstrate a return on investment when the people who stand to benefit have no skin in the game?"

Despite the challenges, researchers can point to a number of promising projects that are either already benefiting patients or soon will be. "It's phenomenal to see the work get to this point," says

**Jenna Wiens, PhD**, assistant professor of computer science and engineering at the University of Michigan. "We talk all the time about leveraging EHR data to produce actionable knowledge, but in practice it can be really hard to do. I'm really excited to see where it leads."

## Improving the EHR to Improve Care

For EHRs, like other databases, garbage in will produce garbage out: If doctors and nurses aren't entering data accurately, or aren't keeping the records up-to-date, patient care could suffer. Moreover, narrative notes in EHRs often

bleeding." It is often the first thing the ER physician sees. This important information could be valuable to record in a structured form, but a drop-down menu of chief complaints would be very long and require too much time from nurses in a hurry.

So, using data for 200,000 patients who had been to the ER in the past, Sontag and Horng, along with Sontag's PhD students **Yacine Jernite** and **Yoni Halpern**, trained a machine-learning algorithm to identify what the chief complaint should be for new patients. Implementing the algorithm in an ER setting required that nurses write a 20- to 40-word triage assessment of the patient, in addition to taking vital signs. The machine-learning



hide information that could be useful if it were more structured. So some researchers are using machine learning to improve the accuracy and structure of the EHR—which in turn makes the EHR more valuable for machine learning. It's a great way to tackle some low-hanging fruit, says **David Sontag, PhD**, assistant professor of computer science and data science at New York University.

About seven years ago, Sontag, a specialist in machine learning, began working with **Steven Horng, MD**, associate director in the division of emergency informatics at Beth Israel Deaconess Medical Center in Boston, Massachusetts. They wondered if machine learning could be used to structure the patient's chief complaint as it is entered in the EHR by emergency room (ER) triage nurses. The chief complaint is typically a brief, free-text summary of the patient's condition. For example, it might be "chest pain," "hit by car," "pneumonia," or "uncontrolled

algorithm then uses that information to predict and auto-complete a structured entry for the chief complaint. The algorithm relies on a clearly defined ontology of many hundreds of possible chief complaints. The system, which has been running live for about three years, is much loved by the nursing staff at Beth Israel Deaconess Medical Center. They complain immediately whenever the system goes down, Sontag says. And the quality of the chief complaints has improved, judging from how rarely the nurses and physicians override the algorithm's chief complaint suggestions, he says. Moreover, with the chief complaint recorded as structured data, it becomes possible to apply more advanced machine-learning approaches to the data—ones that might seek to classify ER patients at highest risk of death, for example.

The approach can be used to improve the structure of EHRs in other contexts as well. For example, Sontag's group used



machine learning to predict what should be added to or removed from the EHR's patient problem list. This list of a patient's current health issues provides valuable contextual information when a patient presents with a new problem, but it is hard to maintain and keep up to date.

"These are simple examples," Sontag says, "And they demonstrate that even the simplest of machine-learning methodologies can have a significant impact on healthcare."

Taking these efforts further, Sontag has a vision to create a foundation for the next generation of EHRs. To be able to deduce a patient's past and present as well as predict the future requires structured information that doesn't exist in current EHRs. So Sontag wants to use machine learning to automatically convert unstructured data into structured data. It's not an easy task. Machine-learning algorithms typically require training data that has been labeled by experts. That's hard to come by in healthcare settings, Sontag says, and it often doesn't transfer well from one institution to another. So Sontag came up with a solution he calls the "anchor and learn framework." It uses prior medical knowledge known to an expert to identify an anchor in the EHR, (e.g., the fact that seeing metformin and multiple HbA1c measurements means someone is a diabetic) and then uses that anchor as a basis for learning. Experts are needed only for determining the best anchors—not for labeling all of the data.

"It's not doing diagnosis," Sontag says. "We're not finding something someone

doesn't already know. We're just getting a piece of knowledge that's important into a structured form." For example, if a patient who is being prescribed antibiotics is from a nursing home—a context where antibiotic resistant bacteria often develop—the EHR could flag that and then offer a popup asking "are you sure the patient doesn't have antibiotic resistant bacteria?" But the EHR can only do that if being "from a nursing home" is known. And Sontag's system can figure that out.

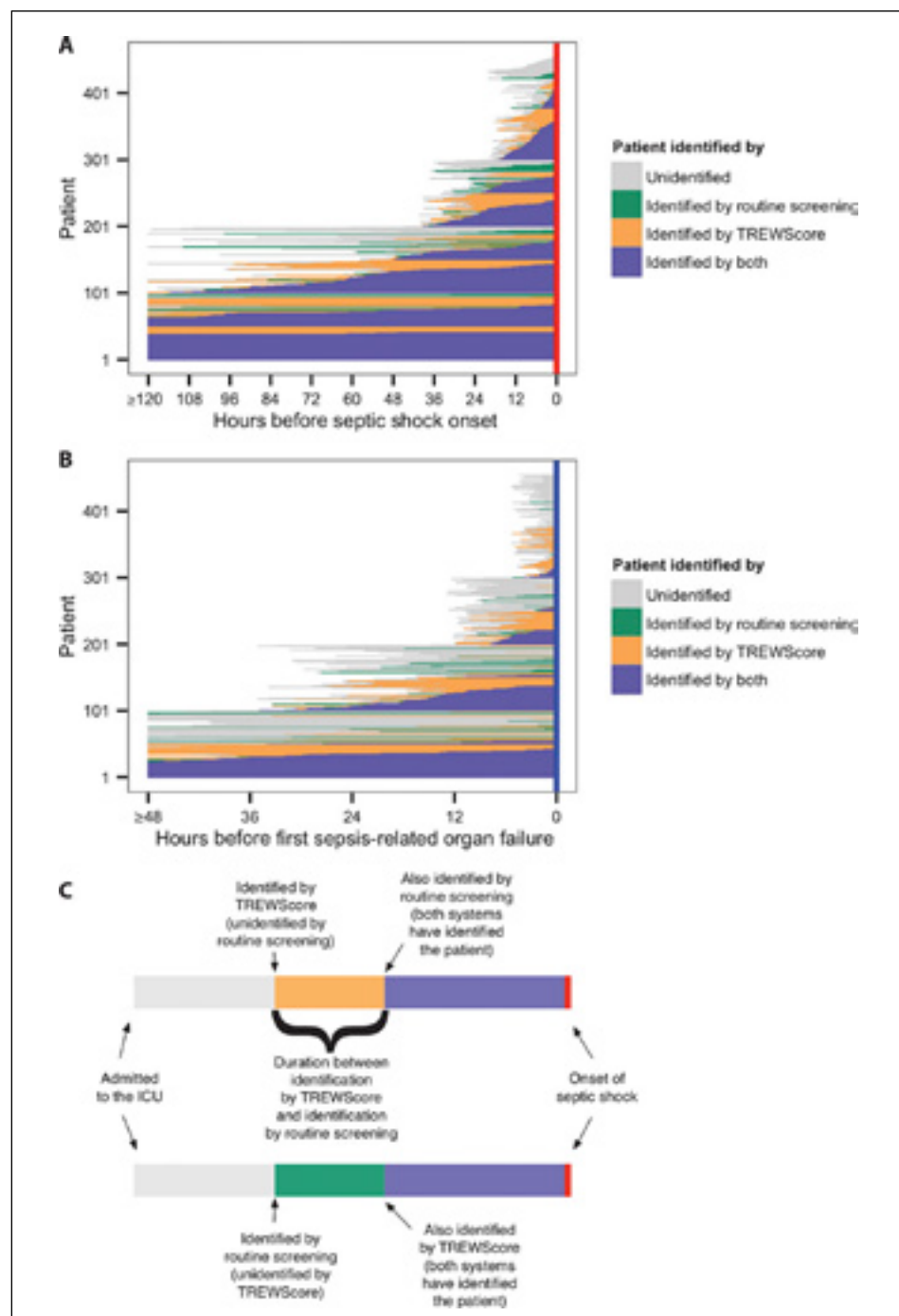
Sontag is also looking into using the anchor framework to predict future

events. For example, researchers can look at people who died and then project backward to identify key characteristics in their EHRs several hours or days earlier. These characteristics could then be used as anchors to predict a current patient's likelihood of dying.

## Individual Risk Stratification: Predicting Chance of Infection

In hospital settings, patients often face an amplified risk of infection either

*Saria and her colleagues compared routine screening procedures to their machine learning-based TREWScore predictions of septic shock during the 120 hour period before septic shock onset (A) and of sepsis-related organ failure during the 48 hours before it occurred (B). Each patient in graphs A & B is represented by a single line (C), with colors reflecting the point at which either routine screening (green) or the TREWScore (orange) or both (purple) predicted septic shock. Thus the quantity of orange in the graphs reflects the success of the TREWScore compared with the quantity of green (routine screening). From KE Henry, DN Hager, PJ Pronovost, S Saria, A targeted real-time early warning score (TREWScore) for septic shock, Science Translational Medicine 7:299:122 (2015). Reprinted with permission from AAAS.*



because they have an underlying disease, their immune systems are compromised, or they've been overtreated with antibiotics, creating a hospitable environment for antibiotic-resistant bacteria. Predicting which patients are most vulnerable could allow healthcare providers to intervene sooner to prevent or control infections. Already researchers are using EHR data to predict two of the most challenging in-hospital infections: sepsis and *C. difficile*.

Sepsis occurs when the body's response to infection begins to shut down the body's organ systems. It's associated with 20 to 30 percent of all hospital deaths each year in the United States—that's about 750,000 people. Automated screening tools have been used to predict that a patient is experiencing sepsis, but none can predict it in advance. "The question was, 'How can you detect sepsis without having to suspect it?'" said **Suchi Saria, PhD**, assistant professor of computer science at Johns Hopkins University, at the Big Data in Biomedicine Conference at Stanford University. She and her colleagues set out to determine whether EHR-based predictions could outperform the standard of care. They developed a score—the TREWScore—that relies on continuous sampling of the EHR. If the score crosses a certain threshold, it is highly predictive of septic shock.

"Using routinely collected data we were able to predict individuals who experience septic shock on average 25 hours early," Saria said. "That's a huge window for intervention." The work was published in *Science Translational Medicine* in August 2015. Further, she adds, "TREWScore is only a starting point. A lot more can be done to target TREWScore to the individual." Her team is actively working on this and she already sees promise.

Wiens and **Erica Shenoy, MD, PhD**, of Massachusetts General Hospital (MGH) took on a different problem that plagues hospital inpatients: *C. difficile* infection (CDI), which causes diarrhea and colitis. CDI is often caused by antibiotic treatment that eliminates the good bacteria in a person's gut, leaving them

vulnerable to the *C. difficile* bacterium.

Like Saria's sepsis work, Wiens's CDI work generates a score for the probability that a patient will test positive for the infection at a later time during the hospital visit. Her algorithm uses two modeling approaches jointly: a time-invariant predictive model that pools data over several days prior to a positive *C. difficile* test, as well as individual daily models that evaluate which parameters are important on each day leading up to the diagnosis. "Other approaches assume a pattern," she says. "We just let the data speak."

The work, which was published in the *Journal of Machine Learning Research* in 2016, identified both expected and unexpected risk factors that contributed to CDI. Patients taking common antimicrobials or proton pump inhibitors were already known to be at high risk for CDI. More surprising, Wiens says, were factors like location in the hospital and the use of opioids. "It's not clear if that's causal," Wiens says, "But it's a hypothesis that can be tested."

The CDI risk score will be applied next at MGH, and will automatically produce a risk estimate for each patient every day at midnight. Wiens and her colleagues are planning a randomized controlled trial to estimate the potential impact of

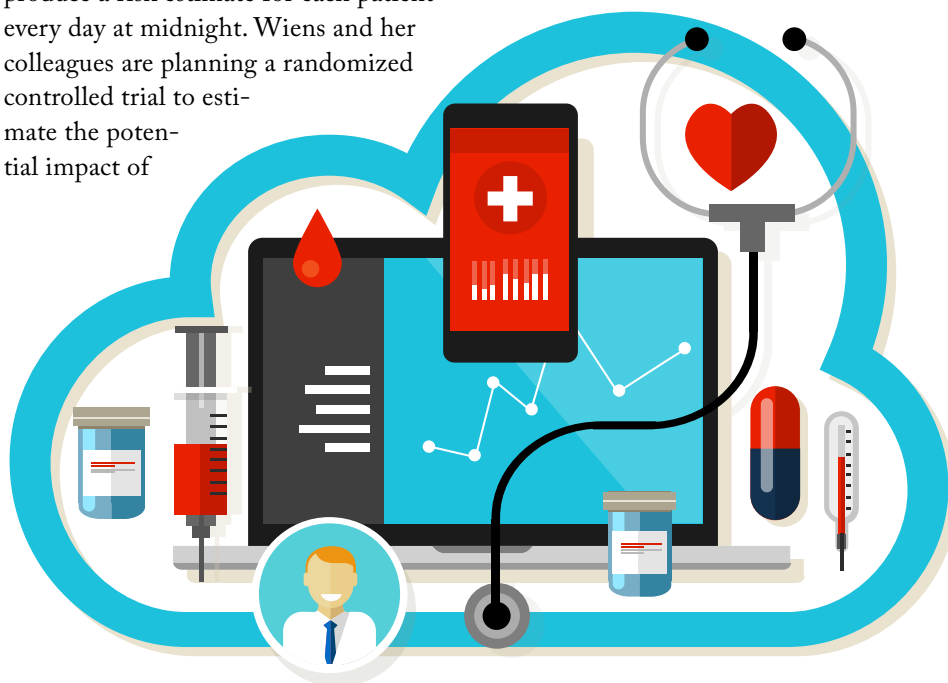
for the two groups and will assess any impact on antimicrobial use and costs.

Modeling each inpatient hospital day and then combining it with a more general model, as Wiens and her colleagues have done for CDI, could prove useful for predicting the progression of other diseases as well. The approach could also generalize more broadly. "You could look at longer time scales to capture how risk factors change over a patient's lifetime," Wiens notes.

Another important direction for the future: combining EHR data with omics data, such as the microbiome. "We're working on that right now," Wiens says. "How much can we predict based on the EHR and microbiome separately versus by combining the two?"

### *The Informatics Consult: Data Analysis for One Patient at a Time*

One of Shah's goals is to develop a medical specialty he calls the "informatics consult." Using machine learning and an EHR warehouse, an informat-



risk-driven interventions.

The planned study will screen for all patients that are at high risk for CDI, but only intervene in a subset of that group. Wiens and her colleagues will then measure the incidence and severity of CDI

ics expert would be available to advise physicians about the prognosis or treatment options for a particular patient. And clinicians would request a consult just as they do from other medical specialists, such as pathologists or radiologists.

To launch a consult, the clinician would describe the patient—Shah posits a 55-year-old Vietnamese woman with asthma and moderate hypertension—and ask for an appropriate treatment intervention. The clinician knows that an antihypertensive medication is appropriate, but which one works for middle-aged asthmatic females who also happen to be Vietnamese? The informatics consultant would then use the EHR to identify similar patients and the most effective treatments for them. If the EHR system contains only five people who match that patient, the consultant might relax the age or ethnicity conditions to get a bigger sample.

“It makes intuitive sense that being able to make decisions using similar patients would lead to better decisions,” Shah says, “but that’s still a hypothesis.” He plans to test that hypothesis in the coming year. The initial pilot will include a limited number of clinicians who will send a consult request over phone or email. “It’s not fully automated and black-box yet,” Shah says. “People might not trust it; and we’re still not at a stage where, technically, we can shrink wrap it and make it into a button.” But the process would be semi-automated in the sense that the informatics expert gets the question, uses a search engine to find a set of similar patients, and then—depending on the question—applies an appropriate statistical method to the EHR data. “There has to be a human in the loop,” Shah says. But in two to four hours, the consult would generate a predesigned report. “That’s my hope for the first pass,” Shah says.

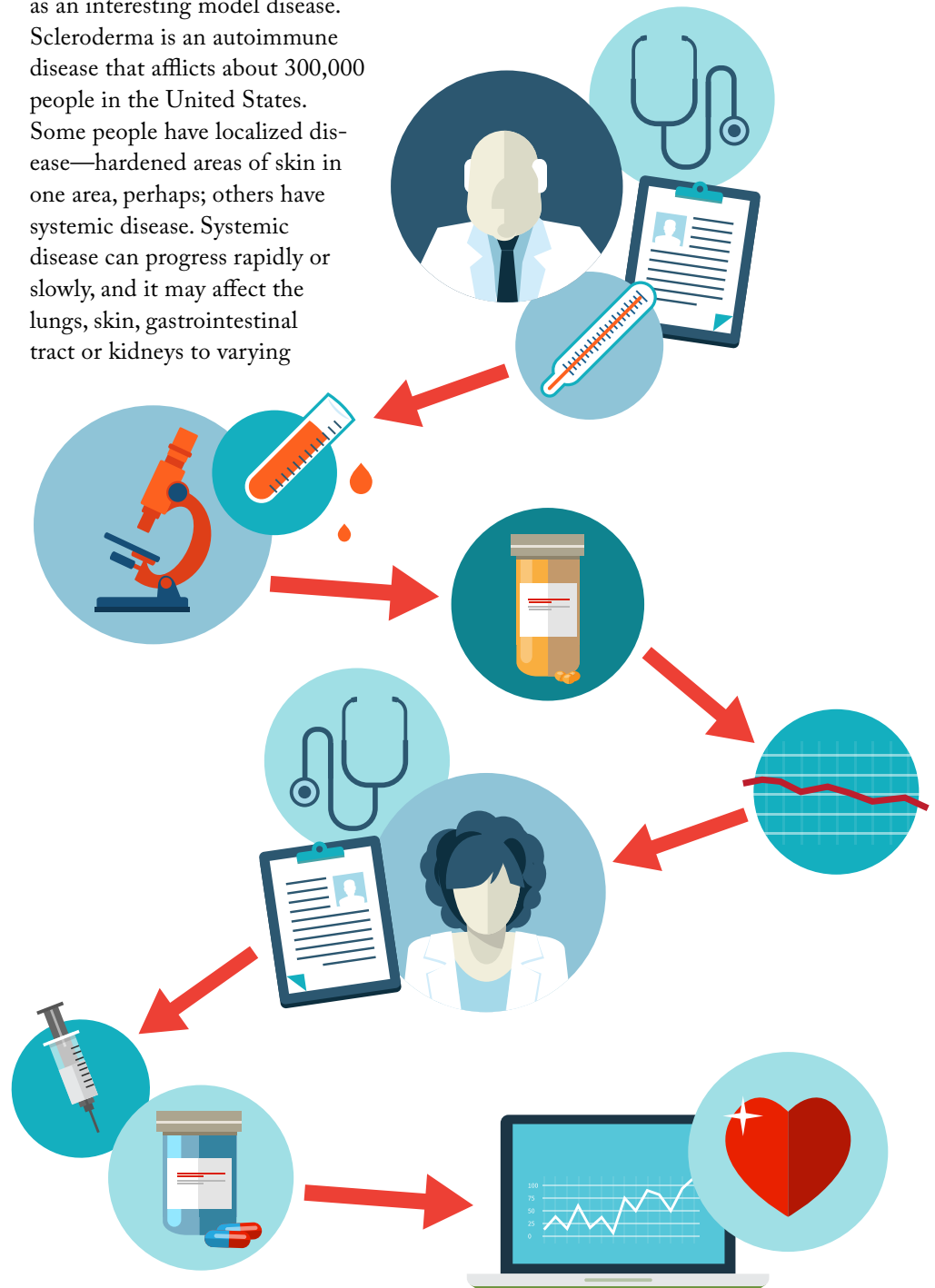
After completing the pilot, they’ll refine the procedure and implement a randomized trial. Some physicians will have access to the consult and others won’t. After a year, Shah’s team will look for differences in outcomes such as the cost of care; speed of recovery; and patient well-being and satisfaction

## Predicting Disease Progression

One of the toughest questions for clinicians to answer is: “How will my disease play out?” So Saria and her

colleagues decided to experiment with establishing a computational framework for predicting disease trajectories in chronic, complex diseases using EHR data. They settled on scleroderma as an interesting model disease. Scleroderma is an autoimmune disease that afflicts about 300,000 people in the United States. Some people have localized disease—hardened areas of skin in one area, perhaps; others have systemic disease. Systemic disease can progress rapidly or slowly, and it may affect the lungs, skin, gastrointestinal tract or kidneys to varying

scleroderma-related lung disease using a measure of lung health called PFVC (percent of predicted force vital capacity). Saria’s team trained a predictive model



extents. For physicians, it can be hard to know what treatments are appropriate.

Lung disease is the leading cause of death among scleroderma patients but the decline in lung function is unpredictable. So Saria’s team honed in on predicting the progression of

using data on 672 individuals collected over a period of 20 years in the Johns Hopkins Scleroderma Center patient registry. Using these data, they were able to uncover several new subtypes of lung disease progression. As time passed, the team could also dynamically personalize



predictions of lung disease progression for specific individuals. Saria has also recently shown how to account for progression in trajectories across many different organ systems in scleroderma, offering the possibility of individualizing management of systemic diseases that, like scleroderma, affect more than just one organ. Saria says the approach could be applied to other complex diseases such as asthma, autism, and cardio-obstructive pulmonary disease (COPD).

## One-Button Predictions: Forecasting ALL Diagnoses

Rather than focus on individual disease risks, Page and his colleagues at the Center for Predictive Computational Phenotyping (CPCP), a Big Data to Knowledge (BD2K) Center of Excellence at the University of Wisconsin, are building a predictive model for every

current or new patients, the trained system calculates the probability each person will be assigned each diagnostic code within the next six months, Page says. The system works well even for predictions six months out, though some diseases can be predicted more accurately than others, he says.

Page hopes that the Marshfield Clinic's EHRs will start to use the system, at least for the most accurately predicted diseases. Perhaps it could offer physicians a pop-up alert if a patient crosses a threshold of risk for particular diagnoses. At the same time, he'd also like to do a careful test of whether physicians actually rely on the pop-ups. "The hope is that the prediction takes into account more features than the doctor can in one visit and can improve care," Page says.

But the work could also be useful in other ways—to help hospitals evaluate how well they are doing at treating high-risk patients system-wide, for example; or to pick potential cohorts for trials of preventive procedures; or to discover unknown long-term effects of treatments. "It could put things on the radar that aren't on there yet," Page says. "We're still at the point now where there's lots of interest and excitement about the possibilities for predictive models in the clinic, but very little translation. This work could speed up that process."

Rather than using random forests to evaluate each patient's risk of every disease, a team of researchers working with **Joel Dudley, PhD**, assistant professor of genetics and genomic Sciences at the Icahn School of Medicine at Mount Sinai in New York City, used neural networks to extract a "deep patient representation" (called Deep Patient) from 700,000 patient records in the Mount Sinai Health System's data warehouse and then tested its ability to predict the likelihood of 78 diseases in more than 70,000 patients. Shah, who did the initial data processing for the project, says Deep Patient created complex features out of the words mentioned in patient records. "It's a representation of the EHR data for risk stratification," Shah says.

Dudley's team found that Deep Patient outperformed a number of other prediction methods at predicting future assignments of disease codes. The research, which was published in *Scientific Reports* in May of 2016, could also be useful for personalizing prescriptions or recommending treatments, the paper suggests.

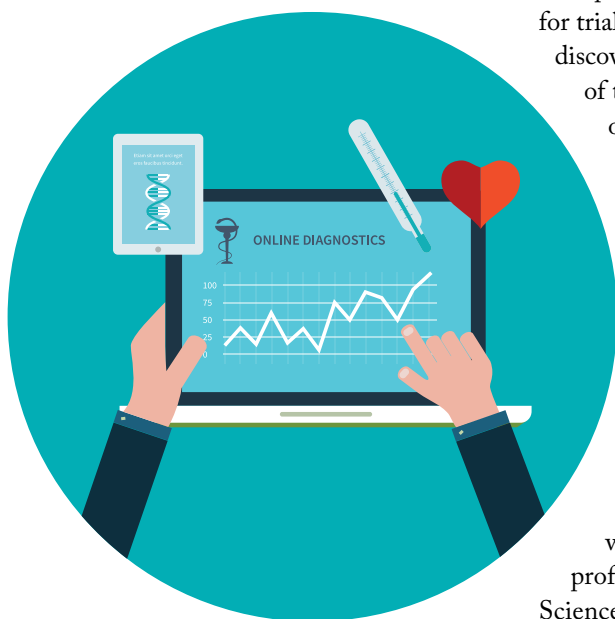
But neural nets have a downside: They don't give users an intuitive sense of what's going on. That's because they are based on finding hidden features in the data. So using Deep Patient, physicians might reliably tell patients their risk of a disease, but they wouldn't be able to point to potential reasons why.

## Getting at Causation

It would be nice to go beyond predictions based on similarity to predictions based on causality, Sontag says. "The machine-learning community has, for the most part, ignored this causal inference question in recent years, but in the healthcare setting it's the most important question," he says.

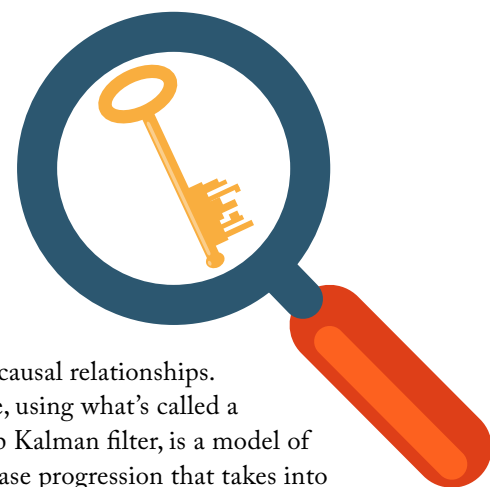
Sontag and his team, including PhD student **Rahul Krishnan** and postdoc **Uri Shalit**, are currently developing several statistical approaches to discover-

ing causal relationships. One, using what's called a deep Kalman filter, is a model of disease progression that takes into consideration how drugs or treatments affect disease progression. The approach would allow researchers to ask, for example, "What would have happened to this patient if he/she had had Treatment B instead of Treatment A?" Sontag says. He's getting initial results now and says: "I view this type of work as the future of precision medicine." □



diagnosis code at a press of the button. The work relies on a high-throughput computing system called HT-Condor and the Marshfield Clinic's EHRs for more than a million patients.

To train their machine-learning algorithm, Page's team used a statistical approach called random forests—essentially a series of decision trees that identify the most informative features for each diagnostic code, then the next most informative and so on. Given a set of







# Taking on the **EXPOSOME**

*Bringing Bioinformatics Tools to the  
Environmental Side of the Health Equation*

BY KRISTIN SAINANI, PhD







When it comes to what kills people, Nurture trumps Nature:

Chronic diseases with overwhelmingly environmental (rather than genetic) causes are responsible for the deaths of two-thirds of the world's population. Yet the investment made in unraveling the environmental side of the health equation pales by comparison to the investment in human genome research.

"In the past 20 years, a lot of effort and funding have pointed toward genome research," says **Paolo Vineis, PhD**, professor of medicine and chair of environmental epidemiology at Imperial College London. "Now, people are suggesting that a similar effort should be put into exposure research, and also that exposures should be investigated systematically as has been done for the genome, such as with Genome-Wide Association Studies, or GWAS."

Though we know some of the biggest players in chronic diseases—air pollution, smoking, poor diet, and lack of exercise—an estimated 50 percent of the environmental drivers remain unknown. "I'm not going to argue that diet or physical activity or smoking don't have a role to play," says **Chirag Patel, PhD**, assistant professor of biomedical informatics at Harvard University. "But it behooves us to explain more of the variation than can be explained by classical environmental factors. We need to look beyond the proverbial lamppost."

Environment-disease research suffers from the same problems that gene-disease research did 20 years ago: Individual labs study hand-picked risk factors one at a time in small studies with inconsistent methodologies; and they are incentivized to report positive findings. The result: a literature rife with spurious findings. "There's a now-famous number being punted around in genetic epidemiology that, prior to GWAS, over 95 percent of the findings from candidate gene studies—that is, your favorite gene in connection with a trait—are false," Patel says. In a 2011 review, researchers found that only 13 of 1,151 purported loci-phenotype associations for eight conditions were replicated in large-scale studies. It took GWAS and related approaches—which consider a multitude of



genes simultaneously in an unbiased, standardized way—to clean up this literature.

We need a similar revolution in the study of environment-disease associations, Patel and others say. In 2005, **Christopher Wild, PhD**, now director of the International Agency on Cancer Research, coined the term “exposome” as a call for high-throughput, systematic approaches to studying how the environment impacts health. Echoing this call, Patel and others coined the term EWAS, or Environment-Wide Association Study, to encourage researchers to apply GWAS-like methods to study health-environment associations.

The exposome encompasses the entirety of a person’s exposures from birth to death. Thus, the first challenge is how to measure it. Fortunately, technological advances are making it possible to measure the exposome at higher resolutions and on larger scales than ever before. Metabolomics measures the chemical ghosts of exposures in our blood; wearable sensors and smartphones track where we go, what we breathe and eat, how we move, and how we feel; social media sites amass records of our moods and social connections; electronic health records store our clinical, personal, and demographic attributes; and geographical information systems and survey data reveal the wider societal factors that influence our health.

The sheer volume and complexity of these data are overwhelming. According to **Gary Miller, PhD**, professor of environmental health at Emory University in Atlanta, Georgia, a geneticist on his staff once commented that after he saw how complicated the exposure data were, she felt like “a wimp” for studying genetics. Whereas genomic data consist of stable linear sequences, exposome data are heterogeneous, non-linear variables that change over time and space. Dense webs of correlation among environmental variables make it hard to tease out causation. And, due to the highly personal nature of the data, privacy and security concerns abound. Exposome researchers can draw heavily on the bioinformatics tools developed for GWAS, but to fully realize the promise of the exposome, they will need new tools for storing, integrating, and analyzing the data.

“It’s daunting. It’s definitely hard,”

Miller says. But it’s also an opportunity for bioinformaticians and computational biologists, he adds. “For people who like wrangling with data, the exposome offers some great challenges.” This article reviews recent progress in exposome research and the challenges that remain for studying everything from the chemicals in our bodies to the quality of our neighborhoods.

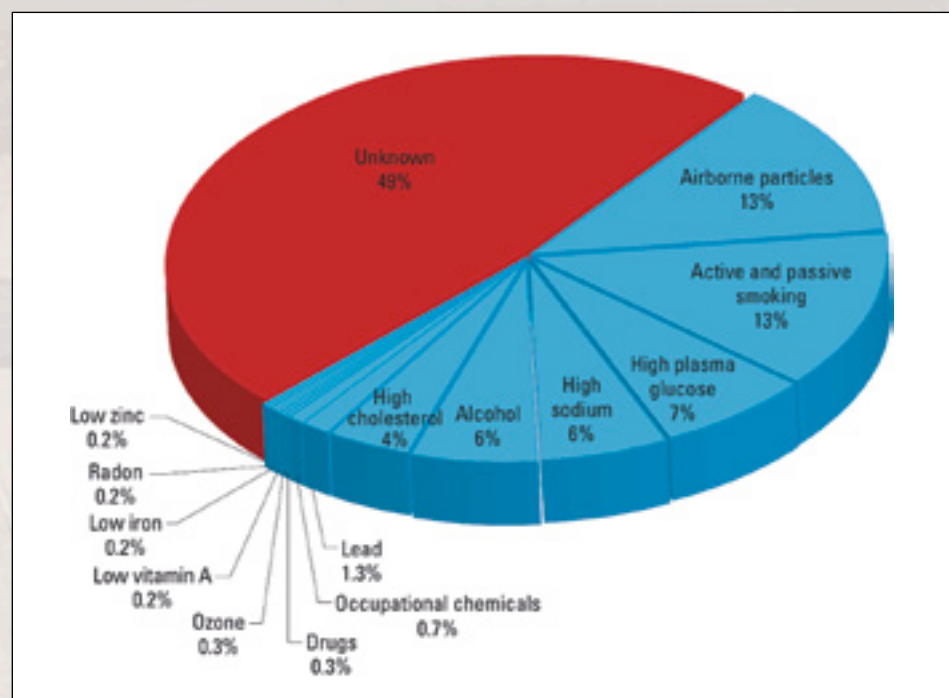
## WHAT’S INSIDE: METABOLOMICS

External exposures leave chemical traces in our bodies. These can provide a convenient window into how those exposures affect health. “Exposures are inherently chemical in nature,” says **Stephen M. Rappaport, PhD**, professor of environmental health sciences at the University of California, Berkeley. “Anything that causes a health effect is either a chemical or is mediated through chemicals.” Food, drugs, and pollutants leave behind metals and small molecules in the blood. “Even psychosocial stress produces hormones and other biologically relevant molecules in the body,” Rappaport says.

Fortunately, researchers who want to

perform large-scale exposome studies can access troves of specimens and associated health outcome data that have been collected and archived by epidemiologic studies and national surveys. In 2010, when Patel was a doctoral student at Stanford, he and his mentors performed the first proof-of-principle EWAS using publicly available data from the National Health and Nutritional Examination Survey (NHANES), which includes data on chemicals in the blood and urine of thousands of participants. When they compared 266 chemicals across participants with and without type 2 diabetes, they turned up four hits: the pollutants polychlorinated biphenyls (PCBs) and heptachlor epoxide and the nutrients vitamin E and beta-carotene (the latter was inversely associated with diabetes). Follow-up studies are needed to determine if any of these factors is causally related to diabetes, Patel stresses. “But by taking a data-driven, agnostic, unbiased approach, EWAS leads to a more reproducible list of hypotheses to prioritize for further study.”

Rappaport concurs: “All we want to do with EWAS is to sort through the thousands of chemicals to which people



**INTO THE UNKNOWN:** Though we know many of the environmental risk factors for chronic diseases, about half remain unknown. This chart shows the percent of total global chronic disease deaths that are believed to be explained by each factor, according to 2010 data from the World Health Organization. Reprinted from Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. 2014. The blood exposome and its role in discovering causes of disease. *Environ Health Perspect* 122:769-774.



are exposed during life and identify those few exposures that may be causes of disease. Then epidemiologists can follow up with focused studies to establish causality. Thus, the exposome paradigm begins with a data-driven EWAS to generate hypotheses and ends with tests of these hypotheses in subsequent stages.”

Patel’s team had developed publicly available software for EWAS (<http://www.chiragipgroup.org/exposome-analytics-course/>) that combines off-the-shelf GWAS tools with cutting-edge machine-learning techniques. “There’s nothing novel in the methods. Rather, we are taking existing methods that statisticians and informaticians have developed for different domains and introducing them to people doing exposure science and epidemiology,” Patel says.

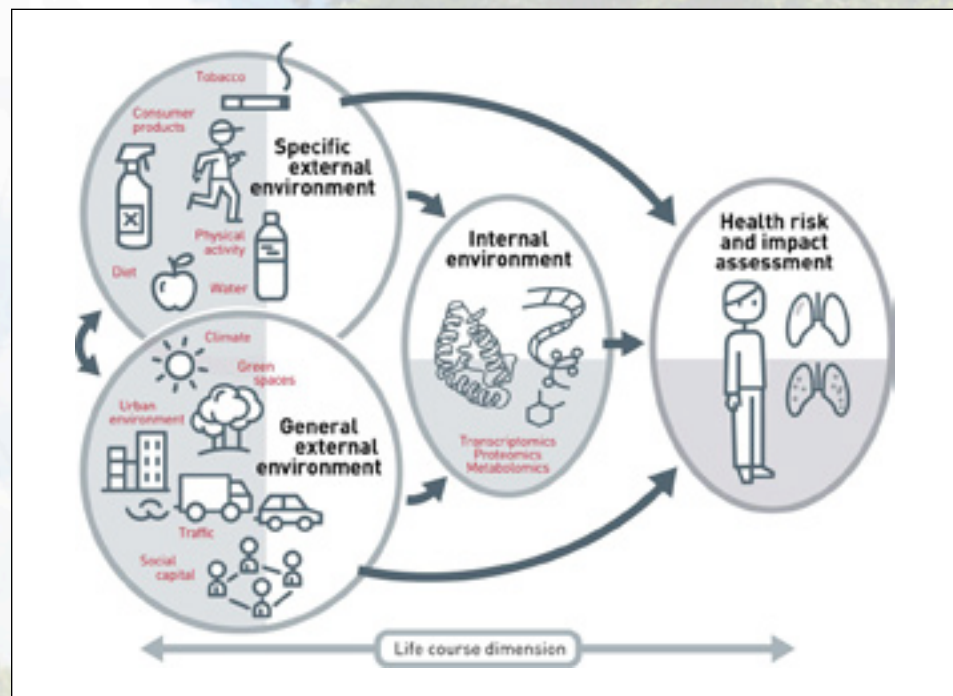
Exposome researchers dream of a day when there is a cost-effective exposome chip akin to the SNP (single nucleotide polymorphism) chips that enabled GWAS studies. “If you could measure even 500 chemicals consistently in human plasma, and you could do it in a cost-effective way at the scale of a GWAS, you would start finding things,” Miller says.

To look for novel triggers of disease, many exposome researchers are also widening their search beyond known chemical markers. They are turning to untargeted metabolomics—using mass spectrometry to explore the vast landscape of unknown chemicals in the blood. Platforms can now measure 100,000 small molecules from a few microliters of blood in 20 minutes, Rappaport says. The catch: Mass spectrometry just gives signatures of chemicals, or spectral peaks; so, once researchers have fished out the most interesting peaks, they still need to work out the identity of the chemicals. Spectral reference libraries exist, but they cover only a small fraction of the metabolome, so chemical identification remains a challenge.

Rappaport’s lab is nevertheless taking this approach. To ensure they are picking up causes rather than effects of disease, they use archived samples from cohorts of people who were healthy at the time of the blood draw. For example, to look for clues to childhood leukemia, Rappaport’s team is using neonatal blood spots

collected on all babies born in California since the mid-1980s. By comparing the metabolomic profiles of 1,000 babies who later developed childhood leukemia with those of 1000 comparable controls, they hope to identify possible pre-natal causes of leukemia. They are also looking for evidence of exposure to damaging reactive molecules by measuring telltale alterations of the blood protein serum albumin (called adductomics). “Adducts from albumin are interesting because they stick around for

eggs—bacteria in our guts convert the fat into trimethylamine N-oxide, or TMAO. Animal studies showed that TMAO clogs arteries. And subsequent human studies showed that individuals with high levels of TMAO are 2.5 times more likely to have major cardiovascular events (heart attack, stroke, or death) than those with low levels. The American Heart Association and American Stroke Association listed TMAO as one of the top 10 advances in heart disease and stroke science for 2013.



**THE EXPOSOME:** The exposome encompasses the entirety of a person’s exposures from birth to death, including internal exposures such as gut bacteria, lifestyle choices such as smoking, and social determinants such as poverty. Reprinted from M Vrijhied, *The exposome: a new paradigm to study the impact of environment on health*, Thorax 69:876-878 (2014) with permission from BMJ Publishing Group Ltd.

a month. So we’ll get a picture of what babies were exposed to during the month preceding delivery,” Rappaport says.

It’s too early to know what Rappaport’s study will turn up. But the power of the metabolomic approach is illustrated by a series of studies from the Cleveland Clinic, including 2010 and 2013 papers in *Nature* and the *New England Journal of Medicine*, respectively. Researchers compared stored blood samples from 150 people who developed a heart attack or stroke with 150 age and gender-matched controls. Following up on the strongest signals from mass spectrometry, they uncovered a key metabolic pathway: When we eat lecithin—a fatty acid found in meat and

“If their hypothesis is correct, I think we’re going to see that this has a major impact on how people diagnose and treat heart disease in the future,” Rappaport says.

Success stories like this have been limited, however, due to the lack of informatics infrastructure. Exposome initiatives in Europe and the United States are building infrastructure such as spectral reference libraries, shared data platforms, and analysis tools. For example, Vineis leads a consortium of 12 European institutions, called EXPOsOMICS (<http://www.exposomicsproject.eu/>), while Miller leads The Emory Health and Exposome Research Center: Understanding Lifetime Exposures (HERCULES, <http://emoryhercules.com/>).







2003. But large-scale exposome studies using behavior trackers remain rare. Since many technologies have only become available recently, scientists are still testing their usability and accuracy. “We’ve spent so much time investigating the reliability of the devices,” Kerr says. Researchers are also grappling with how to deal with the quantity of data. NHANES has seven terabytes worth of accelerometer data, including 150 million data points per person. Besides issues of storage, it’s unclear how to process such data. How do we extract meaning out of 150 million data points—do we look at averages, slopes, standard deviations, or more complicated statistical measures? Two of NIH’s Big Data to Knowledge (BD2K) centers—The Mobilize Center at Stanford and Mobile Sensor Data-to-Knowledge (MD2K) center—are grappling directly with this issue (See *BCR* story: “Wearing Your Health on Your Sleeve”). Privacy is another concern. Kerr outfits study participants with personal cameras, which end up photographing people who are not involved in the study. “Because we have that type of information, we have to handle it in a very secure way. We have to be very careful about our ethical framework,” Kerr says.

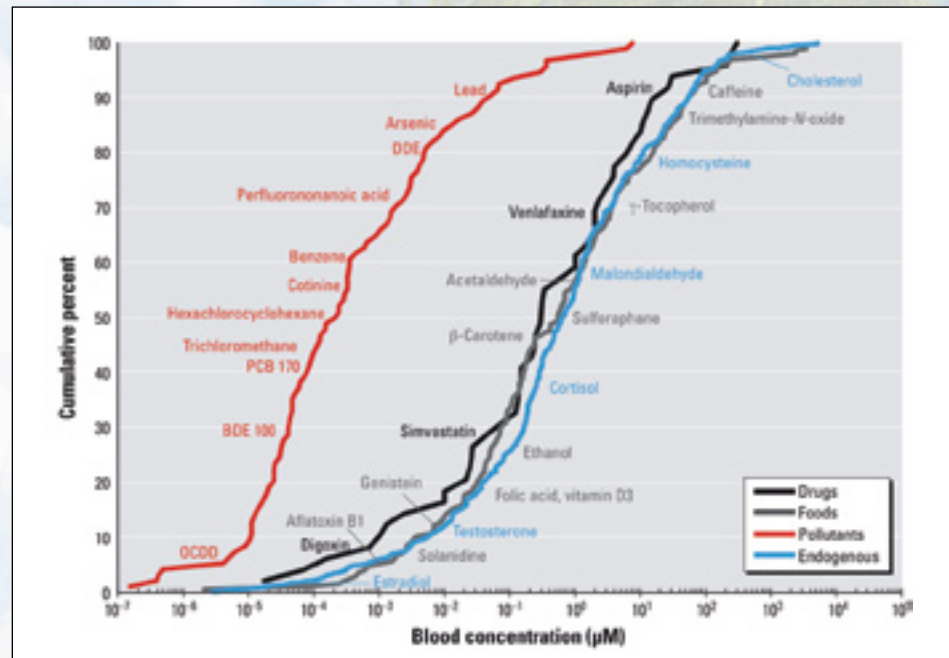
Exposome researchers are also hoping to tap into the massive amounts of personal health data being collected outside of mainstream research. Twenty percent of Americans own a health wearable, such as a fitness band or smartwatch. If just a small fraction is willing to share these data, this translates to huge sample sizes. Many challenges in using and accessing these data remain, however. For one thing, people who are willing to share their data tend to be very different from the average American. “We’ve looked at typical journeys that you might be able to get from Strava, the GPS-based biking system. And they look nothing like the typical journeys that we get in our study participants,” Kerr says. “The data probably don’t represent a lot of the underserved groups that we’re trying to reach.”

Also, the commercial companies that own the data are often unwilling to share, Jacquez says. He hopes to see more “benefit corporations,” or “B-corporations” set up to sell these devices. B-corporations

blend traditional for-profit and non-profit business models—they make money, but are also committed to serving society. Such companies could make user-generated data

predicted mortality as well as high cholesterol and high blood pressure.

Constructs such as stress and social isolation may seem “squishy” and hard



**EXPOSURES IN OUR BLOOD:** Summary of small molecules and metals in human blood. Each curve represents the cumulative distribution of chemical concentrations from a particular source. Concentrations of drugs, foods, and endogenous chemicals are several orders of magnitude higher than concentrations of pollutants. Reprinted from Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. 2014. The blood exposome and its role in discovering causes of disease. *Environ Health Perspect* 122:769-774.

freely available to research scientists. “This would be a model for people sharing their data for the greater good,” Jacquez says.

## HOW WE FEEL AND RELATE: ELECTRONIC FOOTPRINTS

The exposome encompasses a wider set of psychological, social, and behavioral variables that include stress, subjective well-being, personality traits, resilience, social connectedness, and social support. It would be a mistake to neglect these risk factors, says **Nancy Adler, PhD**, professor of medical psychology at the University of California, San Francisco. “The physical environment is concrete and it is related to health, but the effect sizes are small. The associations for some of the social and behavioral variables are actually more powerful.” In one study, her team showed that social isolation

to pin down, but we have well-validated instruments for measuring them from social science and psychology. “We know what the factors are, and we know how to measure them with self-report,” says **Elissa Epel, PhD**, professor of psychiatry at the University of California, San Francisco. The ability to measure these constructs electronically—via mobile phones, social media, and electronic health records—opens the door for their widespread inclusion in exposome research.

Smartphones can measure stress and other emotional states and behaviors in real-time. In Ecological Momentary Assessment (EMA), people are randomly pinged throughout the day and asked questions such as: What’s your mood? How stressed are you? Who are you talking to? Do you have a craving for food? Did you overeat? “We can characterize people in their natural environment in a fresher, closer way to their actual experience,” Epel says.

EMA gives a much richer set of



data than could be obtained from a few questions on a survey. But it also presents challenges for data analysts. “We’re good at collecting masses of data and we haven’t caught up to being able to use it well and create meaning out of it,” Epel says. “We’re in need of data scientists who can manage and make sense of these data. It is a hot new area that we need to be training more scientists in.”

Others are gathering data from social media sites. Sabel uses Twitter to study emotions, for example. People’s tweets objectively reveal their moods, Sabel says. “The idea of mining data from

Many large epidemiologic surveys also include stress-related variables. For example, the Health and Retirement Study—which has been following 20,000 older adults in the United States for nearly a quarter-century—has periodically queried participants about socioeconomic stressors, such as unemployment and financial hardship. Participants also filled out a one-time survey in 2004 that asked about their exposure to stressful life events—such as divorce, loss, or trauma—in both childhood and adulthood. Using an EWAS approach, **Eli Puterman, PhD**, assistant professor of psychiatry at the

University of California, San Francisco, is asking which of 92 variables available in the Health and Retirement Study is most strongly linked to mortality. “I think what’s really exciting about it is that we’re allowing the data to speak for themselves,” Puterman says.

Epel co-leads the Stress Measurement Network, a consortium that aims to deploy more and better measurements of stress in large epidemiologic studies. In particular, more subjective measures of stress are needed, Epel says. “You cannot know how someone

is feeling unless you ask them. That’s one case where we absolutely need self-report.”

Beyond epidemiologic studies, electronic health records (EHRs) offer a huge opportunity for exposome researchers. “If we had interoperable EHR records that had these data in them, we could really start to study the exposome,” Adler says. She participated in an Institute of Medicine panel tasked with recommending social and behavioral measures for inclusion into EHRs. The panel devised an 11-item battery that included one or two questions each on smoking, physical activity, education, race/ethnicity, and home address, as well as four questions on social connection and isolation.

Getting health care providers to implement the battery is challenging, but Adler notes that doctors are increasingly being held accountable for patient outcomes. “Once doctors are on the hook for keeping

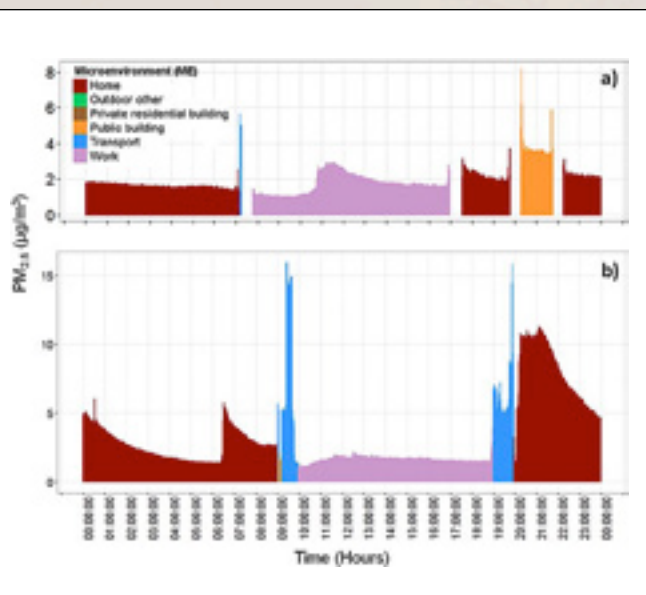
people well, they start to pay much more attention to the things that really drive their health, many of which are social,” she says.

## HOW THE DECK IS STACKED: GEOGRAPHICAL INFORMATION SYSTEMS

Many factors that influence our health operate at the societal rather than individual level: what culture we come from, whether we live in poverty, whether we have access to health care and high-quality education. “There’s a bit of a paradigm shift to say behavior is not just an individual choice. It’s also constrained by the social environment this person is in and their financial resources,” Adler says. To get at these macro-level factors, exposure researchers are using geographical information systems. “Geo-coding is really opening up possibilities of linking what’s going on in neighborhoods and communities to disease outcomes,” Adler says.

For example, **Paul Juarez, PhD**, professor of family and community medicine at Meharry Medical College in Nashville, Tennessee, uses mapping technology to study health disparities. Juarez and his team created the Public Health Exposome Database, which contains 15,000 data points on each of 3,100 counties in the U.S.—including data on water and air pollution; availability of sidewalks and grocery stores; education and poverty; local, state, and federal laws pertinent to health; and health outcomes. “With county level data, you can do some great maps and show the hotspots and patterns,” Juarez says. “People understand maps better they do spreadsheets.”

To analyze the data, “we’ve had to go out and recruit people who have big data skill sets,” Juarez says. For example, he collaborates with **Michael A. Langston, PhD**, professor of electrical engineering and computer science at the University of Tennessee, who uses graph theory to analyze big datasets. “We have these tools that we’ve built over decades and applied to problems that arise in many disciplines. We just need to map them over to the



**PERSONAL EXPOSURE MONITORING:** This figure shows daylong recordings from personal air pollution monitors ( $PM_{2.5}$  = particulate matter smaller than 2.5 micrograms) for two different people. Colors indicate different microenvironments. For example, the person pictured in (b) experienced high levels of pollution while traveling between home and work (transport periods are in blue). Reprinted from S Steinle, S Reis, CE Sabel, et al., *Personal Exposure Monitoring of  $PM_{2.5}$  in indoor and outdoor microenvironments*, *Science of the Total Environment* 508:383-394 (2015).

Twitter is that it’s like you’re looking at them without them knowing that you’re listening.” He looks for positive emotions expressed in tweets and links these to the locations people are tweeting from (from GPS). One drawback with Twitter data is that only about two percent of Twitter users agree to make their location data publicly available, so the sample may not be representative, Sabel says.



exposome setting rather than redesigning them from scratch,” Langston says.

In graph theory, variables are viewed as points in space. Langston’s algorithm examines all pairs of variables in the dataset; if two variables are highly correlated, he’ll put an edge between them. “I have all these points and edges floating around in space and what our algorithms do is find the dense regions—areas where there are a whole bunch of edges, meaning all these variables are moving together.” These dense regions, called paraclicques, can then be correlated with disease outcomes. More refined statistical analyses are then applied to try to isolate the causative factors from the mere confounders.

In one example, Juarez and Langston studied variations in the rates of premature births across counties. The lowest prematurity was found in Marin County, California, and the highest in Hinds County, Mississippi. They considered 590 variables, representing indicators from the economic, health care, physical, and social environments. Of 48 paraclicques extracted, 17 correlated highly with prematurity rates. From there, traditional regression techniques identified race, obesity and diabetes, sexually transmitted disease rates, mother’s age, income, marriage rates, pollution, and health insurance as key drivers of disparities in prematurity rates.

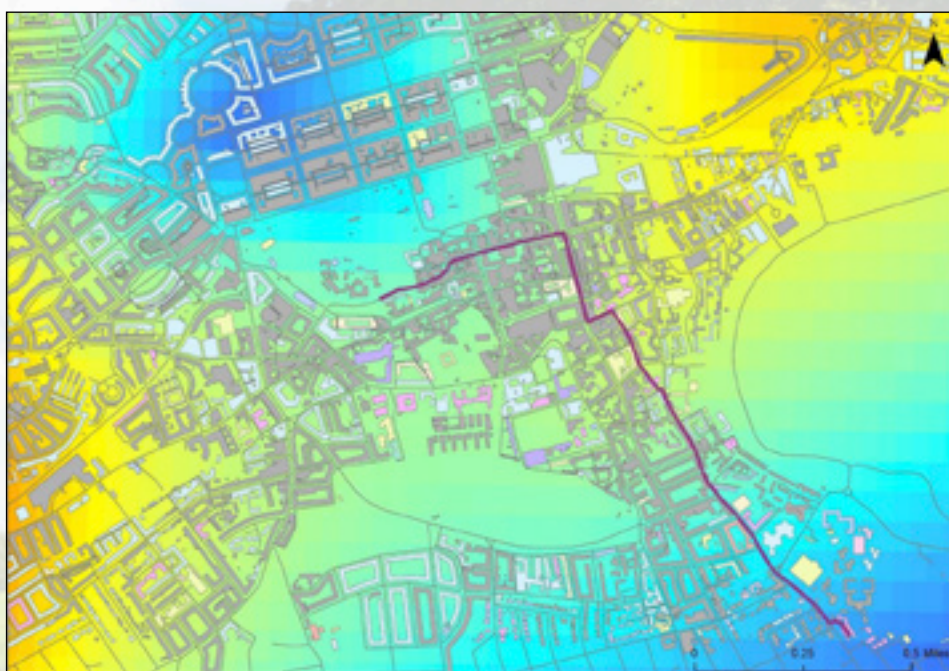
In another example, Juarez and Langston showed that disparities in lung cancer mortality for white men and women were largely driven by variations in smoking rates; but, surprisingly, disparities in lung cancer mortality for black men and women were driven more by differences in poverty, overall health, and access to health care. “The advantage of this data-driven approach is that it allows you to see patterns that you may not have thought about before with a hypothesis-driven approach,” Juarez says.

The lack of high-quality data management tools remains a critical obstacle. “The up-front handling of the data is still back in the stone ages,” Langston says. “Research scientists are going through files by hand, trying to move columns around,” he says. “We learned in biology years ago, if you’re going to deal with large volumes of data, then you’ve got to

bring on board a database administrator and a data curator so the domain experts can concentrate on the science,” he says.

## ASSEMBLING THE EXPOSOME

Bit by bit, researchers are making inroads into the human exposome. But much remains to be done. Besides meeting the challenges already detailed, researchers also must figure out how to integrate all the layers of data—from the chemicals in our blood to the laws in our counties—and also link them to genome data, to get at gene-environment interactions.



**HAZARD MAP:** This map overlays a person’s GPS-recorded travels with a hazard map showing the concentrations of the pollutant nitrogen dioxide (as measured at fixed pollution stations). The map can be used to estimate a cumulative daily exposure for the individual. Concentrations of nitrogen dioxide are lowest in light blue areas and highest in dark yellow/orange areas. Courtesy of Clive Sabel, Bristol University.

The exposome community needs to adopt a “big science” approach akin to the Human Genome Project, comments **Christopher Austin, MD**, director of the National Center for Advancing Translational Sciences at the National Institutes of Health. To “kick it up to this level,” he advises exposome researchers to heed some lessons from the genome community. For example, he says, the exposome community should invest in improving measurement technologies, just as the Human Genome Project did for sequencing technologies; establish a public data

repository similar to GenBank, but for exposures; and agree on standards such as for variable names, meta-data, and security.

The key is to make the data easy to access and use, Austin says. “Otherwise, it becomes what a friend of mine calls ‘data composting’—you just put it on a pile and hope that, if it sits there long enough, something magic will happen.”

On top of all that, Austin says, the exposome community needs strong project management and leadership. With individual-investigator projects, you can make things up as you go along, Austin says. “The building isn’t that big, so if you need to build a foundation halfway through,

you just do it.” But big science projects need to be methodically planned and executed or they risk catastrophic collapse.

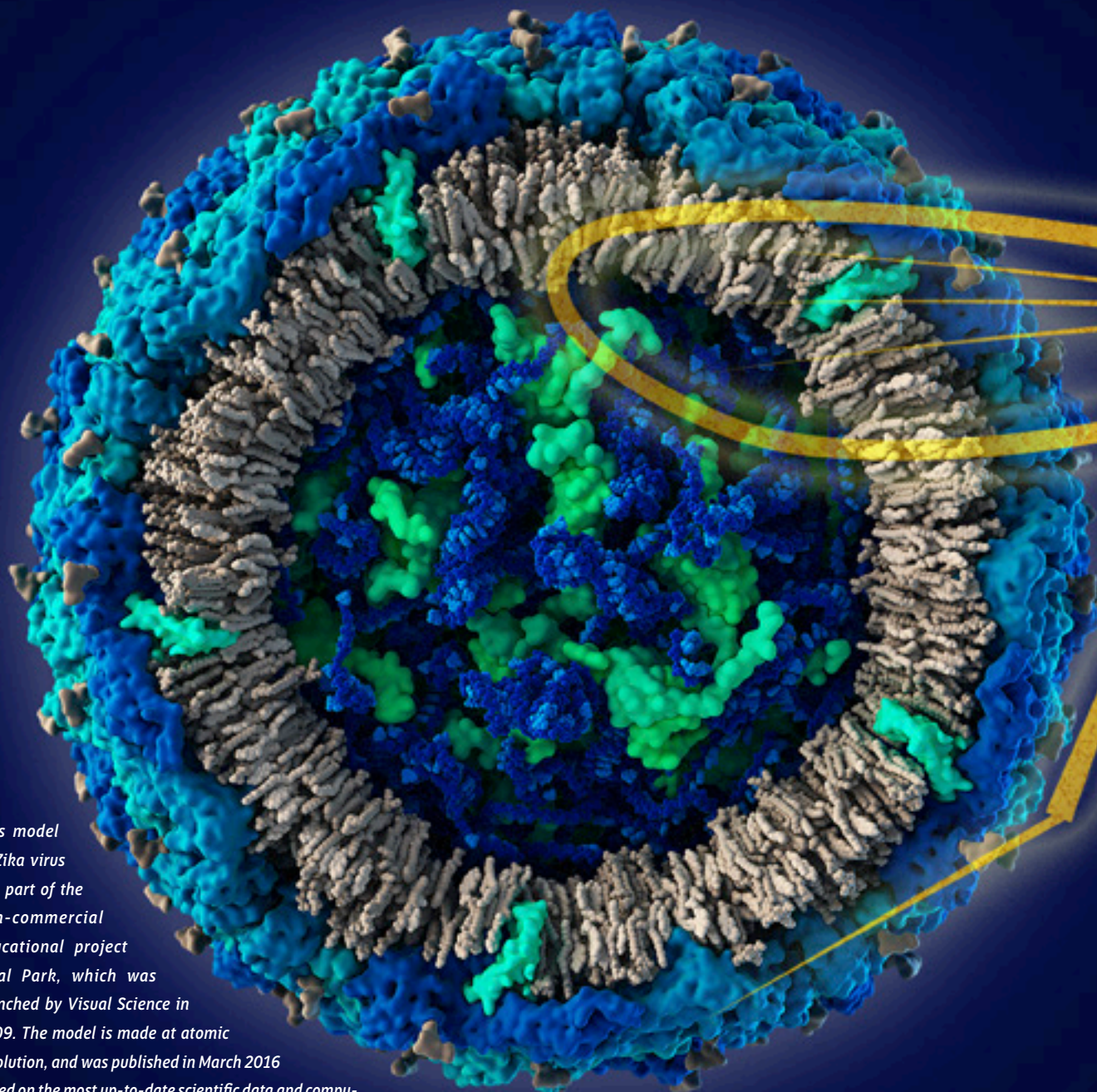
Understanding the exposome is an ambitious idea, Miller says. But it is far from impossible. In the early 1990s, people estimated that it would take 130 years to sequence the human genome. “But once the scientific community said, ‘Okay, we’re going to do it, and we’re going to invest money in it,’ they were able to rapidly accelerate progress and get it done under budget and under time,” he says. “It was really amazing what happened.” □



# Zika!

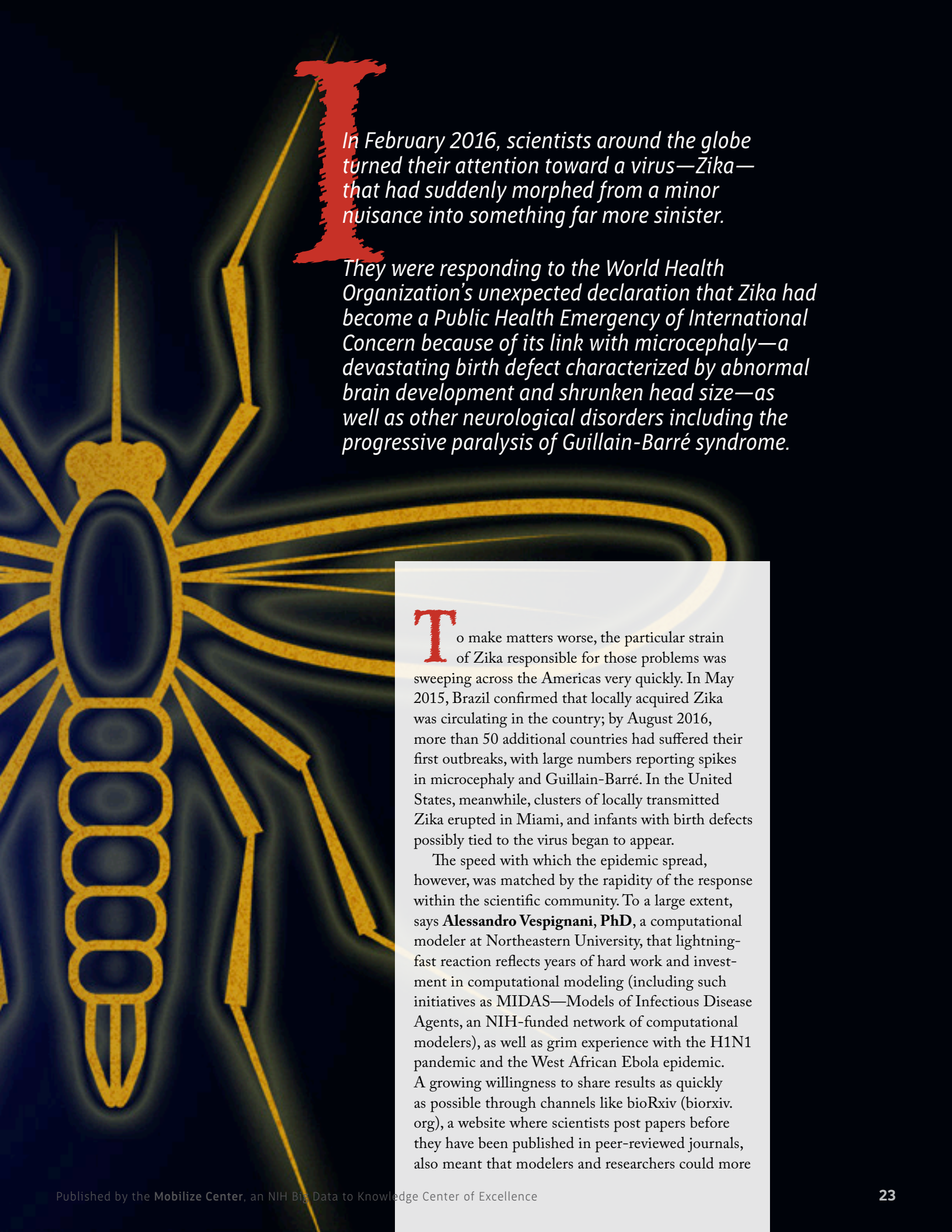
BY ALEXANDER GELFAND

## Computational Biology to the Rescue



*This model of Zika virus is a part of the non-commercial educational project Viral Park, which was launched by Visual Science in 2009. The model is made at atomic resolution, and was published in March 2016 based on the most up-to-date scientific data and computational biology simulations available at that time. Reprinted with permission from Visual Science and creator Ivan Konstantinov. The original image and animation may be found at <http://visual-science.com/projects/zika/3d-model/>*





**I**n February 2016, scientists around the globe turned their attention toward a virus—Zika—that had suddenly morphed from a minor nuisance into something far more sinister.

They were responding to the World Health Organization's unexpected declaration that Zika had become a Public Health Emergency of International Concern because of its link with microcephaly—a devastating birth defect characterized by abnormal brain development and shrunken head size—as well as other neurological disorders including the progressive paralysis of Guillain-Barré syndrome.

**T**o make matters worse, the particular strain of Zika responsible for those problems was sweeping across the Americas very quickly. In May 2015, Brazil confirmed that locally acquired Zika was circulating in the country; by August 2016, more than 50 additional countries had suffered their first outbreaks, with large numbers reporting spikes in microcephaly and Guillain-Barré. In the United States, meanwhile, clusters of locally transmitted Zika erupted in Miami, and infants with birth defects possibly tied to the virus began to appear.

The speed with which the epidemic spread, however, was matched by the rapidity of the response within the scientific community. To a large extent, says **Alessandro Vespignani, PhD**, a computational modeler at Northeastern University, that lightning-fast reaction reflects years of hard work and investment in computational modeling (including such initiatives as MIDAS—Models of Infectious Disease Agents, an NIH-funded network of computational modelers), as well as grim experience with the H1N1 pandemic and the West African Ebola epidemic. A growing willingness to share results as quickly as possible through channels like bioRxiv ([biorxiv.org](http://biorxiv.org)), a website where scientists post papers before they have been published in peer-reviewed journals, also meant that modelers and researchers could more

speedily adapt their computational tools as new information became available.

## First Approximations

Alex Perkins, PhD, a researcher at the University of Notre Dame who studies the dynamics of infectious disease transmission and control, was one of the first modelers to swing into action. His goal: to try to quickly predict

for exploring all of those questions.

Unlike diseases such as Ebola or influenza, which are passed directly from person to person, Zika is a vector-borne disease that is primarily transmitted through the bite of the *Aedes aegypti* mosquito, an insect that thrives in the tropics and whose range and numbers are closely determined by climate. It can also be transmitted sexually, and by another

acquire the virus, then transmit it to uninfected people). But it does provide an opportunity to incorporate finely grained demographic and climate data, including details such as local population and birthrate; average daily temperatures, which govern where and how long the mosquitoes can live and, therefore, how many people they can infect; and even income levels.

---

**Perkins' model, which he first described in a paper posted to bioRxiv less than 2 weeks after the WHO declared a public health emergency, predicts that 93.4 million people, including 1.65 million childbearing women, could be infected before the first wave of the epidemic comes to an end.**

---

the course of the Zika epidemic and the likely number of victims both at home and abroad—information that governments and public health officials could use to plan interventions.

Perkins had for some time been contemplating the problem of integrating disease data collected at different scales. On the one hand, richly detailed local data related to such things as population and climate helps researchers understand factors affecting disease transmission. On the other hand, case reports—i.e., the number of suspected and confirmed cases tallied by hospitals—tend to be collected at the state- or country-wide level.

What, Perkins wondered, might be accomplished if the first bucket of data could be related to the second? And how could scientists predict the course of an epidemic before significant amounts of case data had begun to accumulate? Could they perhaps anticipate the total number of people who might be affected in particular geographic locales while early interventions could still have the greatest possible impact?

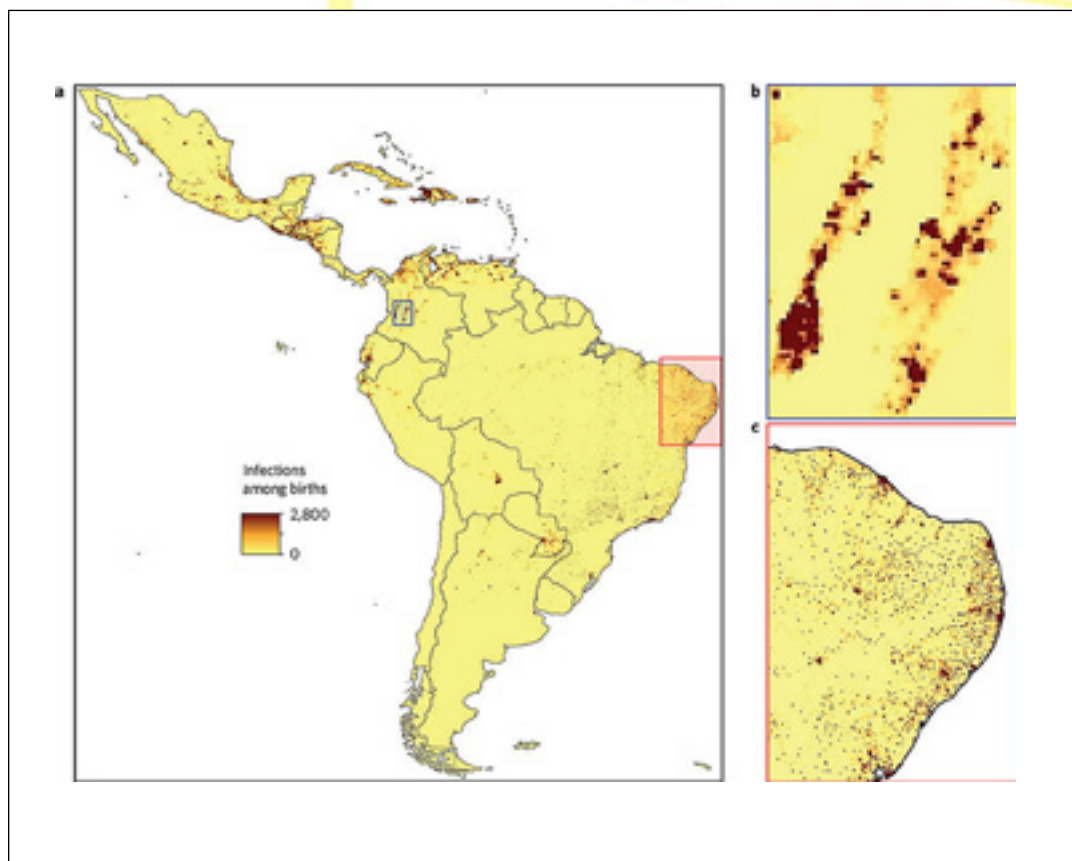
Zika offered both urgent and fertile ground

species of mosquito called *Ae. albopictus*.

This makes Zika difficult to model, since there is more than one infected species to deal with and the chain of transmission is complicated (uninfected mosquitoes bite infected people,

(Affluent people enjoy air conditioning and window screens, which reduce exposure to mosquitoes. Poor people do not, and are therefore at greater risk.)

In addition, while it was first identified more than 60 years ago, Zika remained



Gridded spatial projections of median infections among childbearing women at 5x5 km resolution across Latin America and the Caribbean as a whole (a) and in two specific areas: Cali, Colombia (b) and Recife, Brazil (c). Each grid cell is shaded according to the median number of infections for that cell based on 1,000 simulations. Reprinted by permission from Macmillan Publishers Ltd: TA Perkins, AS Siraj, CW Ruktanonchai, MUG Kraemer, AJ Tatem, Model-based projections of Zika virus infections in childbearing women in the Americas, Nature Microbiology 1, Article 16126 (2016).



for most of that time a neglected tropical disease by reason of its mild symptoms (low fever, rash) and lack of known complications. Consequently, when the World Health Organization (WHO) declared a global public health emergency, the scientific community confronted a disease about which it knew remarkably little. Like many, Perkins therefore had to rely on data that had already been collected for other mosquito-borne diseases such as dengue and chikungunya.

Dengue is a member of the same family of viruses as Zika, and all three illnesses are transmitted by *Ae. aegypti*. So Perkins used some of the basic transmission parameters that had already been established for dengue, and looked to a series of previous chikungunya epidemics to get an idea of the total number of people who might be infected. He then used that information, along with mosquito distribution maps and the aforementioned local data, to build a model that could capture the intersection between where people lived, where the mosquitoes were likely to be, and where the conditions were most suitable for transmission—a model that could project the number of infections among the general population, and among pregnant women in particular, across Latin America and the Caribbean, at a resolution of 5km by 5km.

Perkins is quick to note that his model is static rather than dynamic, and therefore estimates only how many people could become infected and not how long that might take. (A less geographically precise dynamic model developed by **Neil Ferguson, PhD**, at Imperial College London does provide such a timeline, predicting, for example, that the current epidemic will burn itself out within three years.) Moreover, it works best at the level of cities, but probably overestimates the total count at the country-wide level.

Even with those caveats in mind, the numbers are eye-popping: Perkins' model, which he first described in a paper posted to bioRxiv less than two weeks after the WHO declared a public health emergency, predicts that 93.4 million people, including 1.65 million childbearing women,

could be infected before the first wave of the epidemic comes to an end.

## **Adventures in the Fourth Dimension**

Perkins himself says that the dynamic model developed by Vespignani at Northeastern offers the best of both worlds: geographically specific estimates of how many people could be infected, and at what speed.

Ironically, when the NIH-funded Center for Inference and Dynamics of Infectious Diseases initially invited Vespignani, who has previously modeled Ebola and the H1N1 virus, to try his hand at Zika, his first reaction was an emphatic no. The reason: He didn't want to have to deal with the mosquitoes.

Vespignani and his colleagues simulate the spread of epidemics across time and space using the Global Epidemic and Mobility Model (GLEAM), a stochastic modeling platform that randomly moves simulated populations of individuals through a series of epidemiological states (susceptible, infected, recovered), generating ensembles of possible scenarios from which the most likely future path of an epidemic can be estimated. It even takes into account the way people travel from place to place, spreading disease as they go.

That already adds up to a lot of complexity, even for diseases that are transmitted directly between people. Throw in a couple of vectors like *Ae. aegypti* and *Ae. albopictus*, which can't travel very far (a typical mosquito only flies an average of 400 meters in its lifetime), and you suddenly have to simulate a whole new population of disease-bearing individuals and their movements at a very high level of detail—individuals whose range and lifespan depend heavily on temperature, and may therefore change drastically from season to season. (Mosquitoes die more quickly in winter than in summer, and if their lifespan drops below Zika's incubation period, they cannot transmit the virus at all.)

Vespignani initially assumed that achieving that level of detail in GLEAM would be impossible, and only changed his mind when he saw the rich

mosquito-related data that vector biologists at the Centers for Disease Control (CDC) and elsewhere had pulled together. "It was really a learning experience," he says, adding that having expanded GLEAM to accommodate one vector-borne disease, he and his collaborators should now be able to simulate others.

Vespignani and his team took into account many of the same factors (e.g., mosquito distribution, wealth) that Perkins' model used. Because GLEAM is able to simulate the course of an epidemic over time, however, Vespignani asked somewhat different questions: What would the timeline of the Zika outbreak look like from place to place? What would its impact look like at specific points in time? And when, exactly, did the virus first arrive in Brazil?

For a modeler, that last question is important, since the reliability of a model's projections depends on its ability to reconstruct the past. "You must get the past right in order to get the future

---

**GLEAM [the approach used by Vespignani's team] consistently predicted a slow-moving epidemic that would manifest in multiple waves in some places (Honduras, Mexico, Puerto Rico) due to seasonal effects.**

---

right," Vespignani says. He was therefore reassured when GLEAM determined that Zika was most likely introduced to Brazil in 2013, a finding that agreed with the results of phylogenetic and molecular clock analyses performed by **Oliver Pybus, PhD**, of Oxford University, and colleagues in Brazil.

Yet even with a well-calibrated model, forecasting the course of the epidemic was not straightforward. Like Perkins, for example, Vespignani had to cadge some of his transmission parameters from dengue, introducing a degree of uncertainty into his calculations. Because Zika is

passed from humans to mosquitoes and back again, there is also some fuzziness surrounding the serial interval, or the time between one infection and the next. And no one really knows how much of a role *Ae. albopictus* plays in spreading the disease. As a result, Vespignani performed several rounds of sensitivity analysis, essentially playing with small variations in

of infections across the U.S. on a state-by-state basis, a task that requires performing millions of simulations on 30,000 processors on a cloud computing platform. Their efforts generated headlines when GLEAM estimated that there could be 25 times the number of travel-related cases reported by the CDC, but Vespignani says that wasn't really

on the ground through diagnosis and treatment is another. Yet here, too, computation is playing an important role.

Standard diagnostic methods such as antibody detection aren't ideal for Zika because false positives can arise among people who have previously been infected by a related virus such as dengue. But the cost and complexity

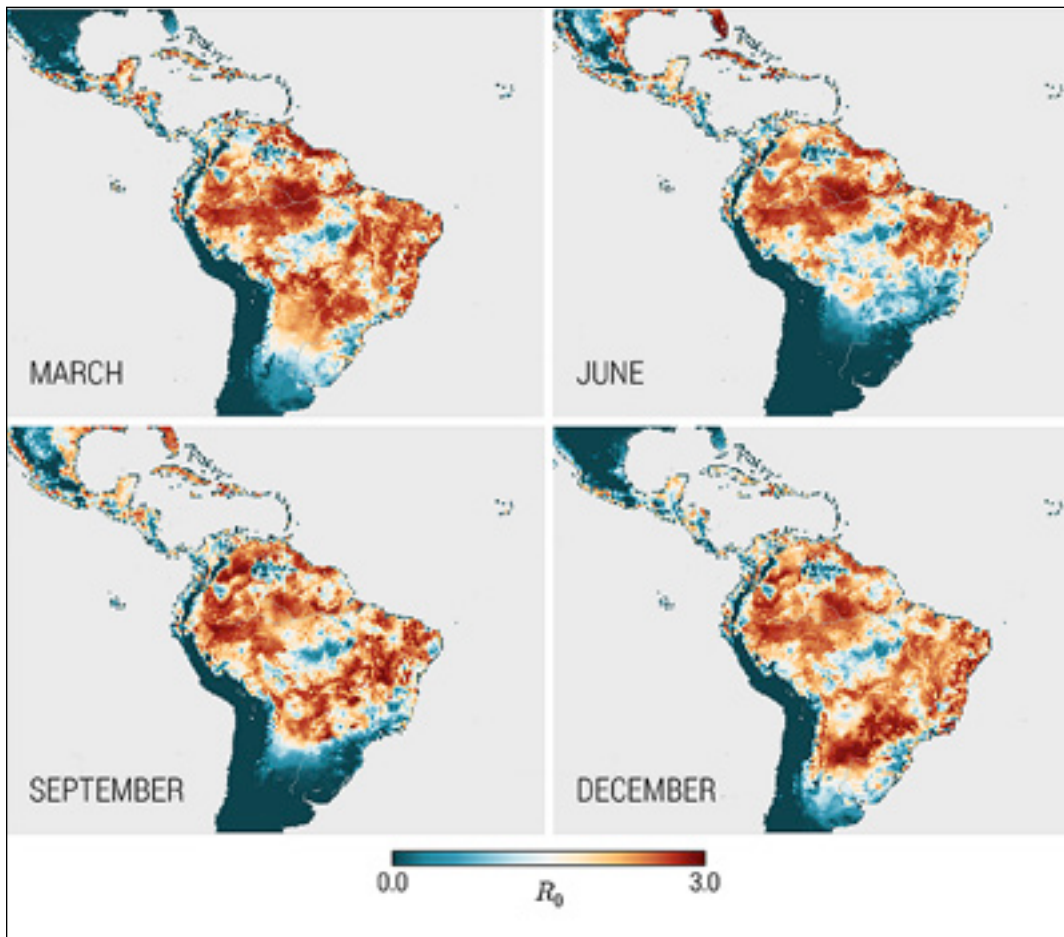
of more accurate methods such as DNA or RNA detection puts them beyond the reach of basic health clinics in poor, remote areas.

Now, however, a team of scientists assembled by **James J. Collins, PhD**, of MIT and Harvard's Wyss Institute, is changing that. Together, they have created a cheap, quick, and highly sensitive RNA test that could be used practically anywhere.

Originally developed to detect Ebola, the test relies on two pieces of technology: programmable RNA sensors called toehold switches that can be designed to detect virtually any RNA sequence; and a freeze-dried, paper-based platform that allows those toehold switches to be stored at room temperature on little paper discs, and activated simply by adding a bit of blood plasma and some water.

The switches are made of synthetic strands of RNA that encode a reporter protein that can make the paper change color from yellow to purple. But the switches also contain a hairpin structure called a stem that physically prevents the RNA from being translated unless the stem itself is unwound. In an ingenious twist, the switches are also configured to be perfectly complementary to specific target sequences of RNA. Only when the switches encounter their targets do their stems unwind, allowing the reporter protein to be produced and causing the paper to change color. Diagnosis: positive.

**Alexander Green, PhD**, who developed the switches as a postdoctoral fellow at the Wyss and is now on faculty



*Monthly seasonality for the time- and location-dependent basic reproduction number—the number of people that each infected individual is expected to infect. The Equatorial region presents less seasonality than the non-Equatorial regions, where the changes of the season have a strong impact over the temperature and consequently over the basic reproduction number. Reprinted from Q Zhang, K Sun, M Chinazzi, et al., Projected spread of Zika virus in the Americas, <http://dx.doi.org/10.1101/066456>.*

parameters—changing the serial interval, for example, or removing *Ae. albopictus* from the picture altogether—to see if the model would break down. (The various scenarios can be seen at [zika-model.org](http://zika-model.org).)

It didn't. Instead, GLEAM consistently predicted a slow-moving epidemic that would manifest in multiple waves in some places (Honduras, Mexico, Puerto Rico) due to seasonal effects.

Vespignani and his team are currently projecting the total possible number

surprising: Only 20 percent of infected people show symptoms, and those tend to be so mild that Vespignani himself doubts that he would go to the hospital if he had them. More reassuringly, the model predicts that this country will only see relatively small outbreaks of the sort that have already occurred in Florida.

### **A Quick and Easy Test**

Anticipating the course of an epidemic is one thing; dealing with it



at Arizona State University, explains that computation is involved at several levels.

For one thing, in order for the sensors to detect the minute quantities of Zika

viruses such as dengue or to human RNA. With a list of candidate target sequences in hand, they then simulated every

toehold switch that could conceivably bind to those potential targets, and evaluated which combinations of primer and switch would work best.

It took less than a day to construct and test the computationally optimized switches, which were sensitive enough to detect Zika in blood plasma samples and specific enough not to be fooled by dengue. And manufacturing a disk of freeze-

dried paper loaded with switches and amplification materials costs only a dollar.

Greene and his colleagues hope to make the test even quicker and less expensive. They also plan to validate

their system using human samples, and to extend its range so that it can detect other pathogens as well.

## Mapping the Mechanisms of Disease

Of course, once you've diagnosed a disease, the next step is treating it. Which is why researchers are also trying to understand exactly how Zika causes microcephaly and other neurological disorders, and are working to find drugs that can fight it.

**Yi Ren, PhD**, a cell biologist at Florida State University who studies inflammation, is one of those trying to get a handle on how Zika does its damage. Her FSU colleague **Hengli Tang, PhD**, was among the first to explain how Zika could cause microcephaly in fetuses—namely, by

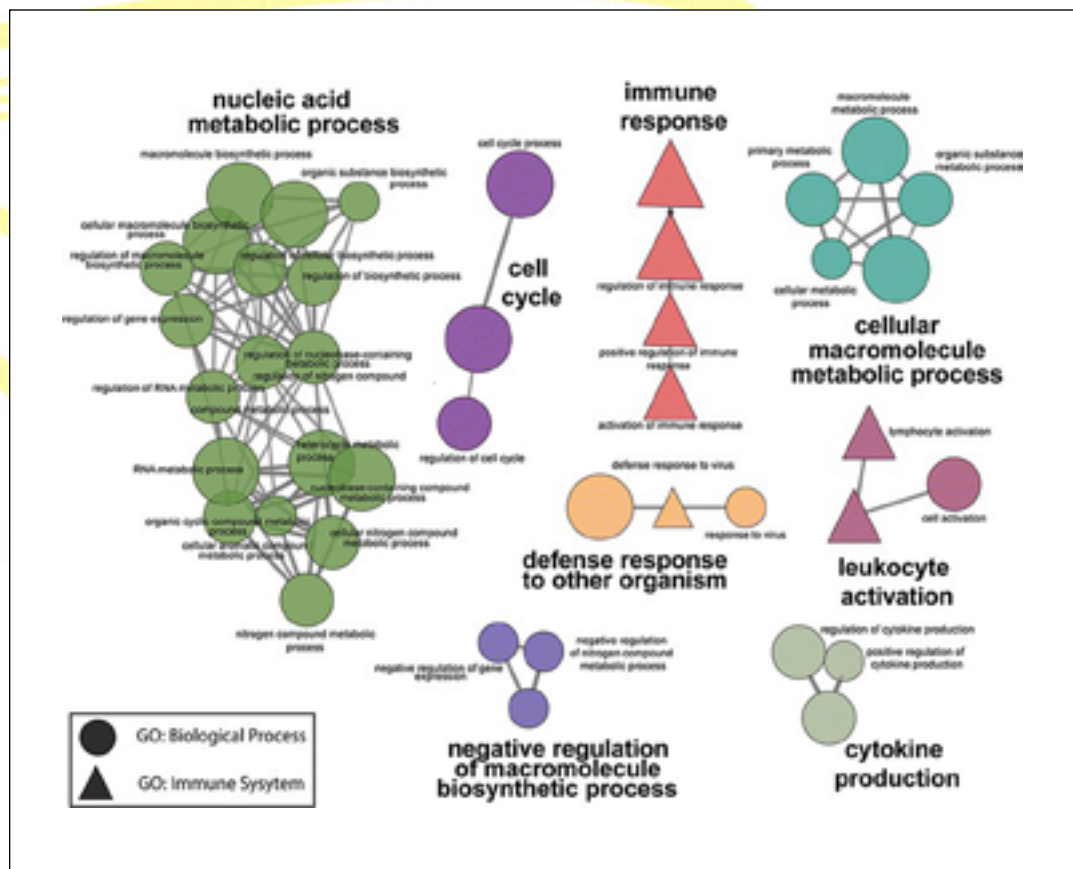
## A team of scientists assembled by James J. Collins, PhD, of MIT and Harvard's Wyss Institute, ... have created a cheap, quick, and highly sensitive RNA test [of Zika infection] that could be used practically anywhere.

RNA present in the blood of an infected person, those target sequences must first be amplified. Yet amplification itself makes use of short nucleotide sequences called primers, and if those aren't chosen wisely, trouble may ensue. If the primers aren't specific enough, for instance, they may also amplify other, similar sequences, like those belonging to dengue.

For another, not all RNA sequences are equally well-suited to detection by toehold switches. When a switch meets its target, the two strands of RNA intertwine, their bases binding to one another; and that interaction can interfere with the performance of the switch itself. Not all switches are equally sturdy, either; and if the stem is too weak, it might unwind even in the absence of target RNA.

Green and his colleagues therefore used several different algorithms and software tools—some custom-built, others open-source—to rationalize both primer selection and switch design.

First, they used their toolkit to screen the Zika genome for regions that were compatible with RNA amplification, filtering out those that were too similar to closely related



When human neural stem cells are infected with the Zika virus, certain genes are over- and under-expressed compared with normal cells. This visual map shows the various networks of biological processes and immune system responses related to those over- and under-expressed genes. These networks were generated using the Gene Ontology (GO) database and the open-source software platform Cytoscape. Circles represent biological processes (e.g., metabolic processes) in the Gene Ontology database, while triangles represent immune system responses. Groupings with less than three connections were excluded from the final list of networks. Reprinted from AJ Rolfe, DB Bosco, J Wang et al., *Bioinformatic analysis reveals the expression of unique transcriptomic signatures in Zika virus infected human neural stem cells*, *Cell & Bioscience* 6:42 (2016).

disrupting cell division and causing death among the neural progenitor cells that give rise to the various components of the nervous system—and Ren, in turn, wondered what inflammatory pathways the virus might activate.

**Alyssa Rolfe**, a PhD student in Ren's lab, explains that she and her colleagues used a variety of bioinformatic tools to analyze the RNA sequence data from Tang's Zika-infected human neural progenitor cells (hNPCs) in order to learn more about how the virus does its dirty work—and to suggest potential strategies for thwarting it.

After assembling a list of all of the genes that were either over-expressed or under-expressed in Tang's Zika-infected cells, the team used the Gene Ontology database to figure out which basic cellular functions those differentially expressed genes might be affecting. They also compared their list with the genes associated with six different neurological diseases in MalaCards, a searchable database of human diseases and disorders; and used an open-source software platform called Cytoscape to create a visual map of all the networks of intracellular biological processes and immune system responses associated with the up- and down-regulated genes. They even compared the gene expression profile of the Zika-infected cells to the profile of hNPCs that were infected with cytomegalovirus (CMV), which can cause a battery of birth defects including microcephaly.

The results were intriguing and, at times, unexpected. For example, the MalaCards search indicated that the pattern of gene expression in the Zika-infected cells had more in common with a suite of congenital nervous system disorders than it did with Guillain-Barré syndrome. And there was little correlation between the immune response pathways that were up- or down-regulated in the Zika-infected cells and their CMV-infected counterparts, suggesting that while the two viruses can cause comparable birth defects, they do so through different mechanisms. Moreover, four of the eight networks the team identified through

visual mapping were associated with immune responses—a surprise, says Rolfe, since one wouldn't expect hNPCs to have any significant interaction with the immune system at all.

The real shocker, however, came when the team dug deeper into the immune and inflammatory pathways associated with their list of genes. Rolfe and her colleagues discovered that a number of genes that one would only expect to see expressed in various kinds of immune cells, such as T-cells and dendritic cells, were in fact over- and under-expressed in the infected neural progenitor cells. “You wouldn't think those genes would have any function in hNPCs,” Rolfe says.

It's possible, she explains, that Zika is either pushing those cells to differentiate into some unknown state; or that the virus is somehow encouraging hNPCs, which do have an innate capacity to modify or regulate immune functions—producing proteins called cytokines, for instance, that normally promote healthy neural development—to shift from an anti-inflammatory role to a pro-inflammatory one. The second possibility, in particular, raises the question of what that shift might do to a developing fetus, and whether moderating the resulting inflammation might limit the negative consequences of infection.

Rolfe says that further investigation in a wet lab will be necessary to sort all of that out. But she hopes that the bioinformatic analysis she and her colleagues have already done will give other researchers useful clues for mitigating Zika's impact.

### **Going Viral on the Grid**

While Rolfe and the rest of Ren's team are probing for insights that could lead to fresh strategies for fighting Zika and its terrible effects, the researchers behind OpenZika ([openzika.ufg.br](http://openzika.ufg.br)) are using computation to virtually screen millions of existing compounds for ones that might already do the trick. The idea, explains **Joel S. Freundlich, PhD**, a chemist at Rutgers New Jersey Medical School who is collaborating on the project, is to jumpstart drug discovery by computationally whittling down the massive list of possible drug candidates

to a more manageable set of likely prospects that can be tested in the lab.

Given the numbers involved, that winnowing process is important.

**Alexander L. Perryman, PhD**, a co-principal investigator on the project who works as a senior researcher in Freundlich's lab, points out that even using high-throughput methods, most laboratories can only screen a couple thousand to a few hundred thousand compounds at a go, with Big Pharma pumping that number up to “a couple of million.” The OpenZika team, on the other hand, is screening 8000 FDA- and EU-approved drugs and NIH drug candidates, plus another 6 million compounds pooled from various sources to see if any are likely to disable or kill the Zika virus, with an additional 38 million compounds waiting in the wings.

OpenZika performs virtual experiments known as docking calculations that predict how small, drug-like molecules will bind and interact with the proteins that scientists suspect allow Zika to infect its victims and replicate inside them. And it does so on IBM's World Community Grid (WCG), which draws its computational horsepower from more than 700,000 volunteers in 80 countries who donate processing time on their idle computers, smart phones, and tablets, creating what Perryman calls “one of the largest supercomputers on the planet.” (Perryman previously used WCG to drive computational drug discovery projects for malaria and HIV/AIDS.)

The team employs a program called AutoDock Vina to predict the interactions between the small molecules in its compound libraries and various Zika proteins, virtually “docking” flexible 3-D atomic-scale models of the former to the latter in hopes of identifying molecules that can inhibit the virus's ability to function.

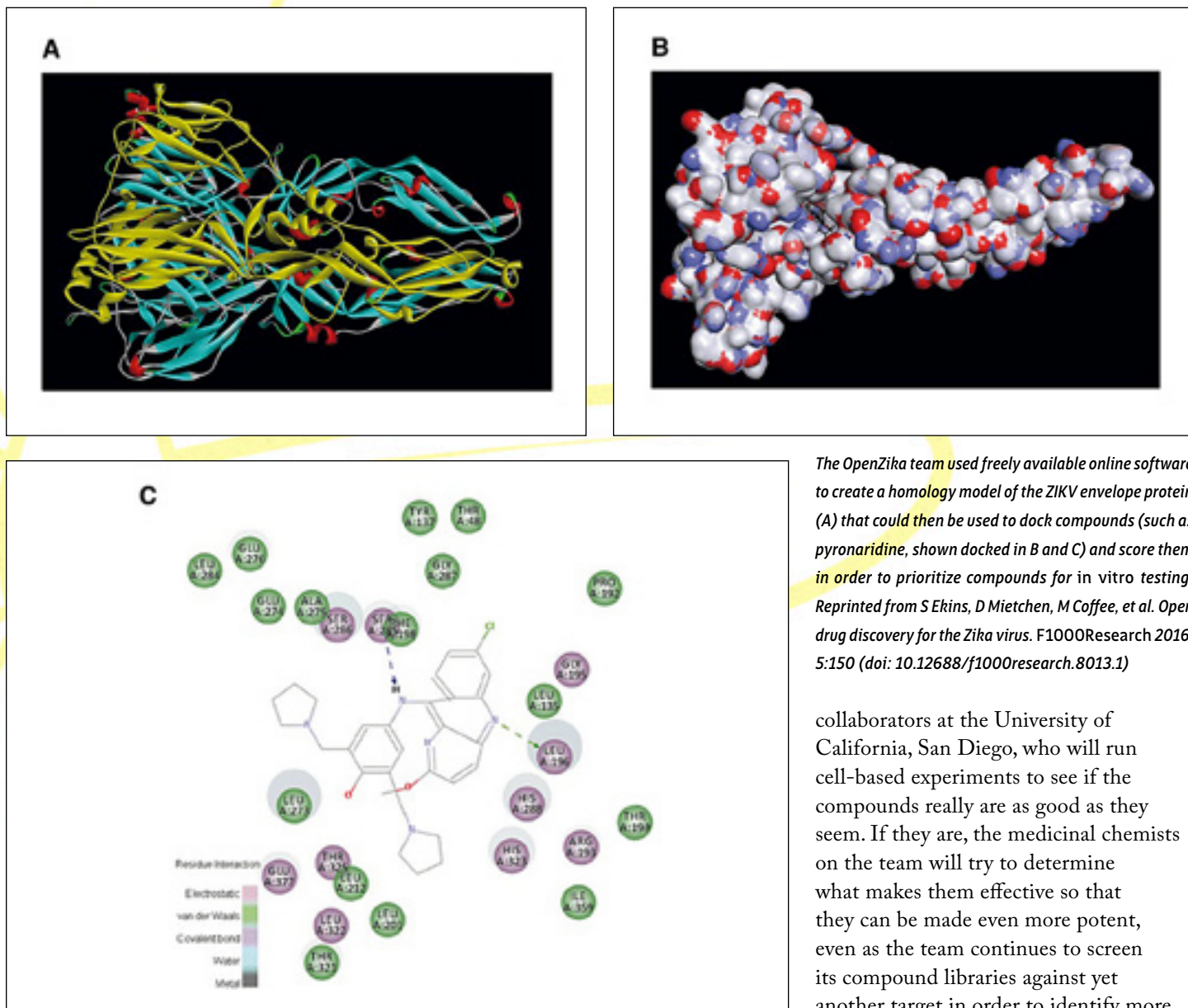
Each virtual experiment, or docking job, calculates the interactions between a single binding site on one protein, and one small molecule that is placed in a variety of positions, conformations, and orientations. That adds up to a lot of calculations. But WCG can handle it: Whereas most researchers



who use supercomputers measure their allotted processing time in thousands of CPU hours, the OpenZika team counts its share in thousands of CPU years. Within the first three months of the project, Perryman had submitted approximately 900 million docking jobs and received 439 million results.

viruses such as dengue and yellow fever. As scientists began to generate structures for the Zika proteins themselves, the OpenZika researchers incorporated those as well; but they continue to use data from related viruses in part because they hope to find broad-spectrum antivirals that will work against more than one.

example, AutoDock Vina thinned the herd to 160; Perryman trimmed it to 15; and Freundlich and Ekins used their chemical expertise to eliminate all but 8. (Eventually, they plan to use Bayesian machine-learning algorithms to do more of that filtering for them.) Of those, five will be sent to



*The OpenZika team used freely available online software to create a homology model of the ZIKV envelope protein (A) that could then be used to dock compounds (such as pyronaridine, shown docked in B and C) and score them in order to prioritize compounds for in vitro testing. Reprinted from S Ekins, D Mitchen, M Coffee, et al. Open drug discovery for the Zika virus. F1000Research 2016, 5:150 (doi: 10.12688/f1000research.8013.1)*

Because no one had bothered to determine the physical structure of the various components of the Zika virus before the current epidemic began, Perryman and his colleagues initially had to rely on speculative 3-D computational renderings, or homology models, of the Zika proteins that various team members created using the Zika genome and structural data gleaned from related

The software scores the performance of each compound, estimating the likelihood that it will stop Zika in its tracks. After that, the humans step in, visually inspecting the highest-scoring compounds to determine which might be the best drug candidates. Of the 8,000 drugs and drug candidates that have been screened against one particularly promising target, for

collaborators at the University of California, San Diego, who will run cell-based experiments to see if the compounds really are as good as they seem. If they are, the medicinal chemists on the team will try to determine what makes them effective so that they can be made even more potent, even as the team continues to screen its compound libraries against yet another target in order to identify more prospects for lab testing. The goal is to get the most promising candidates into the lab and back out the door in enhanced form as quickly as possible.

Forecasting. Diagnosis. Basic biology and drug discovery. All have a role to play in dealing with what is shaping up to be one of the greatest global public health crises in recent times. And computation, in turn, is playing a key role in all of them. □

Stanford University  
318 Campus Drive  
Clark Center Room W352  
Stanford, CA 94305-5444

## SeeingScience

BY KATHARINE MILLER

# ANIMATING HYPOTHESES

**I**n addition to illustrating complex biological molecules, animations can sometimes offer insight into how those molecules function.

That's what happened when **Grant Jensen, PhD**, professor of biophysics and biology at Caltech, and **Yi-Wei Chang, PhD**, a research scientist at Caltech, decided to animate their model of the strongest known molecular motor—the bacterial type IVa pilus machine. This motor resides in the cell membranes of many bacteria, including several that cause human diseases such as

meningitis and gonorrhea. It extends and retracts a filament (the pilus) that pulls the bacteria forward. Jensen and Chang knew the components of the machine, but not the details of how it worked. So they used cryotomography to image the machine and assemble a pseudoatomic model.

When they enlisted **Janet Iwasa, PhD**, research assistant professor of biochemistry at the University of Utah, to animate the structure, Jensen and Chang had a pretty good idea of how the machine worked and even storyboarded most of it. But the animation took them further. “It made us think about the details more carefully than we had,” Jensen says. In fact, the animation revealed that the cage at the base of the machine was too tight for pilin monomers to enter. “That led us to hypothesize that there must be a conformational change that occurs there when pilus assembly starts,” Jensen says.

“Molecular animations are not just entertaining visual candy,” Jensen says. “They are by far the fastest and clearest way to communicate complex hypotheses to a broad audience, and they force us all to think in even greater depth about what might be happening inside cells. Beyond pictures, animations are worth even more than a thousand words.” □

*In Iwasa's animation of the type IVa pilus machine, the ATP-powered assembly mechanism in the inner membrane causes the blue birdcage area to open up, allowing the entry of pilin subunits. The hypothesis is that the protein shown in yellow here binds the subunits and rotates, adding them to the growing pilus as it extends. During retraction, a different ATP-ase steps in and the process reverses itself. The animation is posted at <http://jensenlab.caltech.edu/movies/>. Reprinted from YW Chang, LA Rettberg, A Treuner-Lange, J Iwasa, L Søgaard-Andersen, GJ Jensen, Architecture of the type IVa pilus machine, Science 351:6278 (2016) with permission from AAAS.*

