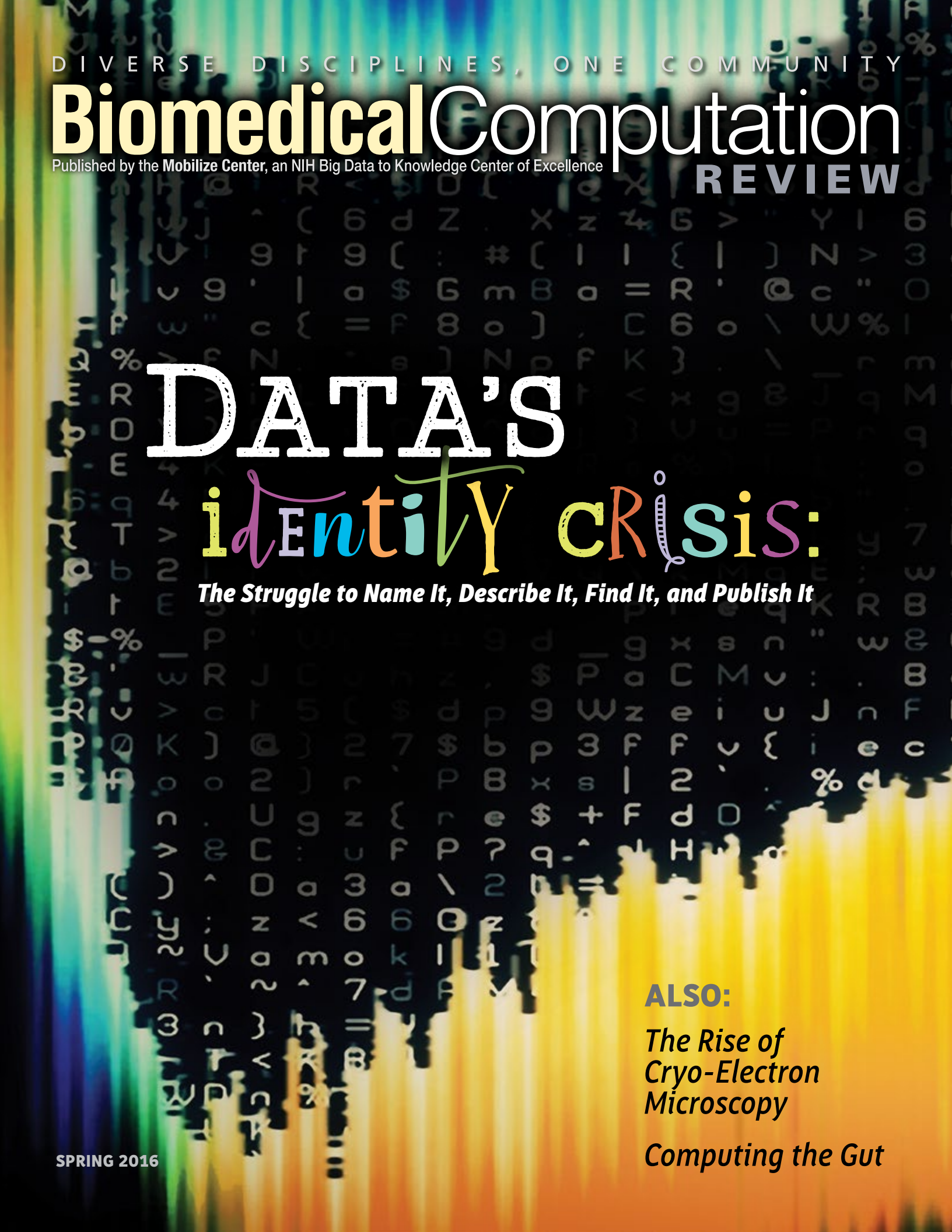
DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

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



DATA'S identity crisis:

The Struggle to Name It, Describe It, Find It, and Publish It

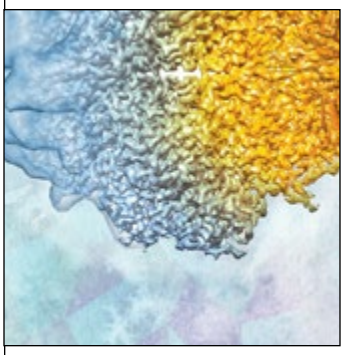
ALSO:

The Rise of Cryo-Electron Microscopy

Computing the Gut

SPRING 2016

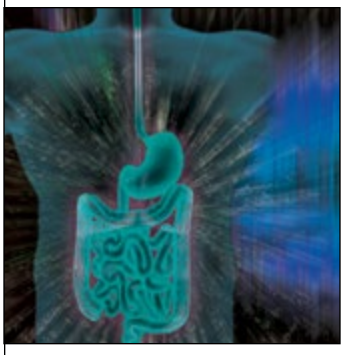
ContentsSpring2016



13
The Rise of Cryo-Electron Microscopy
BY ALEXANDER GELFAND

22
Data's Identity Crisis:
The Struggle to Name It, Describe It, Find It, and Publish It

BY KATHERINE MILLER



28
Computing the Gut
BY ESTHER LANDHUIS

Cover and pages 22-27 Art: Digital background © Alunablue | Dreamstime.com.

Page 28 Art: Created by Rachel Jones of Wink Design Studio using human body image © Dimdimich | Dreamstime.com, and digital background © Alunablue | Dreamstime.com.

DEPARTMENTS

Spring 2016
Volume 12, Issue 1
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
Stephanie Feudjio Feupe, Xiaoqian Jiang, PhD, Lucila Ohno-Machado, PhD, Shuang Wang, PhD

Layout and Design
Wink Design Studio

Printing
AMP Printing

For general inquiries, subscriptions, or letters to the editor, visit our website at 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

mobilize
Center for Mobility Data Integration to Insight

BY LUCILA OHNO-MACHADO, MD, PhD
Professor of Medicine and Associate Dean for Informatics,
University of California San Diego and principal investigator
for bioCADDIE, the BD2K Data Discovery Index Consortium

A TURNING POINT FOR (DATA) SCIENCE

Why open science is essential for scientific progress.

The cover story of this issue of *Biomedical Computation Review* is titled "Data's Identity Crisis"—with good reason. As vast stores of biomedical data are being created on a daily basis, our ability to make thorough use of them is stymied by our failure to share. The result: Scientific progress is radically slower than it needs to be. This seems to me to fit the Webster's dictionary definition of a crisis as "a difficult or dangerous situation that needs serious attention."

Fortunately, there is a solution to this crisis: open science.

Academic researchers take for granted that discoveries need to be published to achieve maximal impact and may be surprised by so much talk about open science. They may not realize that what is at stake in the open science/open data discussion is not whether results should be made public, but whether and how the data and analytical tools that led to those results should be available.

Traditional scientific publishing models were created when providing access to data and software was not possible due the constraints of print media. As we evolved into an era where data and software can be made available online, the belated discussion focuses on how to share them and who has rights and responsibilities to do so.

Provided individual privacy is protected, opening data for further analyses beyond an original study is about diversifying approaches to extract knowledge. It is *not* about witch hunting to destroy the work of those who previously analyzed the data. It is *not* about taking advantage of someone else's work without acknowledgements. It is *not* about removing incentives for good science.

On the contrary, open science is about *reproducibility* so there is no wasted time pursuing approaches that are flawed. It is about eliciting new ideas to reuse data that were collected with funding from taxpayers. It is ultimately about accelerating findings in a time frame that may make a difference for those who are suffering. People don't care *who* discovers a cure for cancer; they just want *someone* to discover it. Soon.

So let's not waste time creating chasms that pit biomedical and behavioral researchers against data scientists. Data scientists are not science "parasites" who use other people's data without attribution and without sufficient knowledge. Medical researchers are not data hoarders who want exclusive rights to discoveries. We all want science to translate into better health for everyone on the planet. Let's focus on what needs to happen to create an environment that promotes rapid discoveries that make a true difference.

People don't care *who* discovers a cure for cancer; they just want *someone* to discover it. Soon.

To achieve the open science ideal, there's a lot of meticulous, time-consuming work that must be done. For example, for datasets to be reused properly, they need to be clearly identified, posted in searchable repositories, and (perhaps most importantly for re-usability and reproducibility) contain descriptions or annotations (so-called metadata) that allow users to understand the context in which data were collected and pre-processed, their potential limitations, and how they can be accessed.

The move to open science is an exciting turning point for scientists everywhere, as it will allow data to be used in many more ways than what we have traditionally envisioned. And plenty of people are already on board: Many scientists, coming from different backgrounds and different specialties, emphasize the importance of maximizing the use of data through systematic annotation and organization. The whole community must unite to help design the ecosystem for data sharing in a way that moves us beyond the ideas of a few researchers and accelerates meaningful biomedical discoveries. □

BY KATHARINE MILLER

WOMEN IN DATA SCIENCE CONFERENCE

What happens when hundreds of talented female data scientists gather in the same place?

In November 2015, the Mobilize Center co-hosted the first Women in Data Science (WiDS) Conference along with Walmart Labs, Stanford University's Institute for Computational & Mathematical Engineering (ICME), and several other Stanford entities, including the department of statistics, the engineering department's computer forum, and the Office of the President.

More than 400 people attended the one-day conference, which was aimed at inspiring, educating, and supporting women in data science—from those just starting out to those who are established leaders across industry, academia, government, and nongovernmental organizations.

And they were inspired.

"When you are surrounded by successful and talented women in a room full of support, encouragement and inspiration, your dreams and goals burst into a reality that pushes you to conquer it all," tweeted attendee Diana Riveros Mello during the conference.

Margot Gerritsen, PhD, director of ICME and a Mobilize Center faculty member, organized the WiDS conference because she recognizes the tremendous talent among women. "I see this every day when I am teaching," she says. "And it would be a real shame if that group of wonderful scientists is underutilized." Gerritsen wants more women to join the field. "It's important for society as a whole…to have a very diverse, inclusive team of people working on data science problems."

Data science involves extracting relevant information from voluminous, heterogeneous, and often messy data streams, and using that information to help inform decisions across all arenas, including research, government and business. "It's everywhere now," Gerritsen says.

The impressive roster of all-female WiDS conference speakers exemplified the field's breadth. About one-third of the speakers came from academia and two-thirds from industry, and the conference covered a diverse set of data science applications, from monitoring individuals with Parkinson's disease, to cancer genomics, cyber security and online marketplaces.

"It's important for society as a whole... to have a very diverse, inclusive team of people working on data science problems," Gerritsen says.

"Just seeing the array of possibilities makes me think, 'Yeah, I can do great things too,'" says **Shenglan Qiao**, a Stanford PhD candidate in physics who attended the conference.

In addition, panels on careers and entrepreneurship offered an opportunity for successful data scientists to reflect on their lives and offer advice to younger

continued on page 6

DETAILS

Video recordings of the November 2015 ***Women in Data Science Conference*** are available online under the 2015 Conference menu at widsconference.org. The next WiDS conference will be held on February 3, 2017.

BY KRISTIN SAINANI, PhD

ACCELERATING SEARCH OF COMPRESSED DATA

Researchers have found that biological data are amenable to fast search.

As biological data grow at an exponential rate, computing power and storage can no longer keep pace. For example, genomic data are increasing by 10-fold per year whereas computing speed is doubling just every 18 months. "We are currently generating massive datasets—so massive that without smart algorithms, we can't effectively analyze them," says **Bonnie Berger, PhD**, professor of applied math and computer science at MIT.

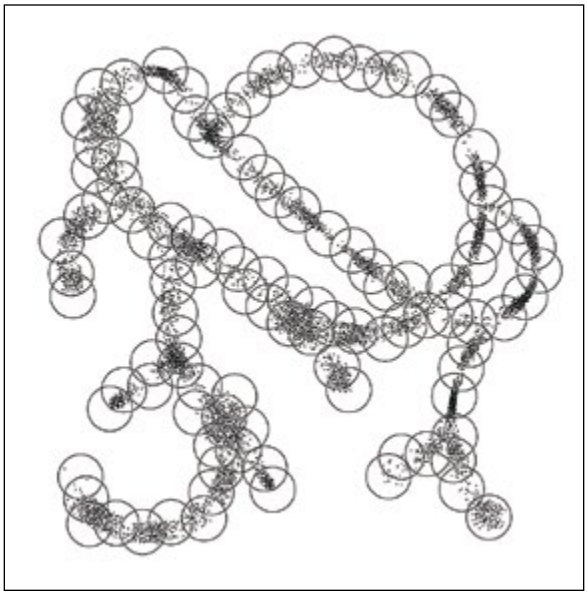
"This whole area of compressive algorithms is really taking off because it's absolutely necessary," Berger says.

Fortunately, biological data are highly redundant, which means that the amount of *novel* data entering these ballooning databases is growing at a much slower rate, Berger says. Her team—including graduate student **William Yu** and postdoc **Noah Daniels, PhD**—has figured out a way to exploit this redundancy to greatly speed up database search. Their framework, called entropy-scaling search, works directly on compressed data. It is described in a 2015 paper in *Cell Systems*.

"Clearly if you wanted to store massive amounts of data, you would compress it—which many people have done," Berger says. "But that isn't enough because eventually we have to look at the data. So we asked ourselves if we could design methods that could search the compressed data."

Sequence data are easily compressed because genomic sequences are much more similar than they are different. For example, two human genomes differ in only 0.1 percent of their nucleotides. To compress the data, Berger's team groups closely related sequences into clusters in a way that exploits the structure of biological data, and represents each cluster with just one sequence. To search the compressed data, they compare the query sequence only to the cluster representatives. This coarse search returns a few clusters that can then be searched more finely.

The search time depends on: (1) how many clusters have to be checked during the coarse search and (2) how many neighboring clusters have to be examined during the fine search. Berger's team showed that biological data has two key properties that limit these two quantities, respectively: low metric entropy and low fractal dimension.



"Octopus-Like" Data. *These data exhibit low metric entropy and low fractal dimension. Though the picture appears 2-D, the data-points traverse multi-dimensional space; circles are spheres that define clusters of data. Low metric entropy means that relatively few spheres are required to cover all the data, limiting the size of the coarse search. Fractal dimension describes the number of neighbors each sphere has: A fractal dimension of 1 means that neighboring clusters lie on a straight line; a fractal dimension of 2 means that each cluster is completely surrounded by neighboring clusters in 2 dimensions; and so on. In real data, this number is fractional—hence the name fractal dimension. Lower fractal dimension means fewer neighbors to search during the fine search. Image courtesy of YW Yu, NM Daniels, DC Danko and B Berger.*

Data with low metric entropy are highly redundant and thus can be represented by a small number of clusters. Take sequence data: Because only a tiny fraction of all possible genomic sequences actually exist in nature, relatively few clusters are required

continued on page 6

BY KATHARINE MILLER

SIMTK RELAUNCH: BACK AT THE CUTTING EDGE

What does it take to make a great biocomputing repository even better?

Way back in 2005, YouTube was brand new; FaceBook had just launched; the idea for Dropbox hadn't yet been hatched; and GitHub wasn't even a glimmer on the horizon. So when SimTK launched that year as a place for the physics-based simulation community to share files and code and control privacy at different access levels, it was offering something quite new and unique.

"SimTK was really cutting edge at that time in terms of enabling sharing and collaboration," says **Joy Ku**, **PhD**, SimTK project manager who also served as Director for Simbios, the National Center for Biomedical Computing that established SimTK. But after 10 years, the site is ready for an upgrade as well as some new functionality to put it back at the cutting edge, Ku says. "The Internet and scholarly publishing have changed dramatically, and we're reinventing SimTK to take advantage of new technologies—new ways of interacting with and using the Web—and new paradigms in research dissemination to help accelerate research."

"The Internet has changed dramatically, and we're reinventing SimTK to take advantage of new technologies—new ways of interacting with and using the Web—to help accelerate research," Ku says.

Some new features, such as social networking capabilities to enhance collaboration, are already up and running. Others, such as opportunities to reproduce simulations in the cloud, and integrated functionality with GitHub and other valuable online resources, are planned for the coming year. New project pages include more information to help determine whether or not a particular resource would be useful, such as when a project was last updated and how many times it has been downloaded. "We want to make it easier for people to make intelligent decisions about the resources they download and use," Ku says.

SimTK's Enduring Appeal

In the years since its launch, SimTK has grown considerably. It now has more than 47,000 members and hosts more than 800 projects—despite the fact that other services now offer some of SimTK's functionality.

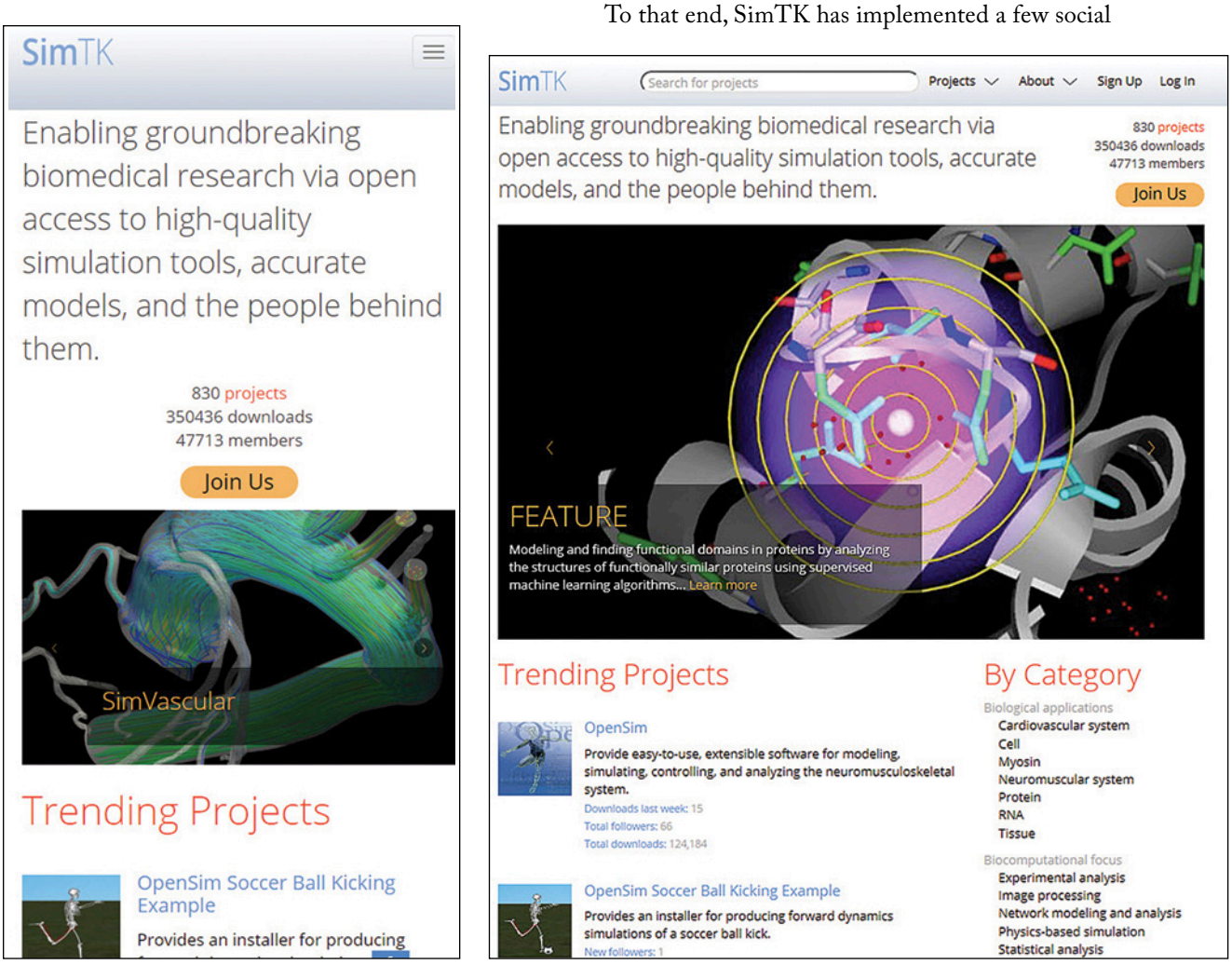
What explains the site's enduring appeal? For many, it's the resources posted for downloading—models, simulations and software—that represent years and years of work. "Being able to download musculoskeletal joint models, run them and analyze them decreases barriers to entry to the discipline because people can use the models rather than redeveloping them," says **Ahmet Erdemir**, **PhD**, director of the Computational Biomodeling Core at the Cleveland Clinic.

SimTK also offers some unique and difficult-to-generate datasets for download, including the one used for the Grand Knee Challenge, a modeling and simulation competition held each year for the past six years. The competition makes available to the community the most complete knee datasets anywhere on the Web. "In terms of human movement analysis, we collected and posted everything including the kitchen sink," says **B.J. Fregly**, **PhD**, professor of computational biomechanics at the University of Florida, who ran the challenge. Participants in the challenge used the SimTK public forum to post questions; and the site's mass emailing capability came in handy as well. "It's one-stop shopping," Fregly says. The knee data has been downloaded by 700 unique users. "People all over the world are using these data. It's not just for the competition."

SimTK also provides a free alternative to building a new research Web site for each of your projects. Erdemir, who has launched all his projects on the site since 2007, appreciates not having to create and maintain a centralized location for posting software, models and documentation. "SimTK offers a better chance of long-term sustainability. And it's more public-friendly, accessible and streamlined," he says. "The site also tracks downloads, which is useful in

grant applications to show the impact on the community," he says.

In addition, Erdemir uses SimTK for open development of models and software. For example, OpenKnee has been completely open from the outset.



The relaunched SimTK is available both for personal computer and mobile devices.

"We disseminate early and frequently, which requires time and effort, but SimTK provides that infrastructure," he says.

Another advantage: By posting their projects, graduating students can ensure that their work is reused. "I make my students put their source code, data and some readme files on SimTK before they graduate," Fregly says. "That way a future student can go in and pick up where they left off. It makes continuity of projects easier."

Social Networking with SimTK

Until now, SimTK has operated primarily as a hub-and-spoke system: A small subset of members contributed resources to the site and a large group of members downloaded them without interacting with each other. "We want to see it become a more interwoven network," Ku says.

To that end, SimTK has implemented a few social networking tools. For example, the "Follow" button, one of the most recently added features, allows members to track a project either privately or publicly. The latter option places the members' names on a list of followers that anyone can explore. "Now you're able to interact with anyone who expresses an interest in the project, not just the project leads and key developers," Ku says. "It's a way to build the community and promote collaboration."

The site also highlights recommendations of other projects to visit based on the project category, keywords, and the viewing behavior of other site

continued on page 6

continued from page 2: WiDS Conference

women. Most often, they encouraged taking risks and being flexible.

"Don't let your fear about your own abilities or fear of being an imposter have any bearing on the kinds of decisions you make," said **Jennifer Chayes**, **PhD**, distinguished scientist and managing director of Microsoft Research New England in Cambridge, Massachusetts. "Take that part of your brain and say thank you for sharing and just put it aside. If I'd listened to that part of my brain, I would have led a very boring life."

Interest in the conference was high: It sold out in less than 20 days with little promotion, and more than 6,000 individuals tuned in to the live-stream.

The vast majority of attendees hope to attend the next conference, which is scheduled for February 2017.

Gerritsen advises women who are interested in computational math or other scientific fields: "Jump in. It's a fabulous field with lots of opportunity." □

continued from page 3: Accelerating Search

to cover all the data.

During the fine search, the best-matching cluster and all neighboring clusters have to be searched. Fractal dimension describes the number of neighbors each cluster has. Lower fractal dimension means fewer neighbors to search. Fortunately, biological data have a fairly low fractal dimension because evolution tends to trace out relatively linear paths (see figure). "Clusters largely tend to extend along the branches of the tree rather than in all directions," Berger explains.

Berger's team showed that their compression and search framework is effective for any data that exhibit low metric entropy and low fractal dimension. Thus, potential applications extend way beyond sequence search. In their *Cell Systems* paper, Berger's team demonstrates orders of magnitude speed-ups for searching databases of chemical compounds, metagenomes, and protein structures.

PubChem is a comprehensive database of 60 million small molecules that can be used for tasks such as repositioning drugs. Until now, it was infeasible to perform even a one-molecule search of all of PubChem on a typical desktop computer. So Berger's team clustered the chemical compounds in PubChem based on the geometric similarity of chemical motifs, and then applied their two-step search process to these data. Compared with the commonly used search tool SMSD (Small Molecule Subgraph Detector), they were able to achieve a 150-fold speed up with 92 percent accuracy.

The team's framework can be wrapped around common search tools, such as SMSD for small molecule search, BLAST for DNA sequence search, and PSI-BLAST for protein sequence search. "The cool thing about all our tools is that they plug right into existing pipelines."

Berger's team has made their tools openly available at: http://cast.csail.mit.edu/, and other groups have begun building upon these tools, she says. "This whole area of compressive algorithms is really taking off because it's absolutely necessary." □

continued from page 5: SimTK Relaunched

visitors. This feature is already encouraging visitors to explore the other projects on SimTK: In the first year after it was implemented, total monthly project visits more than doubled (from 31,000 to 63,000) and 42 percent of project visits were made through the recommendation system. Ku hopes such features will also motivate visitors to host their own projects on the site.

Reproducibility in the Cloud

One major challenge of physics-based simulation is reproducibility—ensuring that researchers using the same data and software can get the same results.

To address this problem, Ku and Erdemir are working on offering members a cloud-based way to reproduce published results on SimTK. "We're hoping to further lower the barrier to entry for modeling and simulation," Erdemir says.

The feature enables users to launch a simulation by simply selecting a server, a model, and a specific software version from dropdown menus. When the results are available either for download or for browsing online, the user receives a notification. As a test case, Erdemir has created a template for running such simulations of OpenKnee. The interface allows users to run a simulation, perhaps apply a different load to the knee, and run a new simulation. Ku and Erdemir hope the cloud-based option will be up and running before the end of the year.

Plug and Play Capability

In the coming year, SimTK will also include the ability to plug-and-play with other online applications. For example, several SimTK projects use GitHub as a way to collaborate on their source code. They might also use another site to track bugs and then they use SimTK to share the software. But SimTK could be the hub that provides ways to pipe information to and from these multiple places, Ku says. Some of the developers of the site's largest projects on SimTK, such as OpenSim, OpenMM and SimVascular, are eager for this improvement. "Users and contributors alike will have one place to go to get quick updates; communicate; know where the project is headed," Ku says.

SimTK: Past, Present and Future

SimTK was novel in 2005 when it started out, Ku says, but 10 years on, "technology and our users' needs have changed, so SimTK is changing, too," she says. Ku hopes the features now being added to SimTK will put it back at the forefront—well ahead of whatever other new new things might come along—and keep it relevant for the community of researchers that flock to its pages. □

SNEAK PEEK

For a sneak peek of the new SimTK site, visit https://simtkalpha.stanford.edu.

BY ALEXANDER GELFAND

GOING VIRAL: MODELING EBOLA

In the midst of the Ebola epidemic, modelers tried to predict its spread, with some success. Now they're reflecting on the lessons learned.

The worst case scenarios were frightening: At the peak of the West African Ebola epidemic of 2014, estimates of the potential death toll ranged from several hundred thousand to more than a million. Experts believe that those dire predictions spurred the international response that helped limit the death toll to fewer than 12,000 to date.

In fact, the estimates themselves were the result of an international response by computational epidemiologists who swung into action as Ebola surged across Guinea, Liberia, and Sierra Leone. Even their simpler models showed that the initial Ebola outbreak wasn't winding down as expected. Other increasingly complex models predicted how the disease might spread, forecast the number of cases that could arise, and simulated possible interventions to help policy makers and public health officials respond effectively to the crisis.

In the end, the models tended to be most reliable when they were used to look two to three weeks ahead. Over longer time spans (e.g., two to three months), the accuracy of the models (with a few exceptions) declined precipitously. This was a good thing for humanity, as the huge death tolls the models foretold were thankfully averted—partly because of policymakers' responses to the modelers' nightmare scenarios, and partly because of gradual changes in the behavior of people on the ground in Africa.

Today, the same modelers are still sorting out exactly what brought the epidemic to a halt, in hopes of developing strategies for the future. In that sense, says **Martial Ndeffo, PhD**, a research scientist at Yale University, their initial predictive work was just a prelude. "Now we're trying to do retrospective analysis," he says. "Can we disentangle the contributions of the different interventions in reducing or averting cases? What kinds of behavior change were there, and can we quantify their contributions to curtailing the spread of the disease?"

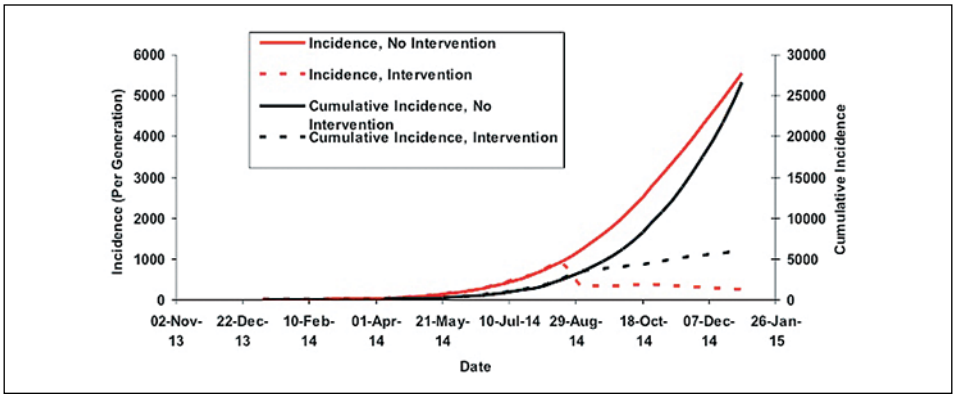
They are also reflecting on the lessons they learned from the epidemic—most notably, that it is difficult to build accurate, well-fed models when human behavior is a key parameter, the situation on the ground is messy, and information is hard to come by. "It's a challenge to model behavior," says **Madhav Marathe, PhD**, director of the Network Dynamics and Simulation Science Laboratory in the Biocomplexity Institute at Virginia Tech. "It's an even bigger challenge to model it in a place where it is very hard to get data."

Ebola's Surprising Virulence

Epidemiologist **David Fisman, MD, MPH**, at the University of Toronto, initially assumed that the first outbreak in Guinea would swiftly peter out, as Ebola incidents have in the past. When it didn't, he employed a mathematical model that he first developed for SARS to illustrate just how quickly the disease appeared to be spreading.

Fisman's model, called IDEA (for Incidence Decay and Exponential

"It's a challenge to model behavior," Marathe says. "It's an even bigger challenge to model it in a place where it is very hard to get data."



Fisman and his colleagues plotted the incidence of Ebola cases (both model-projected incidence per 15-day generation [solid red curve, scale on left y-axis] and cumulative incidence [solid black curve, scale on right y-axis]) against time (x-axis). Unless the disease's decay rate were to increase (as a result of intervention), the model predicted that the outbreak would cause more than 25,000 infections by the end of 2014, with the peak occurring in April 2015 and spread continuing until mid-2016, with more than 140,000 cases. However, if intervention in September 2014 resulted in an increase in the disease's decay rate by just 0.005, incidence and cumulative incidence would drop off significantly as shown by the dashed red and black curves. Reprinted from D Fisman, E Khoo, A Tuite, Early Epidemic Dynamics of the West African 2014 Ebola Outbreak: Estimates Derived with a Simple Two-Parameter Model, PLoS Current Outbreaks, *Sept 8 2014.*

Adjustment), was inspired by financial models that use a so-called discount factor to compensate for the decline in value of money over time. A similar discount factor can be used to predict the course of epidemics, which typically show rapid initial growth followed by similarly rapid decline. When Fisman fit IDEA to the case counts coming out of West Africa, however, he saw what looked like exponential growth tempered by a disturbingly tiny discount factor.

The good news was that Fisman's analysis, which appeared in *PLoS Current Outbreaks* in September 2014, indicated that even a small increase in the discount factor could save tens of thousands of lives. The bad news was that because the model was so simple, it couldn't reveal much about what was driving the epidemic—or what steps might stop it.

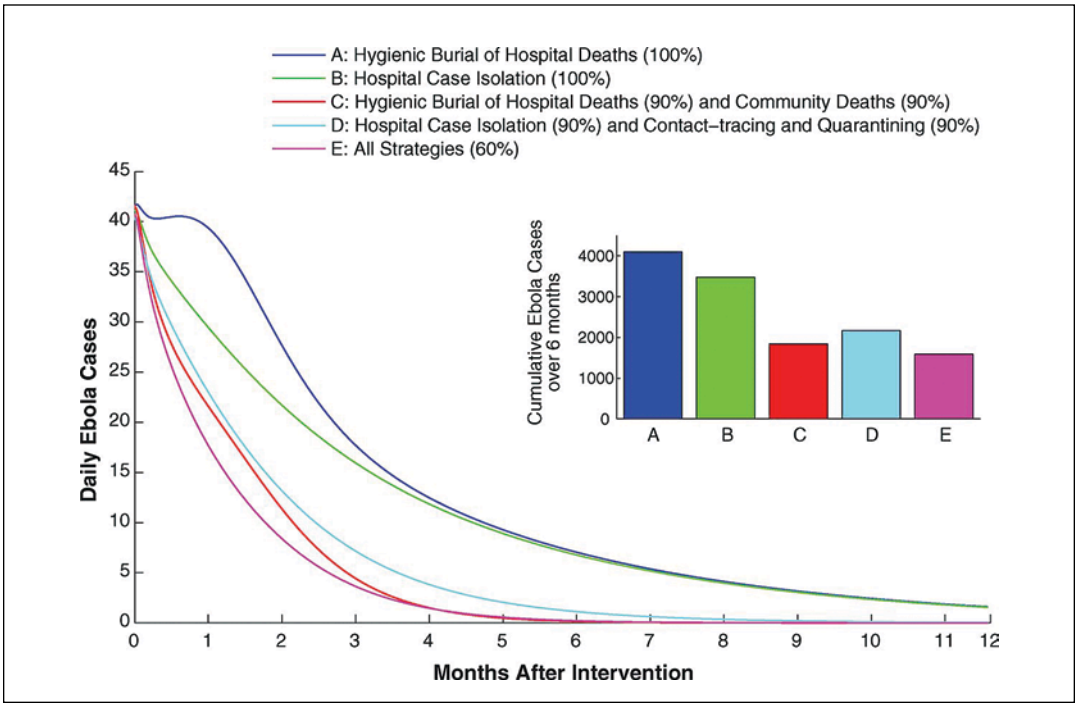
Other researchers, however, were using more complex models to generate those kinds of insights.

Gauging Interventions

At Yale University, a group led by **Alison Galvani, PhD**, director of the Center for Infectious Disease Modeling and Analysis (CIDMA), focused on producing forecasts that could provide timely guidance to international policy makers and local authorities. But they did not have much data to work with, says Ndeffo, who co-authored several papers describing the team's efforts. So they began with a model based on differential equations that could be run relatively quickly with limited parameters.

The model allowed individuals to move from one epidemiological category (e.g., susceptible, exposed, infectious, recovered) to another as they interacted in various settings, including hospitals and funerals—both hotspots for Ebola transmission. According to **Abhishek Pandey, PhD**, postdoctoral associate in epidemiology at Yale University, the model also used probabilistic methods to mimic the uncertainty in the numbers of people who might move from one category or setting to another.

The team used this stochastic model to gauge the potential impact of various interventions, from building more treatment units to distributing personal protective kits and performing sanitary burials. The takeaway was clear: combining more than one response would be far more helpful than just focusing on any single one, no matter how effective it was. (Using a different model that drew on previous efforts to simulate HIV and influenza transmission, Ndeffo and **Dan Yamin, PhD**, also predicted that treating the most severely ill patients within five days of showing symptoms would do the most to halt the spread of the disease—a result that jibed well with what health workers were finding on the ground.) But individual findings were sometimes counterintuitive, like the prediction that distributing protective gear would only have a significant effect once treatment centers were already at capacity. "That was a bit surprising," says Ndeffo, adding that it was "the kind of insight you could only get from a model."



In a paper published in Science *in October 2014, Ndeffo and Galvani compared the effectiveness of various intervention strategies used alone or in combination (graph), and calculated the daily number of new and cumulative cases after 6 months (bar chart) if the following interventions were to be implemented alone or in combination: sanitary burial of hospital deaths, sanitary burial of community deaths, case isolation of hospitalized patients, contact-tracing in the community, and quarantine of infected contacts. The model predicted that use of all strategies in combination would be most effective. Reprinted with permission from from A Pandey, KE Atkins, J Medlock, et al., Strategies for containing Ebola in West Africa,* Science *346(6212): 991-995 (Oct 30 2014).*

Networked Epidemiology

Marathe and his colleagues also began modeling the initial outbreak using differential equations. By October 2014, however, they had shifted gears and developed a multiscale agent-based model to predict the course of what had become a full-blown epidemic.

Building such a model involved what Marathe calls "networked epidemiology." First, Marathe and his team used census data to create synthetic populations that were statistically equivalent to the actual populations of the affected West African

countries. They then linked the individuals in those populations through virtual social networks, allowing them to interact through work, school, and household activities, and to mingle at home, in hospitals, and at funerals. And they used probabilistic methods to inject a realistic element of chance into nearly every aspect of their simulations, from how individuals moved about to how the disease itself progressed (incubation period, time to death, etc.).

The researchers used this stochastic agent-based model to create risk-profiles for other countries in West Africa that might be hit next; to determine which interventions (better contact tracing, new drug therapies) might have the greatest effect; and even to gauge what might happen if Ebola spread to the United States. Like the Yale group, the Virginia Tech team found that a combination of responses worked best. But they also found that even a successful drug intervention wouldn't do much to curb the epidemic.

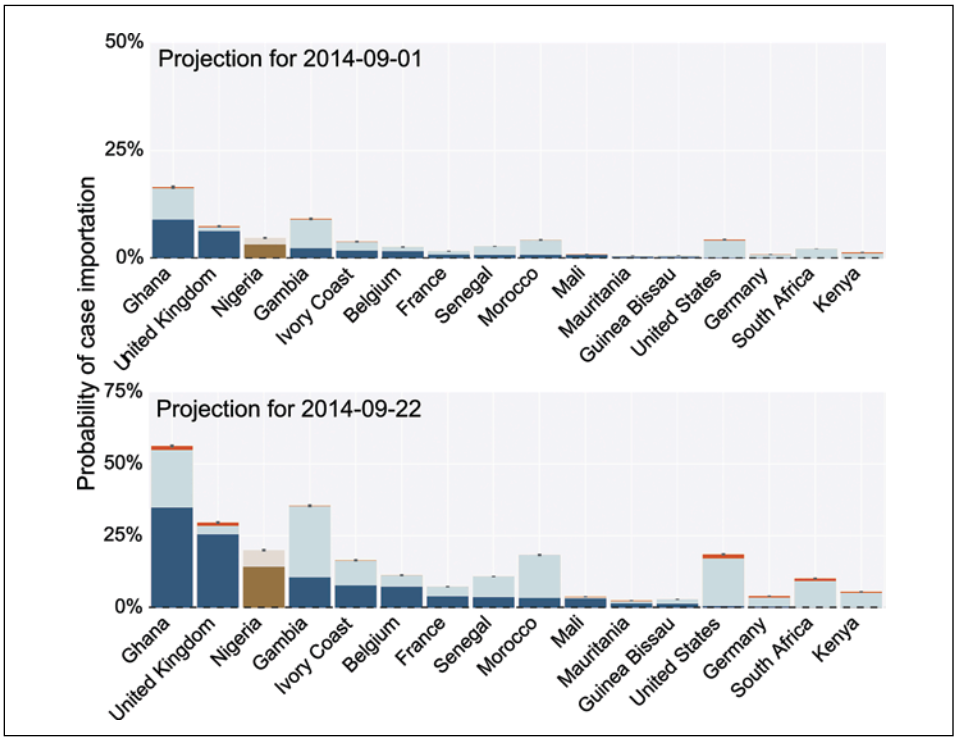
Global Predictions

Marathe's friend **Alessandro Vespignani, PhD**, director of the Laboratory for the Modeling of Biological and Socio-Technical Systems (MOBS-Lab) at Northeastern University, also took a network-based approach to modeling the epidemic. But in his case, Vespignani used a multiscale modeling platform called GLEaM (Global Epidemic and Mobility Model) to predict how Ebola might spread across the globe.

Together with an international team of collaborators, Vespignani integrated an agent-based model of Ebola transmission with two other models: a global population model spanning 220 countries and thousands of subpopulations distributed around the world, and a so-called mobility model that allowed those populations to mix through short-range commuting patterns and long-range air travel. "We talk about individuals and geography," Vespignani says. "But underlying all of that are large network models with people traveling from one point to another."

By simulating the number of passengers traveling daily on each airline connection in the world, Vespignani and his colleagues were able to rank which countries were most at risk of importing Ebola. Their findings were eerily accurate: two of the top three at-risk countries—the United Kingdom and Nigeria—saw cases in 2014.

scenarios prompted—and also to the fact that individual behavior in the affected countries changed markedly as the epidemic progressed, but in ways that modelers couldn't necessarily track. "Behavior change had a major impact on the curtailment of the epidemic," Ndeffo says. "But we were very much ignorant of how behavior was changing on the ground."



Vespignani and his colleagues used the Global Epidemic and Mobility Model (GLEaM) to predict the risk of Ebola Virus Disease (EVD) being imported into various countries. These graphs show the top 16 countries predicted to be at risk of importing at least one EVD case by the first and 22nd of September 2014. The prediction is based on data prior to August 21, 2014 and conditional on the country having no imported cases prior to that date. The dark blue and light blue bars represent the minimum and maximum probability estimates, respectively, according to different models of case detection during travel. The orange area corresponds to the probability maximum assuming the Nigerian outbreak starts to follow the same dynamic of the other West African countries affected by the EVD epidemic. The graph reports the rank of Nigeria as well, although it had already experienced a case importation on July 20, 2014. From FC Marcelo, A Pastore y Piontti, L Rossi, D Chao, I Longini, ME Halloran, A Vespignani, Assessing the International Spreading Risk Associated with the 2014 West African Ebola Outbreak, PLoS Current Outbreaks, *Sept 2, 2014.*

Tracking Behavior Change

Vespignani's predictions worked best over the short term. Beyond two or three weeks, howver, they began to break down. This was true for the other teams, as well: Their long-range nightmare scenarios far exceeded the actual case counts and death tolls. Many believe this was partly due to the massive international effort that those

In addition to sharing data and methods amongst themselves, a number of modeling teams are now also working together to develop better tools for gathering such behavioral information—all in hopes of being ready for action when the next outbreak occurs. "As we learn from each other," Marathe says, "the models will become even more useful." □

BY KRISTIN SAINANI, PhD

SINGLE-CELL GENOMICS: CAN BIOINFORMATICS UNLOCK ITS POTENTIAL?

The tools to sequence the genomes of individual cells yield data that's noisy and somewhat unreliable. What bag of tricks can bioinformaticians use to address these challenges?

To study genomes, researchers have typically pooled the genetic material from thousands of cells together. But this approach can only get at "average genomes" or "average transcriptomes." And sometimes this isn't enough.

"Averages can be very misleading or not meaningful," says **Cole Trapnell, PhD**, assistant professor of genome sciences at the University of Washington. "Plus, there are certain biological questions that you cannot answer unless you take single-cell measurements."

For example, the brain consists of multiple cell types enmeshed with one another; and cancerous tumors are a mix of genetically diverse cells, including some that drive invasion, metastasis, and treatment resistance. To understand what's really going on in such heterogeneous groups of cells, researchers need to study the genomes of individual cells.

Now, using single-cell genomics, researchers have the tools to resolve individual clones within tumors, discover new cell types in the brain, find genetic abnormalities in embryos, and detect rare cancer cells in the blood, among other exciting applications.

But to unlock the full potential of single-cell data, advances in bioinformatics are needed. Many bioinformatics tools that were developed to process and analyze bulk data don't work well when applied to single-cell data. Plus, novel algorithms will be needed to address the biological questions that only single-cell data can answer.

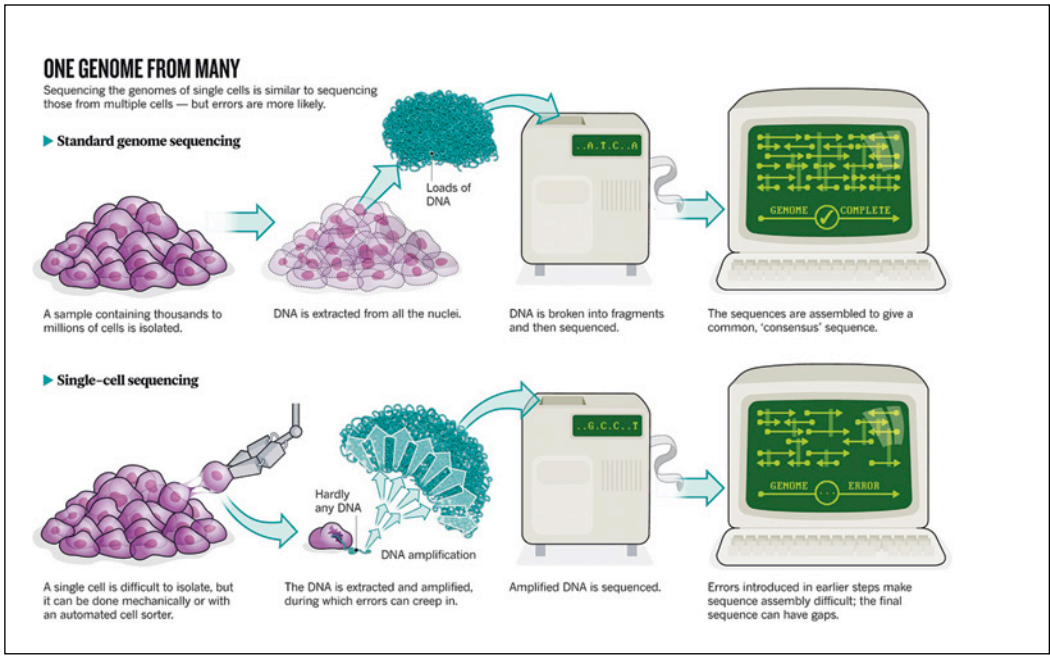
Single Cell Genomics Defined

With recent breakthroughs in isolating individual cells and in amplifying and sequencing DNA and RNA, researchers can now measure the genomes and transcriptomes of thousands of individual cells at once. A single cell carries only two copies of DNA and sometimes just a few messenger RNA (mRNA) transcripts. To sequence the DNA or make sense of its transcripts, researchers must first amplify those numbers as much as a billion-fold using powerful new amplification technologies such as Multiple Annealing and Looping Based Amplification or Multiple Displacement Amplification. This creates a "pool" of genetic material that can then be sequenced and analyzed just as pooled genetic material from multiple cells would be.

But there is a difference: The data are noisier and less complete than the bulk data garnered from multiple cells. Single-cell data are also larger in size and scale—often involving hundreds or thousands of samples rather than the tens of samples typical of bulk experiments. This is where bioinformaticians have their work cut out for them.

Noisy Data

When the small numbers of DNA or mRNA transcripts from a single cell are amplified, some regions of the genome or some transcripts are amplified better than others, leading to distortions. In addition, researchers cannot replicate their work on a single cell—because the same material can't be amplified and measured twice. This makes it harder to separate experimental errors from real biological variation. In an October 2015 paper in *Nature Communications*, researchers in the UK estimated that of the observed variation in gene expression patterns in single-cell



Reprinted with permission from MacMillan Publishers, Brian Owens, Sequencing DNA from individual cells is changing the way that researchers think of humans as a whole, Nature *491, 27-29 (2012).*

genomics experiments, only 18 percent was due to true biological variation. The remainder was due to technical noise.

"Most of the computational efforts so far have been toward trying to separate the true signal from the noise. That's where most of the field is now," says **Peter Kharchenko, PhD**, assistant professor of biomedical informatics at Harvard Medical School.

Experimental tricks can help, Kharchenko notes. For example, researchers can tag each original mRNA transcript with a "unique molecular identifier"—a short random sequence that acts like a barcode—prior to amplification. Since each unique tag corresponds to only one transcript molecule in the original sample, this method can generate an accurate transcript count regardless of amplification errors. "The computational aspects of this are pretty straightforward, but it results in a drastic reduction of the noise in the data," Kharchenko says. Also, researchers can add "spike-in RNAs" to each sample—control RNAs with known composition—to help detect experimental aberrations.

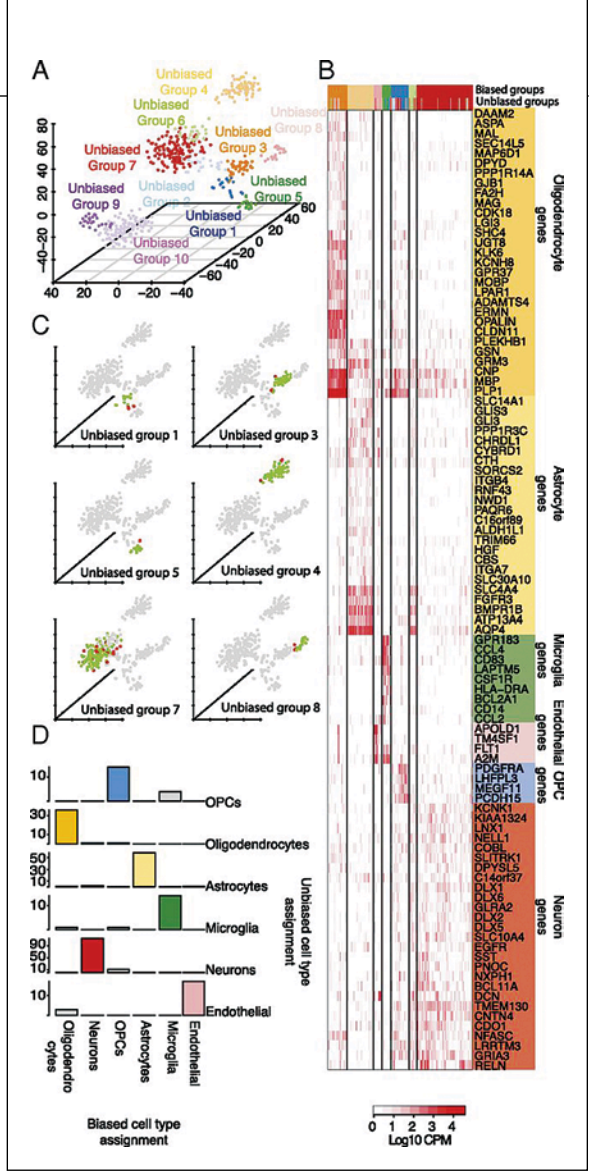
Computational tricks for reducing noise are also being introduced, but remain harder to implement. "Right now, the algorithms aren't packaged in ways that average people can use, so you have to have a bioinformatics person to string it all together," says **Robert C. Jones, MS**, executive vice president for research and development at Fluidigm, a company that makes tools for single-cell genomics. "Someday we hope to provide our customers with a regular pipeline so that people can just turn the crank."

User-friendly software is beginning to emerge. A 2015 paper in *Nature Methods* introduced Ginkgo (http://qb.cshl.edu/ginkgo/), an interactive Web-based program that automatically processes single-cell DNA sequence data, maps the sequences to a reference genome, and creates copy number variant profiles for every cell. The software—which was created in the lab of **Michael Schatz, PhD**, associate professor of quantitative biology at Cold Spring Harbor Laboratory—has built-in algorithms to correct amplification errors.

Missing Data

Amplification is fraught with another key problem: Some regions of the genome or mRNA transcripts may be completely missed. "Your body's very good at copying entire chromosomes. To do so requires this amazingly beautiful orchestrated dance where you bring together many proteins, including for proof-reading and error-correction," Schatz says. "We hijack some of those systems to make copies of the DNA through PCR (polymerase-chain reaction), but it's not nearly as sophisticated as what goes on inside your body." As much as 30 percent of the genome may be unamplified and missed; and as many as 60 percent of heterozygous alleles may be missed. With RNA, the problem is even worse—researchers estimate that some protocols miss as many as 60 to 90 percent of all transcripts present in the cell.

"There's a big zero problem," Trapnell says. "We have to find new ways to deal with that missing data. There's discussion right now in the community about how best to do that: Do we want to fill it in based on our best guess? Do we want to build models that can tolerate a lot of zeros, and don't have a problem with it? It's not obvious what the right way to go forward is."



When Steven Quake's group at Stanford used an unbiased approach to sort single-cell genomic data for 466 individual cells into distinct groups defined by the entirety of their molecular signatures, they identified 10 distinct cell groups (8 adult [1-8] and 2 fetal [9-10]) as shown in the landscape in (A). These classifications compared favorably to a biased approach based on known markers for specific cell types (B-D). In C, for example, the classifications agreed for cells shown in green and did not agree for the small number of cells shown in red. And in D, agreement is shown in colored blocks and disagreement in gray, with the number of cells shown on the y-axis. Reprinted from S Darmanis, SA Sloan, Y Zhang, et al., A survey of human brain transcriptome diversity at the single cell level. PNAS *112(23): 7285-90 (2015).*

Kharchenko's group has developed software called SCDE (Single-Cell Differential Expression) to analyze single-cell RNA-seq data (http://pklab.med.harvard.edu/scde/). The model uses a Bayesian approach that accounts for the likelihood of dropout events. "We had to incorporate explicitly the probability of failing to observe a gene. By predicting the probability of not being able to see a gene in a given cell, then you can propagate that uncertainty further into other analyses," Kharchenko says. "The computation becomes a little more complicated, but you're better off taking into account the uncertainty of the measurement."

Drawing on information gleaned from bulk data—such as the frequency of a particular mutant allele in a tumor—can also give clues as to the impact of dropouts. "So it's not so much about sheer brute-force analytic methods in the single cell—it's also knowing how to bring in

different datasets to help you make better sense of everything," says **Winston Koh**, a doctoral student in bioengineering in Stephen Quake's lab at Stanford University. In a 2014 paper in *PNAS*, Koh and others combined bulk and single-cell genomic data to reconstruct the clonal architecture of childhood acute lymphoblastic leukemia.

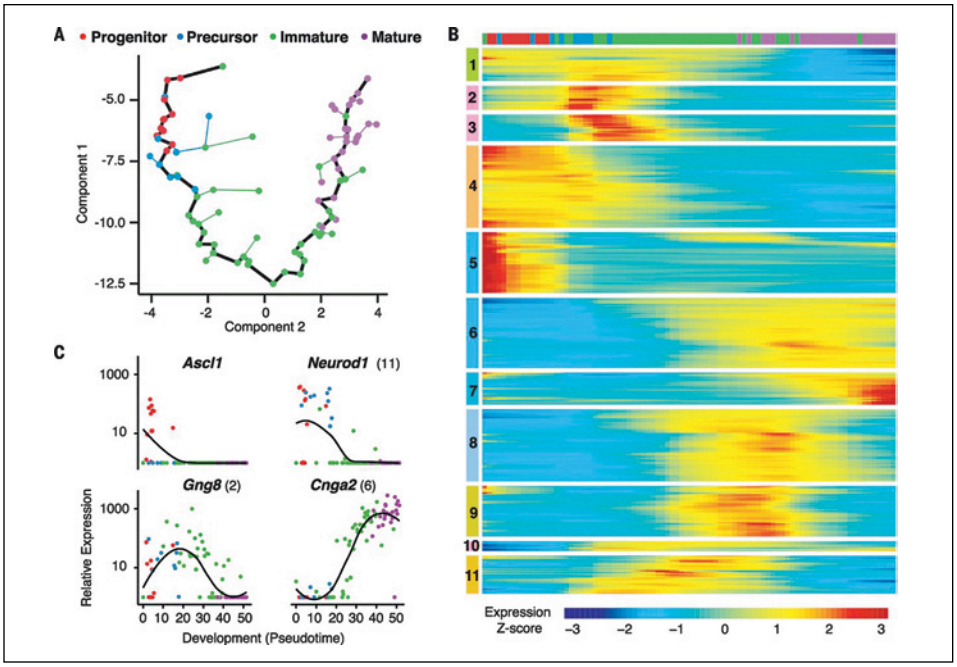
Making Sense of Single-Cell Data

To gain biological insight, researchers start by grouping cells with similar gene or gene activity profiles. But clustering is tricky because single-cell data are high-dimensional (involving thousands of sequences or expression profiles) and involve complex relationships. Traditional clustering algorithms such as Principal Components Analysis (PCA)—which assume a simple linear relationship between variables—aren't optimal for these data. So, bioinformaticians are exploring alternatives, including t-SNE (t-distributed stochastic neighbor embedding), which is well-suited for high-dimensional, non-linear data.

"I'm pretty impressed with how it's all coming together in the field. We are moving beyond simple PCA-type analyses to more sophisticated algorithms," says **Stephen Quake**, **D.Phil**, professor of bioengineering at Stanford University and a founder of Fluidigm. "The whole ecosystem around that is looking very promising." In a 2015 paper in *PNAS*, Quake's team applied t-SNE to single-cell transcriptome data from 466 brain cells to identify 10 distinct cell groups in the adult and fetal brain—eight of which corresponded to known cell types in the brain. They then further classified cells into subpopulations based on their gene expression profiles—a first step towards building a comprehensive atlas of cell types in the brain.

Beyond grouping cells, researchers are also developing algorithms for ordering cells by temporal or developmental stage. This problem is challenging because it may require bioinformaticians to rethink their approach to cell classification, Trapnell says. Rather than trying to classify cells into clear-cut, discrete types, we should think of cells as lying more on a continuum, he says. "There's a desire to put things into nice neat bins. And I think that's not working well so far for single-cell data. So we might just need to be a little bit more flexible about how we analyze this stuff."

Trapnell's lab has developed Monocle, a toolkit for single-cell expression data that reconstructs the trajectory along which cells are presumed to travel, such as during development or differentiation (http://cole-trapnell-lab.github.io/monocle-release/). "Monocle is designed to put cells in continuous order by how differentiated they are—from undifferentiated stem cells to the fully differentiated state," Trapnell says. In a recent paper in *Science*, Trapnell and others used Monocle to track the maturation of nasal olfactory cells in mice. "Because we capture the complete, continuous progression from neuronal progenitors to mature neurons, we can see the exact moment in development that these neurons select which member of a large family of sensory genes to express," Trapnell says.



Using single cell gene expression, Trapnell and his colleagues determined that olfactory neurons exhibit large-scale shifts in gene expression during development. An unsupervised analysis of single-cell gene expression profiles using an algorithm called Monocle revealed a linear trajectory (black line in A) along which cells develop in pseudotime. Coloring of cells based on the expression of developmental markers shows that the trajectory corresponds to a stepwise development from olfactory progenitors to precursors to immature and ultimately mature olfactory sensory neurons. Global analysis of gene expression kinetics along the trajectory identified 3,830 genes that vary significantly during development (B) and can be hierarchically clustered into 11 nonredundant groups that covary over the trajectory. The bar on top shows the locations of individual cells, colored by stage of development, along this developmental trajectory. The Expression Z-score indicates changes in a gene relative to its dynamic range over pseudotime. Kinetic diagrams (C) show the expression of known markers of different developmental stages over the developmental progression. Reprinted with permission from NK Hanchate, K Kondoh, Z Lu, et al., Single-cell transcriptomics reveals receptor transformations during olfactory neurogenesis, Science *350:6265/1251-1255 (2015).*

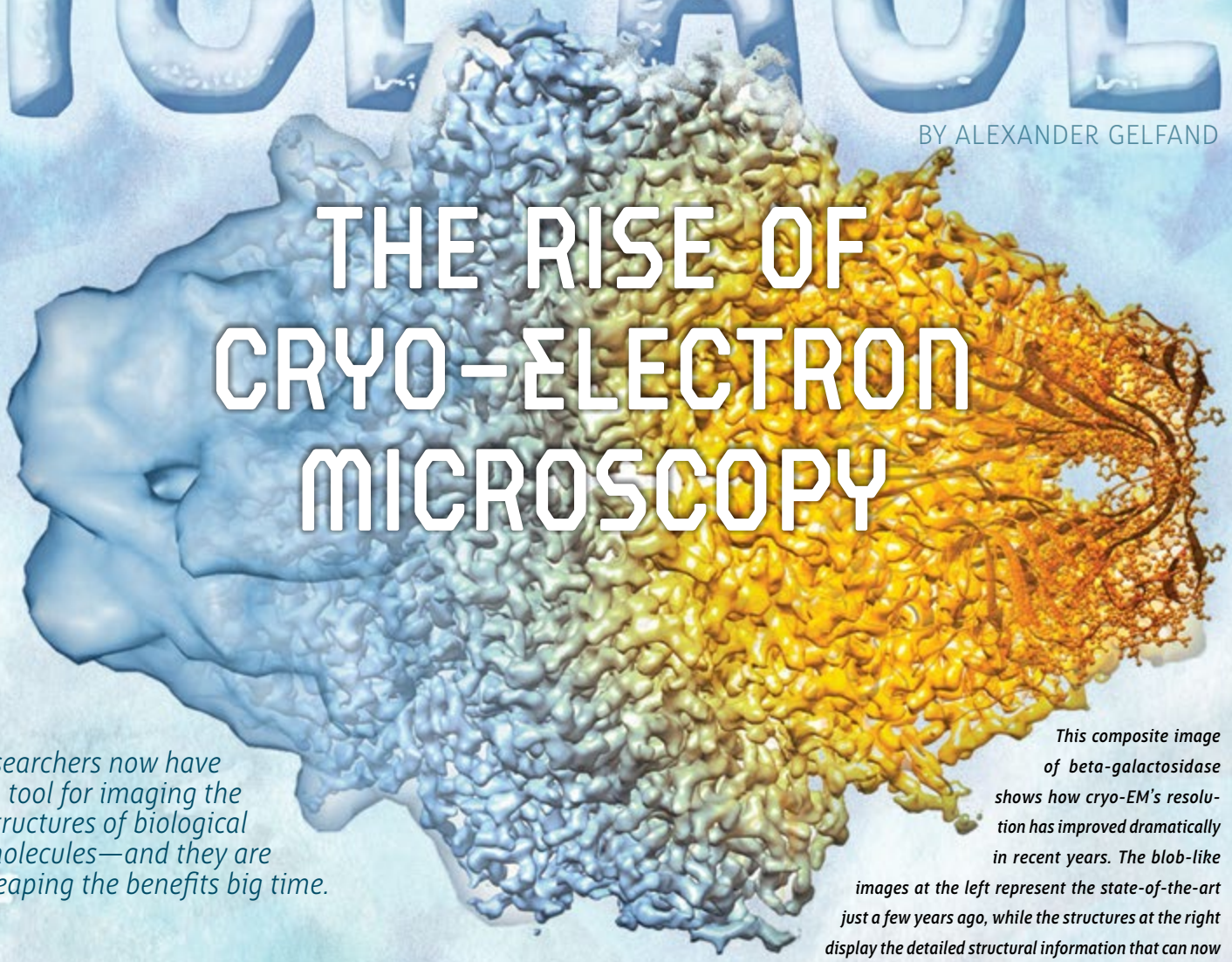
Rapidly Changing Technology

The technology for single-cell genomics is rapidly advancing—and the bioinformatics challenges may shift accordingly. "It's quite frustrating because the approaches that you'll use one day could be totally changed the next day," Schatz says. Just as bioinformaticians will have to keep pace with the technology, biologists will need to stay abreast of the latest bioinformatics innovations, he says. "Researchers who want to use these technologies have to pay really close attention to the state-of-the-art in the field and make sure that they are using all the best practices available at the time." □

ICE AGE

BY ALEXANDER GELFAND

THE RISE OF CRYO-ELECTRON MICROSCOPY



Researchers now have a tool for imaging the structures of biological molecules—and they are reaping the benefits big time.

This composite image of beta-galactosidase shows how cryo-EM's resolution has improved dramatically in recent years. The blob-like images at the left represent the state-of-the-art just a few years ago, while the structures at the right display the detailed structural information that can now be gained using this method, which Nature Methods *named "Method of the Year 2015." Credit: Veronica Falconieri, Subramaniam Lab, National Cancer Institute.*

Structure determines function.

That, at least, is what structural biologists will tell you. And these days they have a powerful new tool—or rather, a vastly improved old one—for determining the structure of biological molecules, and thereby ascertaining what they do and how they do it.

The tool is cryo-electron microscopy (cryo-EM), a suite of methods that allows researchers to construct three-dimensional images of microscopic objects using focused beams of electrons and super-cold temperatures. Until recently, the technique could only produce fuzzy, blob-like images of biological macromolecules with nothing like the fine detail available through methods like x-ray crystallography and nuclear magnetic resonance (NMR). "A few years ago, we were considered the 'blobologists,'" says **Melanie Ohi, PhD**, associate professor of cell and developmental biology and a member of the Center for Structural Biology at Vanderbilt University.

But over the past several years, cryo-EM has begun to produce images with the kind of near-atomic resolution that was once limited to its rivals—but without their drawbacks.

"The field is in a revolution," says **Klaus Schulten, PhD**, a computational biophysicist at the University of Illinois.

Cryo-EM is now producing high-resolution structures of large biomolecules and molecular machines such as chromatin, supercoiled DNA, intracellular vesicles, membrane pores, ion channels, and individual virus particles. At the same time, it is being used to hone in on the atomic-level structures of smaller and smaller individual molecules at an extraordinary level of detail. And on top of that, researchers are figuring out how to extract different conformations of the same molecules from

cryo-EM data—allowing them to create simple animations that illustrate how a molecule's structure changes as it does its job in the cell.

BETTER THAN CRYSTALLOGRAPHY

Before the cryo-EM revolution started, the gold standard for determining molecular structure at atomic resolution was x-ray crystallography. This method requires that proteins be crystallized before they can be scanned, but crystallization is not always possible. Nor is it necessarily desirable. As Schulten points out, crystallizing proteins forces them to line up "like Prussian soldiers in crystal," denying researchers the opportunity to capture the various shapes, or conformations, that the biological molecules assume as they go about their business—conformations that provide the key to a molecule's function. Potential insights into the workings of the largest molecular machines—which tend to be extremely flexible and dynamic systems with many moving parts—have thus remained obscured.

NMR offers resolution on par with x-ray crystallography without the need for crystallization But it has a different limitation: It can only determine the structure of relatively small objects—certainly nothing as large as a molecular machine such as a ribosome or an ion channel, much less the cell that contains it.

Cryo-EM, by comparison, can handle a wide range of scales; though the method can't yet resolve the smallest objects available to NMR, it's getting there, and it can already be used to image much larger ones with ease. In addition, the freezing process used in cryo-EM allows proteins to remain in something resembling their native state, offering the possibility of gleaning more information about function.

THE FREEZE

In cryo-EM, researchers plunge thin films of sample solution into baths of ethane that have been cooled to the temperature of liquid nitrogen (-180˚C). The biological material in the samples is flash-frozen in a delicate layer of glass-like ice, which can then be bombarded with electrons.

Cryo-EM comes in two principal flavors: single-particle analysis and cryo-electron tomography (CET). In single-particle analysis, researchers capture two-dimensional projections of hundreds of thousands, even millions, of the same kinds of biological objects—membrane proteins, cellular complexes, and viral capsids—that have been randomly distributed throughout the ice in different orientations. Image-processing algorithms then sort and average those 2-D–projections to construct a three-dimensional structure.

Cryo-electron tomography, on the other hand, involves taking multiple images of a single object. By tilting the object at various angles relative to the electron beam, researchers can again build up a 3-D structure. The resolution offered by CET is currently lower than that of single-particle analysis, but it can be used to image larger one-of-a-kind objects like organelles or even entire cells. In the most cutting-edge applications of cryo-EM, the two methods are used in combination.

THE REVOLUTION: LEAVING BLOB-OLOGY BEHIND

For many years, cryo-EM could not achieve the kind of near-atomic resolution available to NMR and x-ray crystallography, nor could it handle the smallest molecular structures—in part because researchers had to limit the power of their electron beams in order to avoid destroying the samples they were trying to study.

Recent technological breakthroughs now allow researchers to collect higher resolution images using fewer of the tiny charged particles, however, reducing the amount of damage done to the samples while improving the quality of the resulting structures. Schulten himself helped revolutionize the field by developing computational methods for fitting hi-res structures generated by x-ray crystallography into the relatively blobby, lo-res structures derived from cryo-EM. This so-called hybrid approach uses computational modeling to develop realistic and highly detailed structures that conform to what scientists have learned about the dynamic behavior of biological molecules through decades of computer simulations. Schulten and his colleagues recently used precisely such methods to determine the atomic structure of the *Rous sarcoma* virus, a cancer-causing retrovirus that is used in cancer and HIV research.

Recently, however, cryo-EM has begun to achieve resolutions similar to x-ray crystallography all by its lonesome. This past year, for example, **Sriram Subramaniam, PhD**, a senior investigator at the National Cancer Institute's Center for Cancer Research, imaged a small metabolic enzyme called beta-galactosidase at a resolution of 2.2 Å, or .22 billionths of a meter, using nothing but single-particle analysis. At that level of detail, one can see individual water molecules bound to the protein—something that would have been unimaginable with cryo-EM just a few years ago, and that could eventually assist in drug design.

Much of this progress is due to the development of new, highly sensitive cameras called direct electron detectors, and to the advent of powerful computing clusters that can be used to process the enormous volumes of data they generate. ("We're generating many terabytes every day," Subramaniam says.) But it is also due to the development of new and improved image-processing algorithms—algorithms that play a vital role in virtually every stage of cryo-EM.

In single-particle analysis, for example, image-processing algorithms must determine the relative orientations of enormous numbers of 2-D projections, then align and average them in order to reconstruct a 3-D image. Researchers like **John Briggs, PhD**, at the European Molecular Biology Laboratory in Heidelberg, have been applying similar techniques to improve the resolution of cryo-electron tomography, essentially taking multiple images of repeating structures within a given molecular machine and averaging them, much as multiple images of individual objects are averaged in single-particle analysis. Briggs and his colleagues have used this approach, which is known as subtomographic averaging, to resolve the structures of protein complexes that allow vesicles to travel from one cell compartment to another, and that enable the HIV-1 virus to self-assemble.

MOLECULAR MOVIES

Increasingly, algorithms that use sophisticated statistical techniques like Bayesian hierarchical classification are also being used to sort, select, and categorize the objects that are imaged, determining which ones represent different conformations of the same basic proteins—a process that Ohi refers to as *in silico* purification. With snapshots of enough conformations in hand, researchers can create simple movies that illustrate how molecular machines move through different states as they do their work: transporting cargo through a cell, synthesizing proteins, transcribing RNA. "If you can computationally tease out those states," says Subramaniam, "then you can piece them back together to derive a plausible sequence of what [a machine] actually does."

Researchers are continuing to extend the limits of cryo-EM, achieving ever-higher resolutions and imaging ever-tinier domains. Even now, Subramaniam and his colleagues are preparing to publish a paper in *Cell* describing the structure of a 200-kilodalton ion channel, one of the smallest proteins ever imaged with cryo-EM, at a resolution of 3.8 Å. (One dalton is equivalent to the mass of a single proton or neutron.)

The goal, however, is not just to go smaller, but also to go bigger: to achieve the resolution now possible with single-particle analysis at the scale available to cryo-tomography. "We want to get an atomic-level picture of a cell," says Schulten.

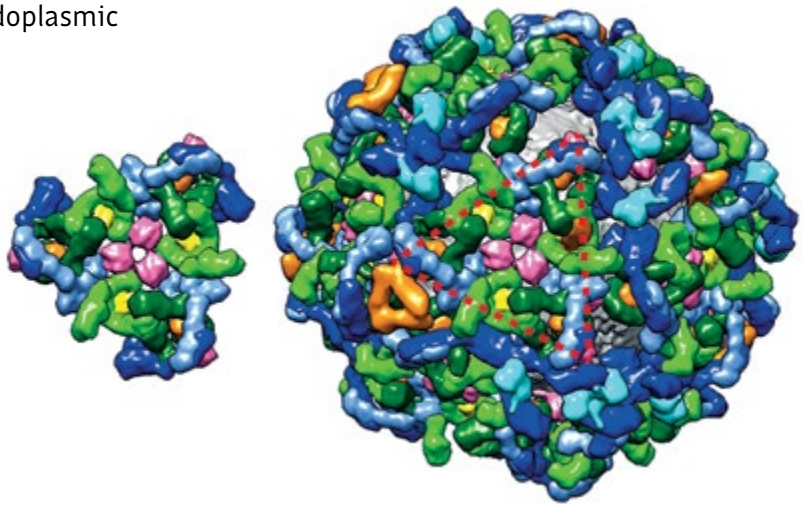
In time, thanks to some (extremely) cool technology, he and his colleagues may get their wish.

But the pictures being generated right now already set the dedicated biologist's skin to tingling. The image gallery presented on the following pages is just a small sampling extracted from the steady stream of cryo-EM research now being published. Much of this work is appearing in high-profile journals because cryo-EM is enabling structural biologists to gain deeper insights about how molecular machines, the workhorses inside our cells, actually function. □

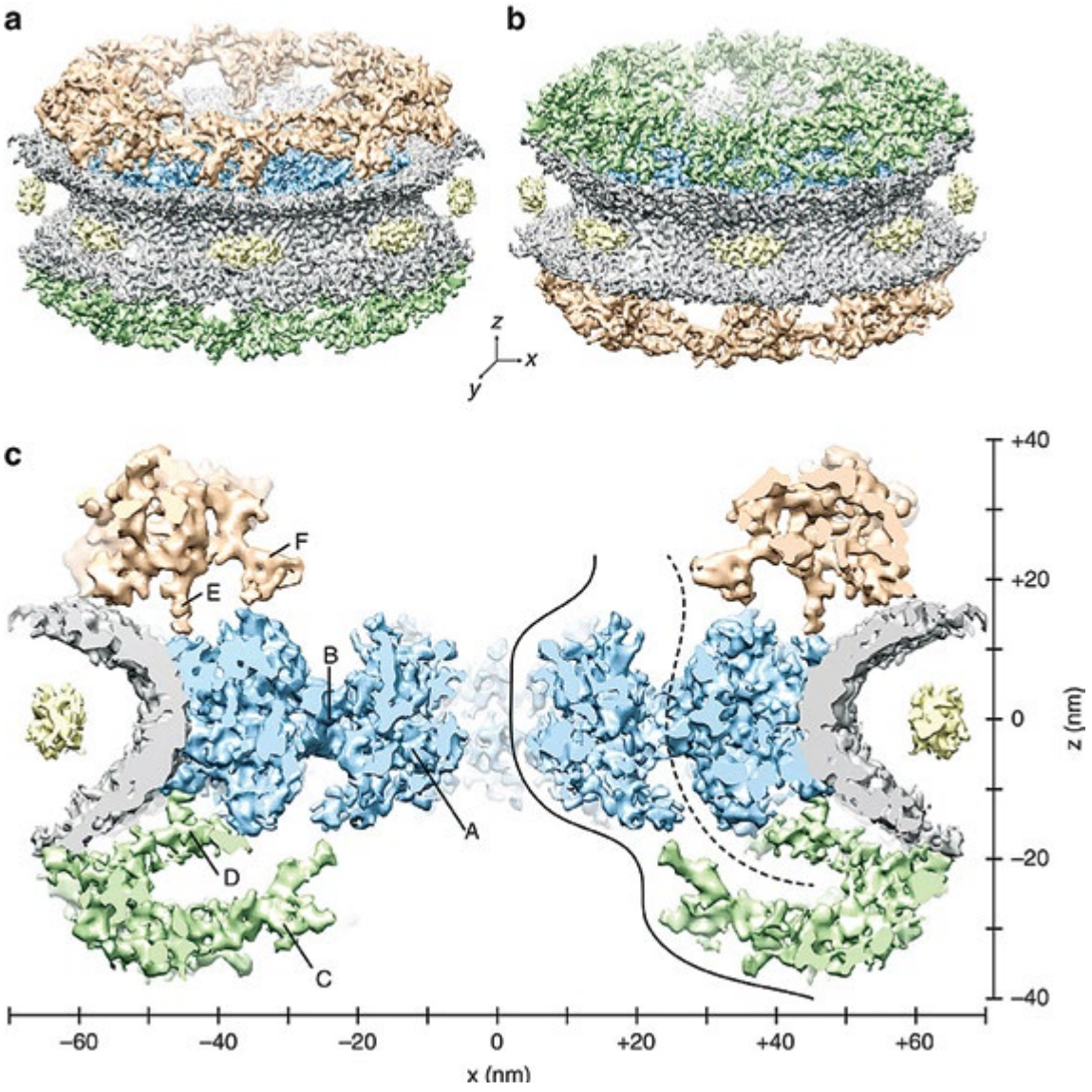
CRYO-EM IMAGE GALLERY

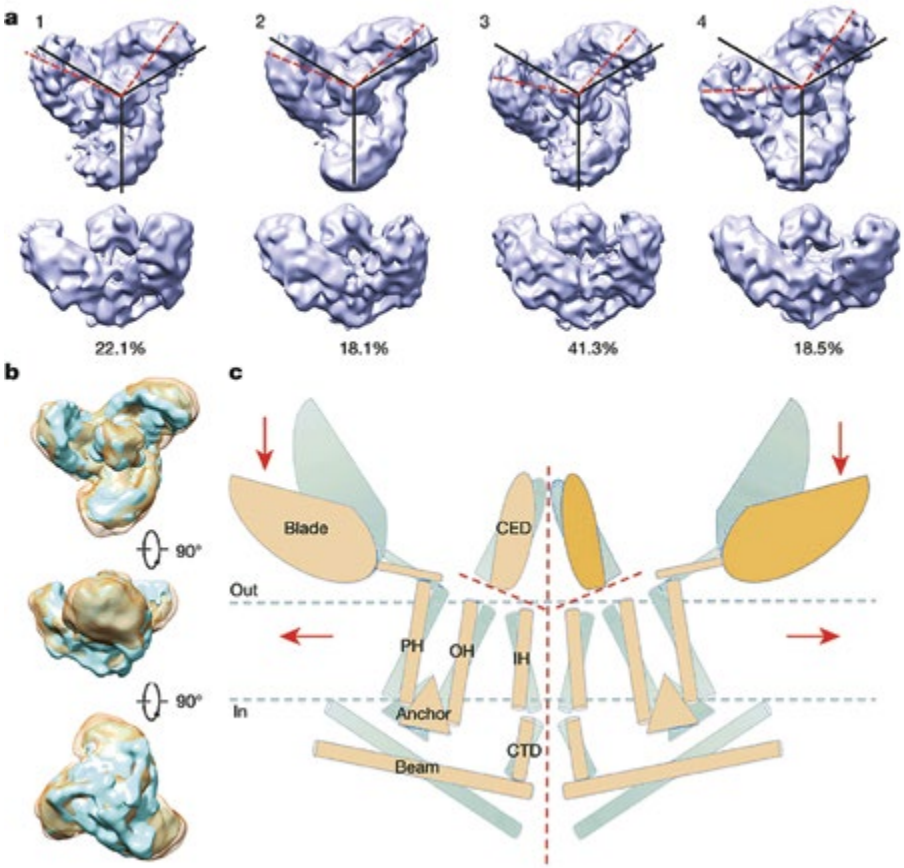
2015 saw a plethora of high-profile journal publications describing how jaw-dropping images of macro-molecular structures (obtained with cryo-EM) are providing novel insights into biological function. A sampling is shown on these pages.

Discerning how coated vesicles form. Trafficking vesicles bud from one cellular compartment and fuse with another to transport material within cells. To form such vesicles, membrane coats localize cargo and polymerize into cages to bend the membrane. Although extensive structural information is available for components of these coats, the heterogeneity of trafficking vesicles has prevented an understanding of how complete membrane coats assemble on the membrane. Using a combination of cryo-electron tomography, subtomogram averaging, and cross-linking mass spectrometry, researchers derived this complete model of the highly interconnected coat protein complex I (COPI), a coat involved in vesicle traffic between the Golgi and the endoplasmic reticulum. At left is a 'triad', the building block of the COPI coat. At right, the complete COPI-coated vesicle made of an assembly of triads. The development of this model provided novel insights into how coated vesicles form. From SO Dodonova et al., A structure of the COPI coat and the role of coat proteins in membrane vesicle assembly, *Science* 349, 195 (2015). Image: Svetlana Dodonova, European Molecular Biology Laboratory (EMBL).



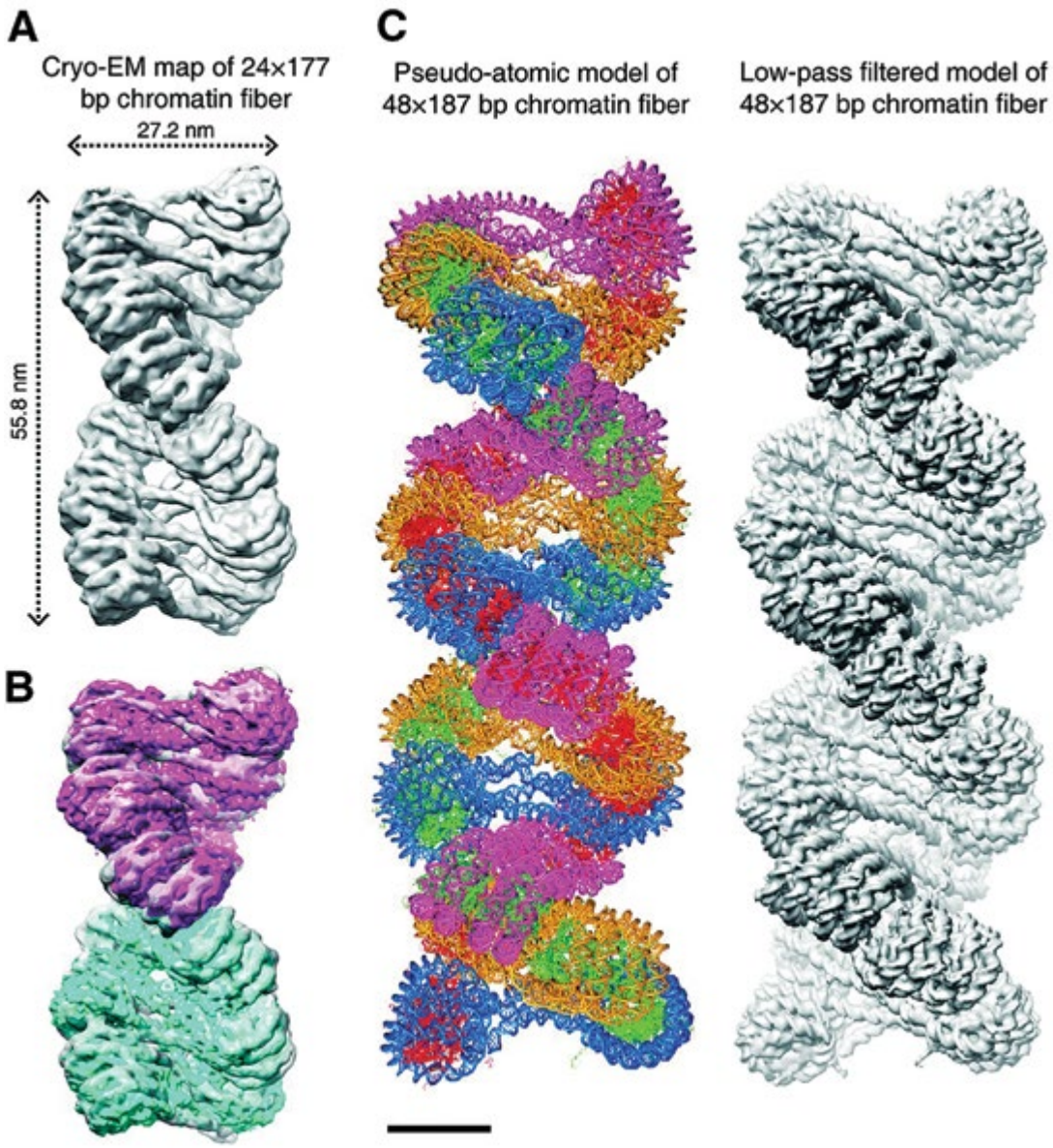
Finding the path through the nuclear pore complex.
The nuclear pore complex (NPC), one of the largest protein complexes in the cell, is responsible for mediating or blocking the exchange of materials between the nucleus and cytoplasm. Cryo-EM was recently used to determine the structure of the NPC as shown above. The NPC is comprised of three layered rings: the cytoplasmic ring (gold), the spoke ring which forms the pore (blue) and the nucleoplasmic ring (green) as shown in (a) and (b) below (inverted views of the same complex). In cross-section (c), extended linker structures protrude from the nucleoplasmic ring (C and D), as well as from the cytoplasmic ring (E). Likely nuclear transport routes pass through the nuclear pore complex barrier, as illustrated by solid and dashed curves. The axes show the dimensions of the NPC in the *x*- and *y*-direction. From M Eibauer, M Pellanda, Y Turgay, A Dubrovsky, A Wild, and O Medalia: Structure and Gating of the Nuclear Pore Complex. *Nature Communications.* June 26, 2015. doi: 10.1038/ncomms8532.





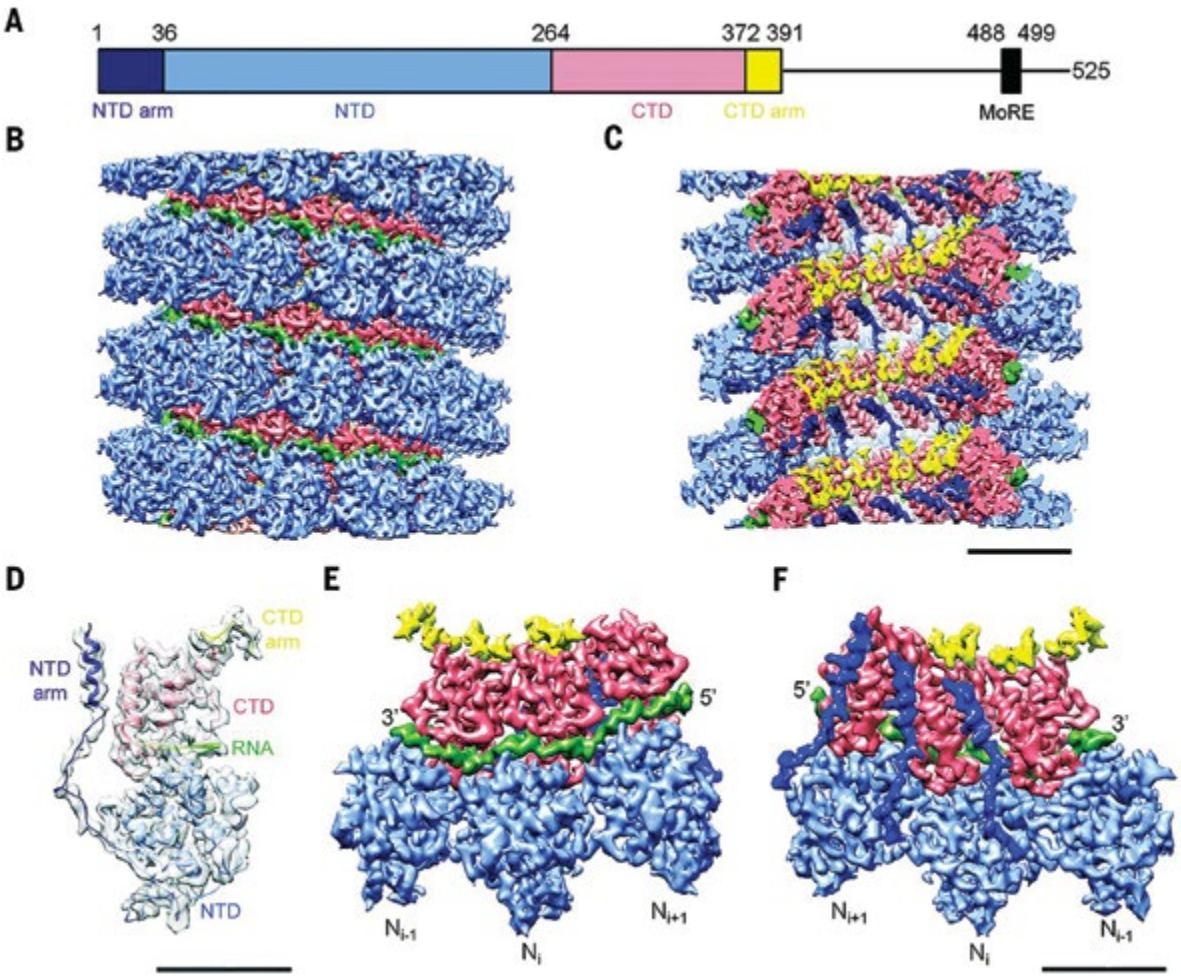
Propellers with a soft touch. Mechanosensitive cation channels serve key roles in converting mechanical stimuli into various biological activities, such as touch, hearing and blood pressure regulation, through a process termed mechanotransduction. The Piezo family of cation channels, in humans and other mammals, plays broad roles in multiple physiological processes, including body proprioception, sensing shear stress of blood flow for proper blood vessel development, regulating red blood cell function and controlling cell migration and differentiation. Until now, researchers have not known the overall structural architecture and gating mechanisms of Piezo channels.

In recent work published in *Nature*, researchers determined the cryo-electron microscopy structure of the full-length (2,547 amino acids) mouse Piezo1 cation channel at a resolution of 4.8 Å. Here, four representative cryo-EM Piezo1 structures are shown (a). They consist of a trimeric propeller-like structure with the extracellular domains resembling three distal blades and a central cap. The rather flexible extracellular blade domains are connected to the central intracellular domain by three long beam-like structures. The red dashed lines, which represent observed positions of the propeller blades, reveal that the blades are not always positioned 120 degrees apart (black solid lines). By overlaying the third and fourth structures (b) in orange and cyan, one can see the centripetal movement of the blades (top) and the tilted movement of the beams relative to the plasma membrane plane (bottom). It's possible that Piezo1 uses its peripheral regions as force sensors to gate the central ion-conducting pore, as diagrammed in (c), where the blue and orange models represent the closed and open state channels, respectively, and red arrows indicate force-induced motion. Red dashed lines indicate the possible ion-conduction pathways. Reprinted by permission from Macmillan Publishers Ltd: J Ge, W Li, Q Zhao, et al., Architecture of the mammalian mechanosensitive Piezo1 channel, *Nature* (2015) doi:10.1038/nature15247.



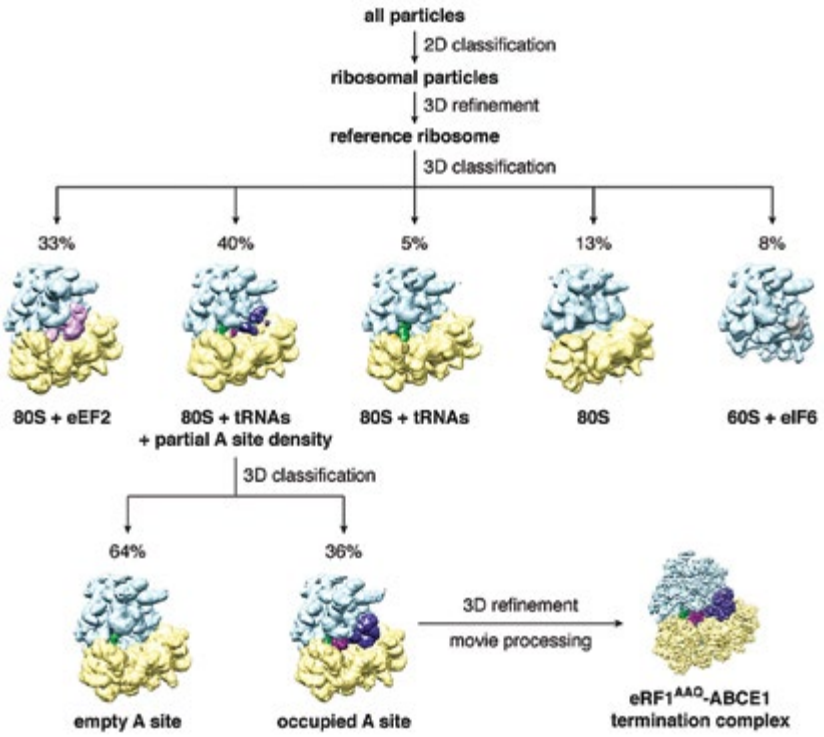
A clearer picture of chromatin structure.
DNA packs itself into the small space inside the cell nucleus by wrapping around histone proteins to form nucleosomes. These basic elements repeat as beads-on-a-string, interconnected by sections of linker DNA. In addition, a linker histone called H1 coils the beads-on-a-string structure into a 30 nm chromatin fiber whose structure has been a matter of debate. Cryo-EM structures recently described in *Science* are now offering a clearer picture. Here we see a cryo-EM map of the 30 nm chromatin fiber (A). This structure was used to build the model of a longer fiber as shown in (C). Reprinted with permission from F Song, P Chen, D Sun et al., Cryo-EM Study of the Chromatin Fiber Reveals a Double Helix Twisted by Tetranucleosomal Units, *Science* 344 (6182), 376-380 (2014).

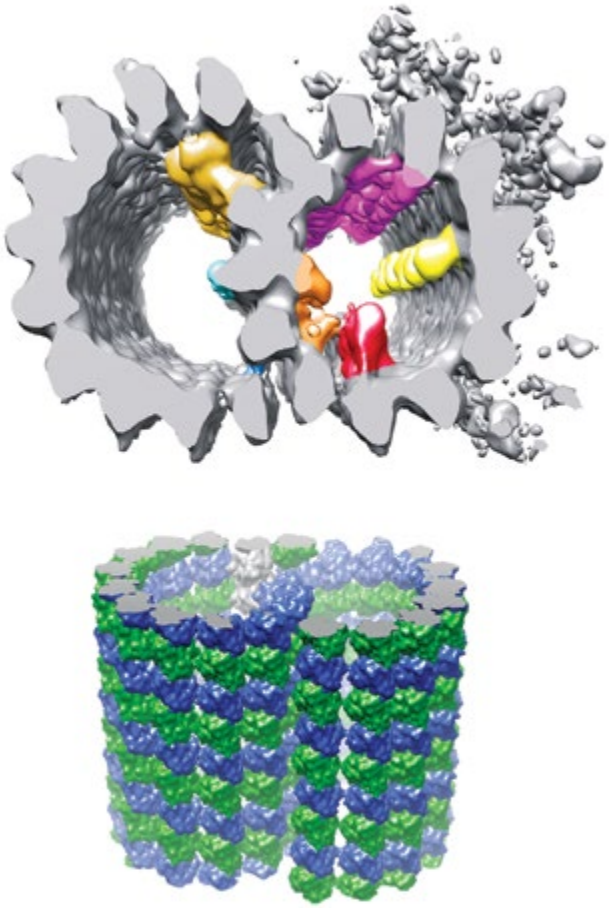
The flexibility of supercoiled DNA revealed.
DNA supercoiling regulates access to the genetic code, which strongly affects DNA metabolism. Researchers recently used cryo-electron tomography together with biochemical analyses and computer simulations to investigate the various shapes taken by individual purified DNA minicircle topoisomers with defined degrees of supercoiling. They found that each topoisomer adopts a unique and surprisingly wide distribution of three-dimensional conformations including circles, handcuffs, "racquets" and figure eights. Molecular dynamics simulations independently confirmed this conformational heterogeneity and provide atomistic insight into the flexibility of supercoiled DNA. These images show the structure of the DNA calculated with the supercomputer simulations (in color); and in the images to the right, superimposed upon the cryo-electron tomography data (in white or yellow). From RN Irobalieva, JM Fogg, DJ Catanese Jr, T Sutthibutpong et al., The Structural Diversity of Supercoiled DNA, *Nature Communications* 6 (2015). Image credit: Thana Sutthibutpong.



Helical measles. Viruses rely on their capsid proteins to package and protect their genome. For the measles virus and other viruses in the same family, multiple capsid proteins together form a helical shell around the viral RNA and are collectively called the nucleocapsid. In recent work, researchers determined the high-resolution cryo-EM structure of the measles virus nucleocapsid at near-atomic resolution. The nucleocapsid consists of a series of connected N-nucleoproteins (D) wrapped around the viral RNA (green). This figure shows the nucleocapsid structure in front view (B) and cutaway view (C), with colors denoting the N-nucleoprotein's two domains [N-terminal domain (blue) and C-terminal domain (pink)], as well as its N-terminal (dark blue) and C-terminal (yellow) arms that hold nucleoproteins together contributing to the stability of the whole architecture. Close-ups of three consecutive N-nucleoproteins from the exterior (E) and interior (F) of the helix reveal the nucleoproteins stacked into a helical shape. The structure reveals how the nucleocapsid assembles and how the nucleo-protein and viral RNA interact, both of which may inform drug design. From I Gutsche, A Desfosses, G Effantin, et al., Near-atomic cryo-EM structure of the helical measles virus nucleocapsid, *Science* 348:6235:704-707 (2015). Reprinted with permission from AAAS.

Understanding how ribosomes spot stop codons.
As a ribosome chugs along a strand of messenger RNA, it adds amino acids to a peptide chain according to the prescription of trios of mRNA nucleotides known as codons. In coordination with molecules called release factors, the ribosome stops adding more amino acids when it comes across one of three universally conserved stop codons: UAA, UAG or UGA. Eukaryotes rely on an omnipotent release factor (eRF1) that recognizes all three stop codons, but avoids each of the 61 sense codons for amino acid addition. To better understand how eRF1 discriminates between stop codons and sense codons, researchers trapped some ribosomal complexes at this crucial recognition step and examined the complexes using cryo-EM. Using a Bayesian classifier, they isolated five distinct classes of ribosomal complexes as shown here. One of these classes was further sub-classified (bottom line) to identify only the eRF1-containing particles, permitting a high-resolution view of the recognition event. By comparing the structures for each of the three stop codons, the researchers could see how eRF1 remodels mRNA in a way that permits their sequence to be queried. The results provide a molecular framework for understanding eukaryotic stop codon recognition. Reprinted by permission from Macmillan Publishers Ltd: A Brown, S Shao, J Murray, RS Hegde, V Ramakrishnan, Structural basis for stop codon recognition in eukaryotes, *Nature* 524:7566 (2015).





Seeing the spokes of a molecular motor.
The main skeleton of cilia and flagella is a microtubule doublet (MTD). The structure of tubulin, the main component proteins of MTD, has been previously solved at atomic resolution. In recent work, researchers analyzed the three-dimensional structure of the entire MTD from *Tetrahymena* cilia at ~19 Å resolution by single particle cryo-electron microscopy to reveal how various proteins such as tubulin isoforms, dyneins (motor proteins), radial spokes and microtubule inner proteins (MIPs) bind to the MTD to generate or regulate force. The image above shows various MIPs (in color) bound to the inside of the MTD as well as an external view of the MTD showing its tubulin subunits. Reprinted with permission from A Maheshwari, JM Obbineni, et al., α- and β-Tubulin Lattice of the Axonemal Microtubule Doublet and Binding Proteins Revealed by Single Particle Cryo-Electron Microscopy and Tomography, *Structure* 23:9:1584-95 (2015). □

BY KATHARINE MILLER

DATA'S identity

How can we make data sharing less daunting in order to address the scientific reproducibility problem?

Biomedical data is undergoing an identity crisis. "How can that be?" you may ask. It's data: bits of information stored on servers somewhere; sequences of nucleotides in a genome; levels of gene expression in lots of different cells and lots of different organisms; images of brains and lungs and hearts; and all of these things tied to particular health problems.

How lost can data be?

Quite lost, in fact. Datasets are often unnamed, undescribed, homeless, and unpublished. And the sheer quantity of biomedical data generated by diverse labs all over the world makes the problem worse: How can you find a needle in a pile of needles if the individual needle can't even be described in a unique way?

As a result of data's identity crisis, researchers can't find or access it in order to reproduce published work or use it in new ways.

"We have to solve reproducibility," says **Anita Bandrowski, PhD**, a specialist at the Center for Research in Biological Systems at the University of California, San Diego. "We're scientists for gosh sakes."

If data had a will of its own, perhaps it could be coaxed to declare its identity, tell us where it came from, where it lives and what knowledge it holds. But data does not have a will of its own. Instead, researchers carry the burden of assigning data identifiers to their data, connecting metadata descriptors to it, putting it in a reliable repository or other predictable location, and publishing it. Unfortunately, the reward system for researchers conspires to keep data in the dark.

"We have to create a culture saying that data sharing is really important," says **Vivien Bonazzi, PhD**, senior advisor for data science technologies and innovation in the Office of the Associate Director for Data Science (ADDS) at the National Institutes of Health (NIH).

Changing the data-sharing culture means changing the incentives—including the reliance on publication in a scientific journal as the sole currency of biomedicine. There must be room for recognizing the value of shared data and software. "Currently, you don't get tenure from data and software," Bonazzi says. "We have to find a way to give credit to the people doing the data wrangling."

There is hope: The NIH is pushing an overarching philosophical shift toward making datasets FAIR—findable, accessible, interoperable and reusable. "Everyone agrees on that," Bonazzi says. As a result, the NIH is funding the development of tools, software and systems to make data sharing and data discovery easier: systems for giving datasets an identifier, simplifying data annotation, and registering datasets in a searchable index. And to make it more real, the NIH wants to connect these efforts with software and supercomputing in an ecosystem called the NIH Commons. And all of those efforts will make it easier to track datasets and give researchers credit for them.

On the publishing end, there are changes afoot as well—changes designed to incentivize linking data to publications, publishing data and metadata in data journals, and even redefining what it means to

Crisis:

The Struggle to Name It, Describe It, Find It, and Publish It

publish scientific results. Some of the most interesting thinking in this area is coming out of FORCE11, the Future of Research Communications and e-Scholarship, a grassroots organization dedicated to transforming scholarly communication through technology, where people are floating lots of ideas for changing the incentives around data sharing.

With projects launched on all fronts, it's still unclear how things will play out. If the work in progress can solve data's identity crisis, it just may have a significant effect on the reproducibility of scientific research as well.

Who Am I?
Giving Data a Name

For datasets to be FAIR, they have to have a name—a way to distinguish them from all other datasets. For some digital objects, such as scientific publications, the DOI (digital object identifier) has become standard. Some research groups have also started issuing DOIs for data and software. In Europe, researchers often use URIs (Uniform Resource Identifiers) issued by identifiers.org under the auspices of the European Molecular Biology Labs (EMBL–EBI) as part of the MIRIAM (Minimum information required in the annotation of models) Registry, a catalog of data collections. But DOIs and URIs are not the only globally unique identifiers out there.

"There are various camps," says Bandrowski. Those in what she calls the "ontology camp" would like to see identifiers for each version of a software tool or dataset. Such an approach would be beneficial when researchers want to reproduce another group's research results—i.e, they might need to know the version of a dataset or software tool that was used. But it could also get cumbersome pretty quickly.

There's also "a less granular camp," Bandrowski says, that would argue for a simpler system—giving unique identifiers to the data associated with particular funding efforts. This is the approach Bandrowski's group has taken with RRIDs (Research Resource Identifiers). They provide a funder, such as the NIH, with a way to track the impact of a project.

It's also possible that multiple options can co-exist. "NIH wants to foster the community discussion and watch for coalescing around it," Bonazzi says.

Regardless of where the community ends up, all agree on the need to incentivize researchers to identify their datasets. The question is: What incentives will actually work? "Maybe if all the publishers said, what's your RRID, then data would be more trackable," Bandrowski says. She collaborated with others to run a pilot project through the Resource Identification Initiative—a Working Group that is linked to FORCE11—to test that idea. They convinced 25 editors in chief of scientific publications to require authors to use RRIDs (unique identifiers for the reagents, tools, data, software and materials used to perform experiments) in the methods section of their research papers. "Journals were willing to buy in because they care about reproducibility," Bandrowski says. It worked: Authors obtained the required IDs, especially when the journal editors were persistent about checking the IDs; and additional publishers signed on to the requirement.

Why was it successful? Bandrowski has a theory:

RRIDs are kind of like an H index—the system used to measure an individual's impact as a researcher. They offer a way to give credit to researchers for creating datasets and software.

The RRID system is also robust, with a strong registry stamp behind it as well as long-term financial backing, Bandrowski says. "There has to be a living entity that takes care of these things," Bandrowski says. The RRID is an accession number that points to a registry page that lists the digital object's funding, description, and people in charge. Automated checkers determine if a link is dead or live. "If it's down for two to three weeks then a human looks," Bandrowski says. If it's permanently gone, then the registry page is changed to say that—"so you don't get the 404 error message," she says.

Members of FORCE11 are still laying the groundwork for data referencing to be done consistently across all the different journals. "We're trying to see if all you need is a number but there may be other things that would make it a lot easier in the future," Bandrowski says. "We're bringing people together to see what they come up with."

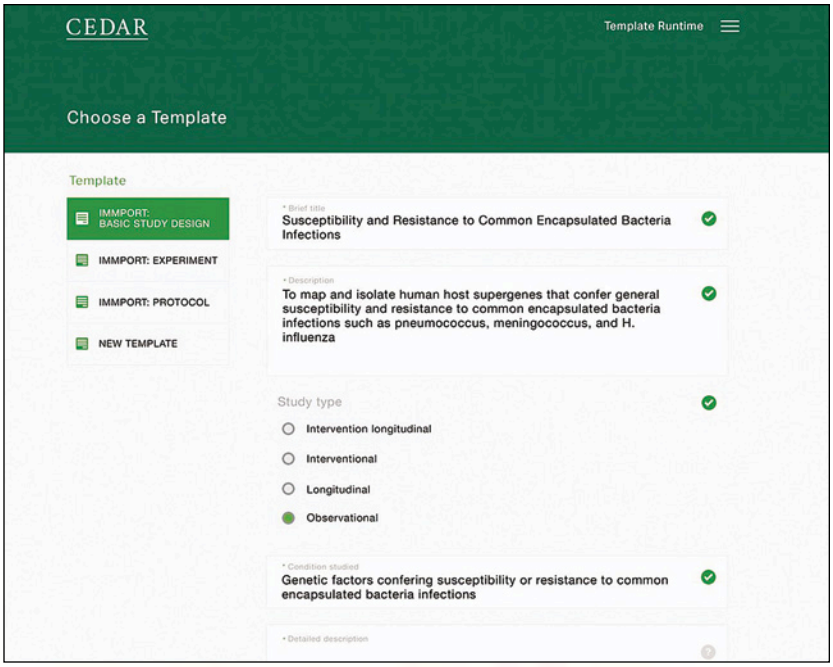
Where Did I Come From?
Metadata Made Simple

A data identifier allows a dataset to say, in essence, "Here I am, I am unique." But it doesn't describe what the researchers did to gather the data. What laboratory procedures did they use? What machines took the measurements, determined sequences, or collected images? What do certain data fields or acronyms mean? All of that is opaque to the viewer of the data itself unless someone has annotated the dataset with metadata—a detailed but concise description of how the data were collected and what they represent.

But if researchers haven't jumped at getting data identifiers, just imagine their reluctance to create metadata in a standard format. Again, incentives matter. "There's no great reward for doing a good job of annotating the data to be useful for others, says **Mark Musen, MD, PhD**, professor of biomedical informatics at Stanford University. In fact, he says, there's a disincentive—the fear of being scooped, or of others finding results you could have gotten yourself. So what could be done to change that?

Again, publishers are playing a role. "They are trying to be agents of change," says **Susanna-Assunta Sansone, PhD**, associate director of the Oxford e-Research Center at Oxford University. One option is the so-called "data journal," which may take many forms. The open source journal *GigaScience,* for example, requires that all supporting data and source code be publicly available and hosted in the journal's database and cloud repository. And the primary article type in *Scientific Data*, a data journal from Nature Publishing Group, and Elsevier's *Genomic Data*, is a data descriptor, designed to make data more discoverable, interpretable and reusable. Some data journals also include a machine-readable description of the dataset in addition to text. These efforts incentivize researchers to publish clear data descriptions—and since publication remains the currency of science, they also spread a little bit of the wealth to those who gather, curate, and wrangle data.

Funding agencies could also play a role in shifting incentives. "What if, to submit a new grant application you have to document that you did the right things with your old data?" Musen wonders. "That might have some teeth!"



As a user inputs information about a dataset, CEDAR's metadata templates customize themselves to fit the situation. In this prototype user interface, the end user has selected the "ImmPort Basic Study Design" template, and has filled in values in the template's slots for brief title, description, study type, and condition studied. Reprinted with permission from MA Musen, CA Bean, KH Cheung et al., The center for expanded data annotation and retrieval, J Amer Med Informatics Assn *(2015).*

So far, granting agencies haven't taken that approach—yet. The NIH is, however, investing in metadata infrastructure as part of its push for data to live up to the FAIR principles. For example, the Center for Expanded Data Annotation and Retrieval (CEDAR), a Big Data to Knowledge (BD2K) Center of Excellence for which Musen serves as principal investigator (PI), is building tools to streamline metadata creation.

After one year in business, CEDAR has a prototype of a user interface. "We're creating a library of hundreds of templates, each for a specific kind of experiment or experimental subject or specific instrument," Musen says. The templates are designed to incorporate standards established for a particular field—standards that have been curated at BioSharing.org, a registry of more than 600 domain-specific minimal information checklists that is run by Sansone's group at Oxford. But, importantly, researchers will be insulated from the technical details of the standards.

"The idea is that you will be guided," says Sansone, a CEDAR co-investigator. "The system will intelligently create the template, customized to the needs of the researcher and the dataset, with the standards hidden from view."

How Can You Find Me?
Data Discovery via DataMed

Good metadata is a first step toward data discovery. The next is an index and a search engine that can find that metadata in response to a researcher's query.

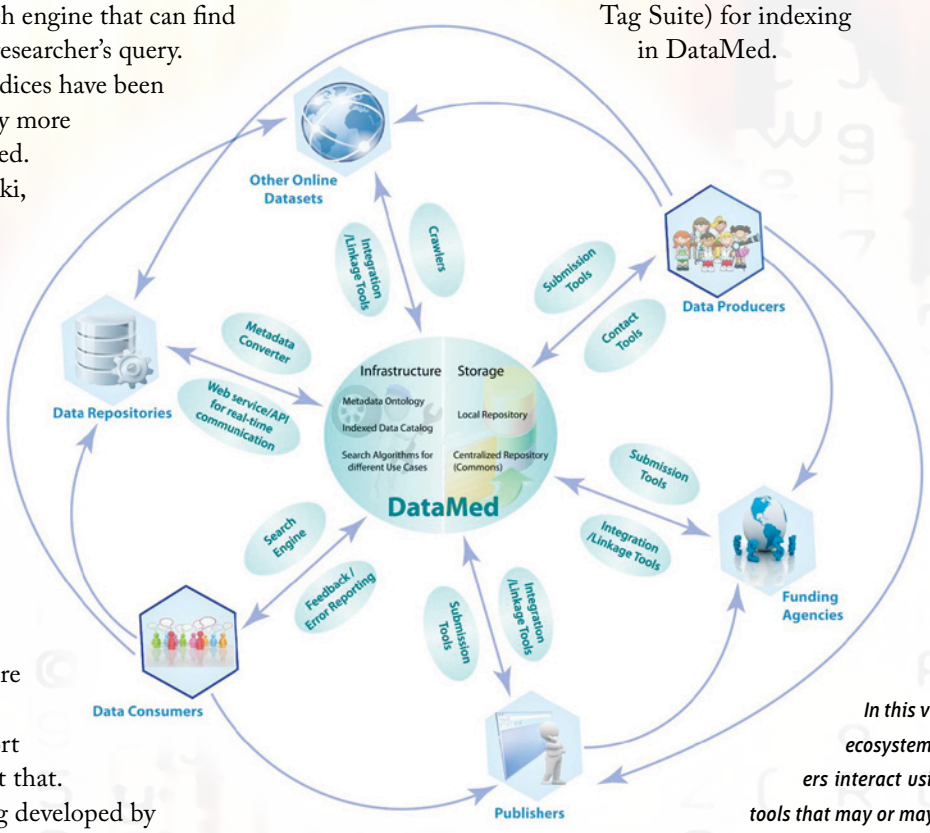
Plenty of domain-specific indices have been created over the years, and many more are still being built and supported. "Unfortunately," says Bandrowski, who helped develop an index called the Neuroscience Information Framework (NIF), "nobody comes to these things." She and her colleagues received positive feedback from researchers whenever they publicized NIF at neurosciences conferences. Yet the next year, they would realize NIF had been forgotten. Her thinking now: "You have to meet the biologists where they are, which is PubMed."

The latest data indexing effort funded by the NIH may do just that. It's a data discovery index being developed by bioCADDIE (biomedical and healthcare Data Discovery Index Ecosystem), a BD2K Center of Excellence. "bioCADDIE is doing for data what PubMed is doing for literature," says Sansone, who is on the bioCADDIE executive and steering committees. "We're calling it DataMed."

Lucila Ohno-Machado, **MD**, **PhD**, professor of biomedical informatics at the University of California, San Diego, who is principal investigator of bioCADDIE, calls the center "an integrator of multiple indexing efforts."

Currently, if researchers go to PubMed to look for research in a specific subject area, they can explore all the publications out there. They cannot, however, go to a single place to explore all the datasets in the world, says Sansone. DataMed will be designed to allow that kind of exploration—with richer filters and data-specific browsing fields than the ones that are currently available in PubMed, she says.

It's an ambitious goal. For the first phase, bioCADDIE has developed a unified way of describing datasets that connects up nicely with the CEDAR metadata templates. In order to map it to the databases that already exist, the team is working with the largest data repository managers. "It needs to become a cultural thing like PubMed indexing," Sansone says. Just as journals create a JATS (Journal Article Tag Suite) file for indexing in PubMed, database creators would create a DATS (Data Tag Suite) for indexing in DataMed.



In this vision of the DataMed ecosystem, multiple stakeholders interact using components and tools that may or may not yet exist.

Don't expect DataMed to allow searches at the level of molecular queries. "It will be able to retrieve datasets or point to another index, but not able to query on gene expression," Sansone says. "You will be able to narrow things down, but you still have to go out to the actual datasets." For example, a researcher might ask for all datasets of Alzheimer's patients that have RNA-seq, behavioral and imaging data available. Or they might ask for all proteomics and metabolomics datasets related to a specific biological process. Or for all data related to the effect of stress on health.

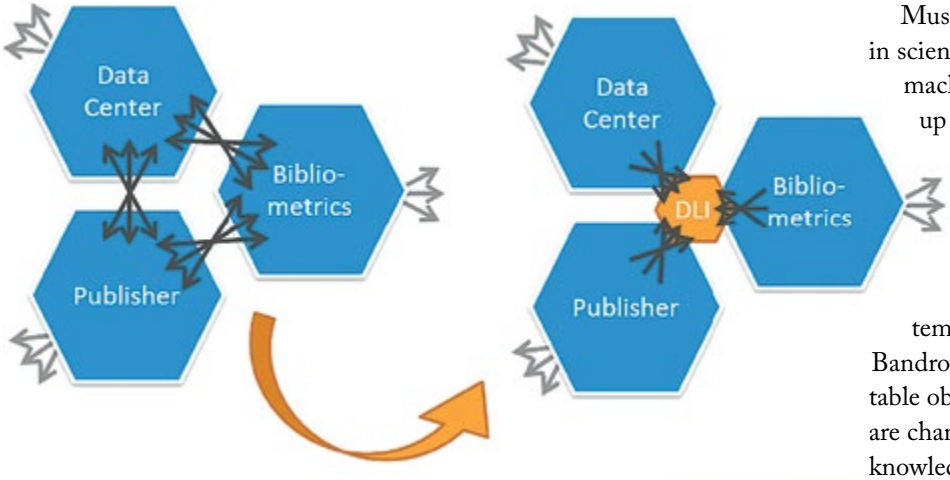
The bioCADDIE data discovery index is very much a work in progress. "It's all in discussion. It's all happening right now," Sansone says. The team expects to release a prototype in the summer of 2016.

Some researchers argue that DataMed should just be inside PubMed. "Scientists live in the literature almost every day," says **Maryann Martone**, **PhD**, past president (2014-15) of FORCE11. Much less frequently, they might be looking for datasets or software programs. "Let's start projecting things into where people actually are as opposed to expecting them to know that we exist," she says.

Why Do I Matter?

Linking Data to Publications

At some point in the future, perhaps datasets will have identifiers and associated metadata, be located at reliable addresses, and be findable in DataMed. But there remains the question of what those data are telling us. What knowledge has already been extracted from them?



The publishing data services working group at the Research Data Alliance is promoting a move from bilateral arrangements between data centers and publishers toward common standards and one-for-all services.

Essentially, datasets need to be linked to publications. All the data identifiers and metadata in the world can't make that happen unless journals require authors to build those links and share their data. As mentioned above, journals have been stepping up their data-linking requirements. But many are not stringent enough about checking that datasets have been submitted to a repository that will live on after the project, Sansone says. "It's a slow process," she says, "but it is happening."

Some publishers and researchers, including a working group at the Research Data Alliance, are also pushing beyond bi-directional linking from data to publications. They'd like to see an overarching service that can combine links from different sources into a common "one-for-all" service model.

Links between publications and data could also cause a beneficial side effect: The ability to give people credit for the value of their data. "If someone generates data that got used 5,000 times or was cited 300 times, there will be a way to recognize that," Bonazzi says.

Some researchers think more significant changes are afoot. "In a recent perspective article in *Nature*, **Philip Bourne**, **PhD**, Associate Director for Data Science at the NIH, and his co-authors wrote: "There is an unnecessary cost in a researcher interpreting data and putting that interpretation into a research paper, only to have a biocurator extract that information from the paper and associate it back with the data. We need tools and rewards that incentivize researchers to submit their data to data resources in ways that maximize both quality and ease of access."

Musen would go further. "Ultimately, publication in science will have to move from prose to something machine processable," he says. "People don't pick up journals anymore and get cozy with them."

While the tools to make this transition do not yet exist in a realistic way, ideas along these lines have been percolating for a long time, especially within FORCE11.

One reality of the current publication system is its inability to deal with change over time, Bandrowski notes. "Essentially, you have these immutable objects [papers] that are referencing things that are changing all the time," she says. Databases grow, knowledge shifts, but papers remain static. There's a disconnect between "the flowing river of the Web and these stable objects that are publications—like rocks in that river," she adds.

Martone agrees: "The minute you publish something or put a dataset out there, there's already something you can say that you didn't say." These days, she's working on an effort to allow instantaneous annotation of anything on the Web using an open

source tool created by Hypothes.is, a nonprofit for which Martone serves as director of biosciences and scholarly communications. Upon selecting text, on an existing Web page, users of the plug-in open a dialog box where they can enter whatever they want—a hyperlink, updates, additional information, tags. "It gives you the capacity to create searchable knowledge," Martone says. She thinks the plug-in can help fix some of the structural problems in biomedicine. For example, it allows people to open up independent communication to update the literature. "Hypothes.is allows scaling of content at the time rate that science happens," she says.

Martone imagines that eventually the links to and from various updates and tags will be data themselves. "We have to be able to read these signals much as Google reads the signals of links. We just have to figure out what those signals mean."

For now, users can install the Hypothes.is plug-in in their browser. "It should be built into everything we have," Martone says. "We'd love it built into PubMed and other browsers." In fact, Hypothes.is is organizing a new coalition: Annotating All Knowledge (https://hypothes.is/annotating-all-knowledge/), that is bringing together publishers and other stakeholders to bring this capacity to all scholarship. "The challenge now is letting people know these capabilities exist," she says.

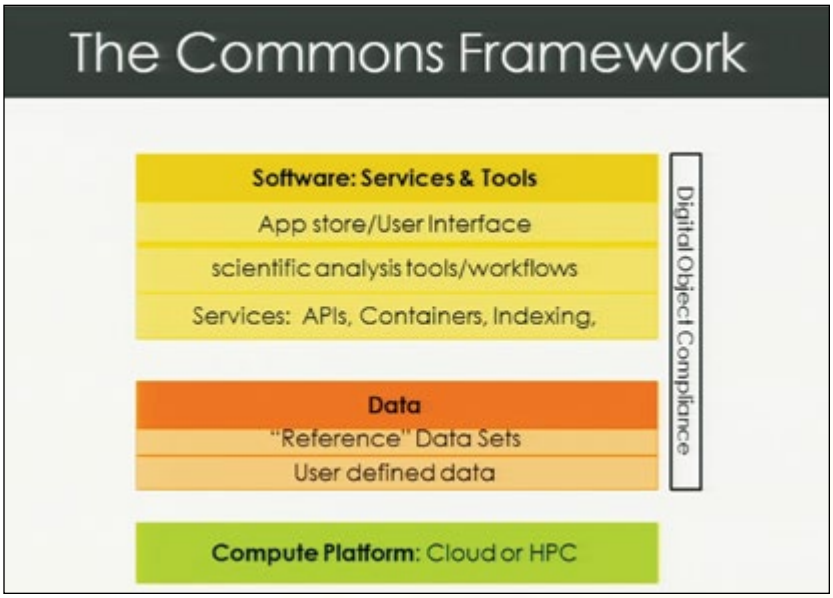
The Commons
An Ecosystem for Data Sharing

In October 2014, Bourne announced plans to create the "NIH Commons" to catalyze the sharing, use, reuse, interoperability and discoverability of shared digital research objects, including data and software.

Bonazzi diagrams the Commons as a layered system consisting of three primary tiers: high performance and cloud computing (at the bottom); data, including both reference datasets and user-defined data (in the middle); and (at the top) services and tools, including APIs, containers, and indexes (DataMed, for example), as well as scientific analysis tools and workflows and—eventually—an app store and interface designed for users who are not bioinformaticians.

To be eligible for use in the Commons, data and software will have to meet the FAIR principles. To make that easier for researchers, the products of all of the BD2K centers will be part of the Commons ecosystem, including DataMed from bioCADDIE and streamlined metadata templates from CEDAR. And to incentivize participation in the Commons, the NIH plans to offer cloud computing credit vouchers that researchers can use with a provider of their choice, so long as the provider complies with the FAIR principles.

The Cloud Credits Model, as it's being called, "democratizes access to data and computational tools," said **George Komatsoulis**, **PhD**, (acting) chief of the informatics resources branch at the National Center for Biotechnology Information (NCBI), when he spoke at the BD2K All Hands Meeting. Right now, researchers access cloud computing with a credit card or through a university, he said. Komatsoulis anticipates that the voucher system will be more cost effective by creating a competitive marketplace for biomedical computing services and reducing redundancy. The voucher system is now being piloted in specific research areas—the Genomic Data Commons, for example. "Credits will be distributed the way the National Science Foundation distributes access to specific facilities such as light sources," Komatsoulis said. "But having an existing NIH grant will be a precondition."



Current plans for the NIH Commons would bring together high performance and cloud computing with data, services and tools. Courtesy of Vivien Bonazzi, NIH.

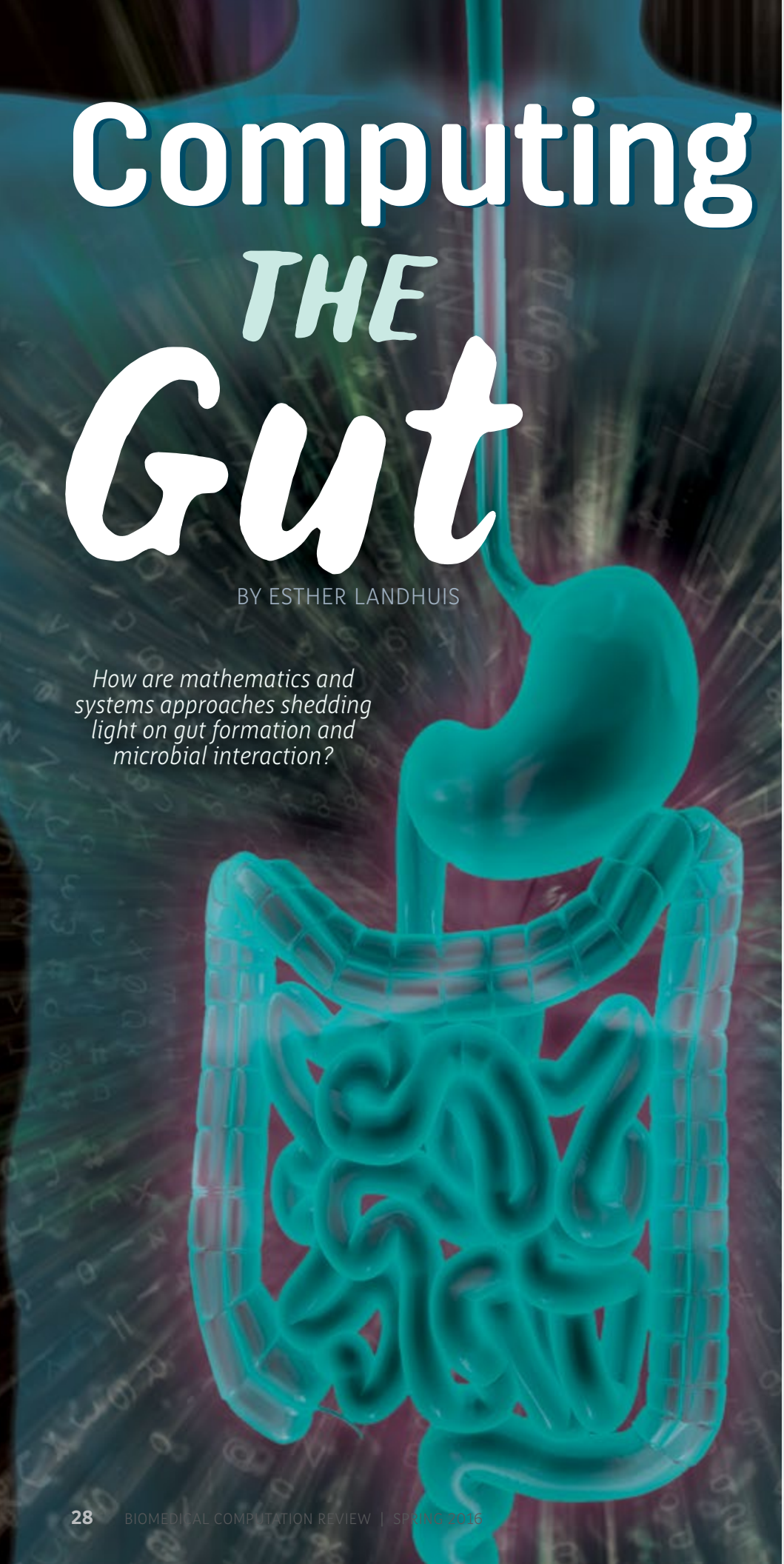
With the Commons, the NIH is feeling its way toward a viable ecosystem for the sharing of big data. "We're testing pieces of the ecosystem out," Bonazzi says. "Does this make sense? What are the pieces that are missing? What still needs doing? And how do we facilitate the community to do those?"

The NIH doesn't want to be in a position of saying here's *the* infrastructure. "That's not going to work," Bonazzi says. "I'm not claiming this is it. I'm saying this is what I'm seeing the community is doing. If this is one step toward coalescing a concept around how we do biomedical science in the future, and that's useful, then let's use it as a point of discussion." □

Computing THE Gut

BY ESTHER LANDHUIS

How are mathematics and systems approaches shedding light on gut formation and microbial interaction?



The heart holds a special place in human history and literature, and the brain may be the organ we most associate with a sense of self. But the proverbial seat of wisdom—the gut—deserves reverence, too.

It is an architectural wonder buzzing with activity. A 20- to 40-foot tube with many tight bends and folds, the gut houses trillions of bacteria working in cahoots with our own cells to extract energy from food and maintain health.

How does this long tube cram inside the belly without becoming a tangled mess? Why doesn't food get stuck in there? And what about all those bacteria? How do they work with gut immune cells to keep us from getting sick?

The sheer complexity of the gut—and the finger-like projections called villi that line the intestinal tissue—is inspiring some scientists to explore how physical forces, such as changes in stress or geometry, influence how the gut is formed. In addition, a growing suite of mathematical models and computational tools is offering insight into how immune cells within this engineering wonder interact with native bacteria and foreign pathogens to regulate health.

Gut Formation: Loops, Wrinkles and Folds

So how does nature design a gut?

"If you took a garden hose and randomly folded and packed it, you would form kinks—and this would be problematic," says **Thierry Savin, PhD**, a biophysicist at the University of Cambridge. "Yet nature has designed a smart, elegant way to put regular loops in the gut without forming kinks."

As a former postdoctoral researcher at the School of Engineering and Applied Sciences at Harvard University, Savin collaborated with Harvard mathematician **Lakshminarayanan Mahadevan, PhD**, and developmental geneticist **Cliff Tabin, PhD**, to explore the physics behind this amazing feat.

Tabin and colleagues had examined dissected embryos of chicks, quails, finches and mice, and seen that the gut forms loops that are strikingly similar in number, size and shape across species. When his team surgically separated the gut from the rest

of the embryonic tissue, the loops remained intact. However, if they cut the gut tube away from its attached membrane, the looping structure disappeared—the tube relaxed into a straight configuration and the membrane shrank. The big question, says Savin, was, "How do you form this shape? What is the strategy nature uses to make the loops?" Could it be that the tube grows faster than the attached membrane, which gets stretched and forces the gut to coil?

Experiments with common lab materials gave the team a sense for how this might play out. They stitched a straight rubber tube to a stretched latex membrane, then let the structure relax. It spontaneously adopted a helical pattern that looks like the biological gut. What happens at the scale of a single loop is the same as what happens with a taut bow. "If you cut the string, it becomes straight," says Savin. "This convinced us that elastic forces originating from differential growth between the tube and membrane are responsible for shaping the gut."

Savin and his colleagues produced a graphical simulation of gut looping in a chick embryo using a model based on geometry, the mechanical properties of the tissues, and the relative growth rate of the gut tube and the mesentery (bottom). The simulation compared favorably with both the rubber model (middle) and an actual chick gut (top). Image courtesy of T Savin and A Shyer.

Further experiments with the rubber-latex structure helped the researchers work out mathematical equations to account for altering specific parameters—for instance, membrane stiffness, tube size and radius—to produce distinct looping patterns in the gut. The team made similar measurements in gut tissue from chick, quail, finch and mouse embryos at various stages of development to refine and confirm their mathematical model.

More recently, Tabin, Mahadevan and colleagues extended their modeling to incorporate genetics. In a 2015 *Cell* paper, the researchers report how mechanical forces in the developing gut activate molecular signals that position intestinal stem cells at the base of villi, where they give rise to the other cell types in the gut lining.

Another group of interdisciplinary researchers has also used mathematics and computational tools to examine gut formation. However, rather than study the looping structure of the gut, they focused on the formation of epithelial patterns during embryonic development of the gut's inner layers—the endoderm and mesoderm—by modeling them as concentric tubes. The work by **Pasquale Ciarletta, PhD**, an applied mathematician at the Université Paris 6 and Politecnico di Milano, **Valentina Balbi, PhD**, Ciarletta's graduate student at the time, and **Ellen Kuhl, PhD**, a bioengineer at Stanford University, was published in December 2014 in *Physical Review Letters* and February 2015 in the *Journal of the Mechanics and Physics of Solids*. Their model explains how the tubes' elastic and geometric properties influence wrinkling and folding patterns in the epithelia of the esophagus, intestines and other gastrointestinal tissues—traits that contribute not only to development but also disorders of the intestines, such as food allergies.

The team started by collecting existing experimental measures of the thickness, elasticity and growth rate of the gastrointestinal tract of chick and turkey embryos at different stages of development. From these measurements they calculated parameters that drive key pattern transitions during development. For example, certain geometric and mechanical properties triggered development of ridges on day 13 and caused villi to form on day 14, Ciarletta says. The model explained how the esophagus develops longitudinal folds with a thick and stiff outer layer, while circumferential folds emerge in the jejunum with a thinner and softer outer layer.

Researchers can use the model to explain and predict changes in gut morphology that lead to digestive disorders. In people with food allergies, for example, local inflammation can cause atypical wrinkling that is a hallmark of disease.

Insights from modeling point toward potential treatments that tweak the tissue's mechanical properties—for example, osmotic drugs to restore the homeostatic condition, Ciarletta says.

Gut Microbiome Variation

Moving beyond architecture, some scientists are developing computational methods to survey the constituents of the gut—specifically, its cells and microbes. Our bodies have about as many microbes as cells, and microbiomes vary dramatically between individuals. With advances in genomic sequencing and analytical methods, researchers have compared samples of gut bacteria from different people and found vast differences in which species are present and which genes they encode.

Research suggests that microbiome variation may influence many aspects of health. Gut bacteria shape immune system development and can affect how well we digest certain foods and how easily we gain weight, research suggests.

Yet gut microbes aren't the whole story. "It's not only which players are there but how they interact with each other and with the host. It's important to study [the gut] as a complex system," says **Elhanan Borenstein, PhD**. A computational biologist, Borenstein runs a lab in the Department of Genome Sciences at the University of Washington in Seattle. His group hopes to gain an improved, systems-level, mechanistic understanding of the microbiome using systems biology approaches and computational modeling.

One question that intrigued **Sharon Greenblum**, **PhD**, during her graduate studies in the Borenstein lab, was the extent to which the gut microbiome varies across individuals at the strain level. This information could be important because different strains of the same species of bacteria could encode different genes and may therefore perform different functions in the gut. They might also have more or fewer copies of particular genes.

Many studies of the gut microbiome use methods that are not sensitive enough to characterize the bacteria at the strain level. To study strain-level differences, Borenstein and Greenblum, who is now an evolutionary genetics postdoc at Stanford, used a different approach. Their method involved sequencing short stretches of DNA in the sample and counting how many map to a specific gene in a particular species. They used this method to analyze gut microbiome data from previously published studies of fecal samples from healthy, obese and inflammatory bowel disease (IBD)–afflicted people.

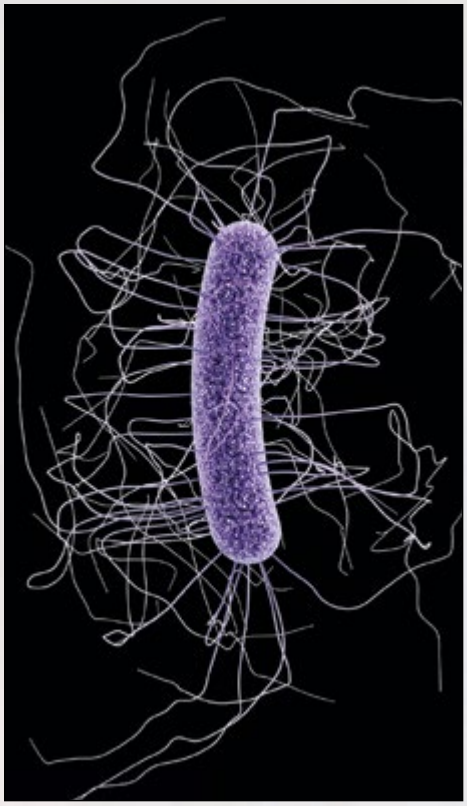
First they had to determine whether a gene was more abundant in a particular individual simply because the sample contained a greater number of species each encoding the gene, or because that individual's strain of the particular species contains more copies of the gene. Indeed, the team wondered: When comparing individuals with the same bacterial species, could one person's strains have more copies of a certain gene while another person's strains have fewer copies?

To address these issues, the researchers developed algorithms to map each shotgun sequence to the bacterial genome it came from and determine— for each sample, and for each bacterial genome in that sample—the copy number for each gene. Next, they compared between samples, asking if the copy number of gene X in species Y is the same as it is in other samples. The goal was to identify cases where a specific gene in a particular species is present in different numbers of copies across individuals.

As reported in February 2015 in a *Cell* paper, there was "tremendous variation" among individuals, Borenstein says. For most species analyzed, individuals had copy number differences in many genes. Moreover, for some genes, one person could have a single copy while another had 15.

"We were surprised by the amount of variation," Borenstein says.

And copy number variation did seem to impact function—particularly for genes associated with responding to the environment, such as those encoding proteins that transport metabolites in and out of cells. This makes sense: In a nutrient-poor environment, a higher copy number of a specific set of transporters might be advantageous for feeding the cell, Borenstein notes.



This illustration, which is based on photomicrographic data, depicts the morphology of a single Clostridium difficile *bacillus, a common cause of antibiotic-associated diarrhea. Over the past several years nationwide, states have reported increased rates of* C. difficile *infection, as well as more severe disease symptoms and an associated increase in mortality. Credit: Centers for Disease Control/James Archer.*

By quantifying the extent to which the gut microbiome varies between individuals, Borenstein and his colleagues have taken a first step toward the lab's eventual goal: personally tailored interventions. "We want to be able to design a specific perturbation to create a specific phenotype," Borenstein says.

Today, to coax a patient's microbiota toward a healthy microbome composition, physicians use fecal transplantation: They take a stool sample from a healthy person, "transfer it into a diseased person and hope it works," Borenstein says. But Borenstein's lab is striving for rational design rather than trial and error. And to achieve that, he says, "We need to build accurate, mechanistic models of the microbiome."

Tackling Superbugs

Broadening the analyses further, some research groups are building computer models to study dynamic interactions within the gut—not just among microbes but also the immune cells that live and work alongside. The body's ability to fight dangerous pathogens depends on coordinated interplay between microbial and immune systems, each consisting of diverse cell types. Sometimes a menacing microbe can throw this network out of whack. One such culprit is *Clostridium difficile*—a "superbug" that infects some 600,000 people in the US each year, killing 29,000. Healthcare costs associated with *C. difficile* infections top $3.2 billion. Worse yet, these numbers are on the rise.

So it may be disconcerting to learn that *C. difficile* are actually found everywhere. They can even live in a normal human gut—though "usually in low quantities and kept in check by good bugs," says **Steven Steinway**, **PhD**, an MD/PhD candidate at Pennsylvania State University working with biophysicist **Reka Albert**, **PhD**, also at Penn State, and biomedical engineer **Jason Papin**, **PhD**, at the University of Virginia.

C. difficile only becomes a problem when antibiotics prescribed to fight one infection deplete the body of other bacteria, many of them beneficial. That gives superbugs a chance to grow and dominate—which calls for another round of therapeutics. "What's ridiculous is that *C. difficile* infection (CDI) is caused by antimicrobial treatment, yet the treatment for CDI is another set of antimicrobials," says **Josep Bassaganya-Riera**, **DVM**, **PhD**, director of the Nutritional Immunology

and Molecular Medicine Laboratory (NIMML) at Virginia Tech in Blacksburg. "There is an unmet clinical need for safer and more effective therapeutics for CDI, and modeling can accelerate the development of such new treatments."

As described below, these research teams are using computational tools to find strategies for tackling *C. difficile* infection—one focused on finding good bacteria to do the job and the other focused on boosting immune defense.

Beneficial Bacteria

As reported in *PLoS Computational Biology* in June 2015, Steinway and his collaborators modeled metabolic interactions in the gut microbiome in order to identify specific bacterial strains that act to suppress *C. difficile* growth. His team hopes the insights will lead to the development of probiotics to supplement conventional antimicrobials for people battling a CDI.

Steinway says the model views the intestinal community as an ecological niche—sort of like a rainforest—with diverse organisms that interact in predator-prey relationships. However, in a microbial community, the bacteria are not necessarily preying on each other but "produce chemicals that can help or suppress the growth of other bacteria," Steinway says.

His team's mathematical model was built from mouse data showing that treatment with the antibiotic clindamycin makes animals more susceptible to *C. difficile* infection relative to untreated controls. The researchers measured quantities of different bacteria in the mouse gut and monitored changes in these populations over time.

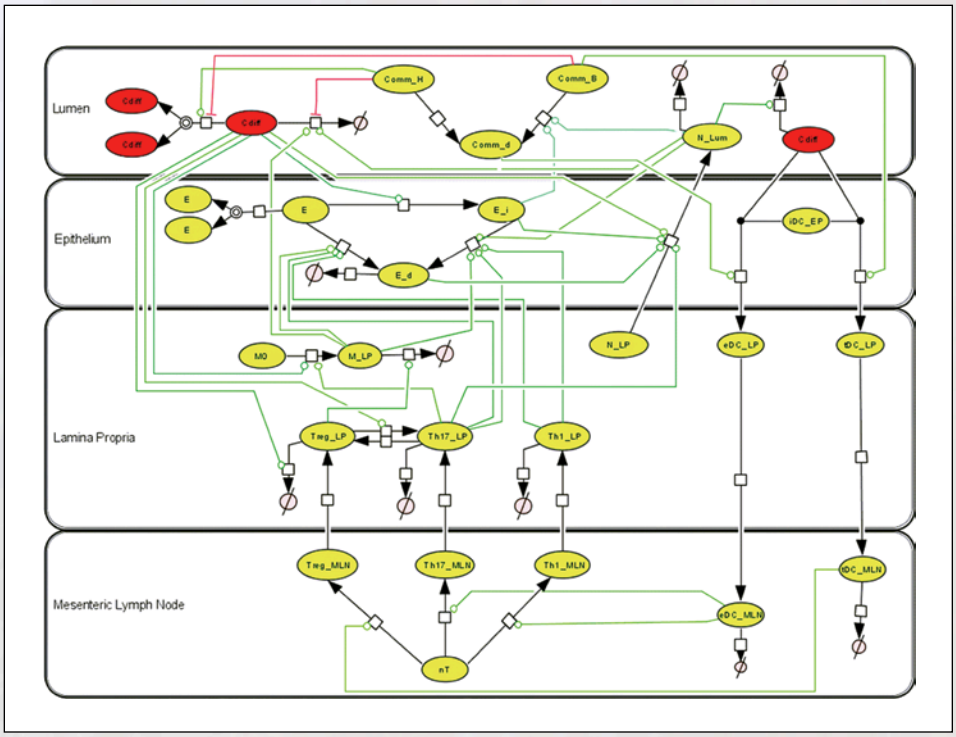
To model cause-and-effect relationships among bacteria, the team used a binary approach: For each timepoint in the mouse data, the researchers determined which bacteria were present and which were absent. These data were then crunched by machine learning algorithms to reveal which strains were likely activating or inhibiting other bacteria.

Their model identified a strain of normal gut bacteria, *Barnesiella intestinihominis*, that inhibits *C. difficile* growth—a result that has been confirmed by lab co-culture experiments. If the tests pass muster in mice, the team hopes to move toward human trials of the probiotic.

An Immune Boost

Rather than identifying good bugs to counteract *C. difficile*, Bassaganya-Riera's team sought to understand how to help specific immune cells do a better job of keeping the superbug in check. To figure that out, they needed to know how *C. difficile* disrupts the balance between the branch of the immune system that promotes inflammation (the effector branch) and the regulatory branch that suppresses it. This question is hard to address with traditional experimental approaches because the relationships among the players are networked rather than uni- or bi-directional.

The team therefore turned to computational modeling to explore interactions between pathogens and the host's gut bacteria and immune cells. They began by modeling a network of 23 interacting entities across the four-part architecture of the gut's mucosal immune system: the lumen—the inner part of the intestine—where beneficial bugs and pathogens are located; the epithelium—the layer of cells that separates the body from its external environment; the layer beneath that, called the lamina propria, where most immune cells reside; and lymph nodes, where immune reactions begin. To build and calibrate the model, they used data from immune cell populations analyzed individually in *C. difficile*–infected mice over the course of an infection. The model relies on ordinary differential equations to describe the cell dynamics during



Bassaganya-Riera and his colleagues created a network model of the gut immune response to Clostridium difficile *infection (*CDI*) across four compartments of the intestinal mucosa (black boxes) as diagramed here. In the model,* C. difficile *interacts with other bacteria as well as immune cells in various ways. For example, interactions could activate the bacterium to start proliferating—or inhibit or kill the bacterium. Alternatively, interactions could modify various other reactions among the participants. Species include* C. difficile *(Cdiff, in red), infection-exacerbating commensal bacteria (CommH), protective commensal bacteria (CommB), dead commensal bacteria (CommD), epithelial cells (E), inflamed epithelial cells (Ei), neutrophils (N), macrophages (M), dendritic cells (tDC and eDC), T cells (nT, Treg, Th17, Th1) existing in multiple compartments: lumen (Lum), epithelium (EP), lamina propria (LP), and mesenteric lymph node (MLN). Reprinted from A Leber, M Viladomiu, R Hontecillas et al., Systems Modeling of Interactions between Mucosal Immunity and the Gut Microbiome during Clostridium difficile Infection,* PLoS One, *DOI:10.1371/journal.pone.0134849 (2015).*

the infection as well as the effect of the bacteria-killing chemicals some of the cells were producing.

And as it turns out, churning out such chemicals—or antimicrobials—wasn't necessarily a good thing. Secreting more bacteria-killing compounds did not dampen CDI but rather sustained it by preventing regrowth of beneficial bacteria that could have quashed the superbug. "A significant amount of damage during CD infection is not caused by the pathogen itself but rather by the overzealous host immune response," Bassaganya-Riera says.

Published in July 2015 in *PLoS ONE*, the model is steering the researchers' attention toward therapeutic responses that manipulate the host rather than the bacterium. The goal: to allow the host immune system to co-exist with bacteria such as *C. difficile*, Bassaganya-Riera says.

His team will now begin testing the model's predictions in his multidisciplinary lab. "We have computer scientists, mathematicians, and physicists but also immunologists and lab technicians," Bassaganya-Riera says. Penn State's MD/PhD program and Virginia Tech's NIMML create "researchers who can navigate the interface between experimental and computational work—that is, spend the morning writing code and in the afternoon perform studies in mice or analyze clinical specimens."

Gut Tissue Modeling

Gary An, MD, associate professor of surgery at the University of Chicago, also straddles multiple disciplines. An trained as a trauma surgeon in the mid-1990s but grew frustrated by decades of failed attempts to develop treatments for sepsis—a life-threatening illness caused by disordered systemic inflammation. Around that time he learned about complexity science and agent-based modeling, an emerging approach for studying systems with interacting components that can behave in unexpected ways.

In the entertainment industry, such models are used to create virtual worlds in video games and movies—for example, battle scenes in *Lord of the Rings*—where individuals operate under similar guidelines yet behave differently moment to moment, leading to unanticipated outcomes for the group. An considers cellular interactions within the human gut an analogous situation. "One of the huge advantages of agent-based models is the ability to construct spatial representations that look real," An says. "This is why it's used for battle scenes in movies. That's why it's used to model birds flocking and traffic and things that have a spatial pattern to them."

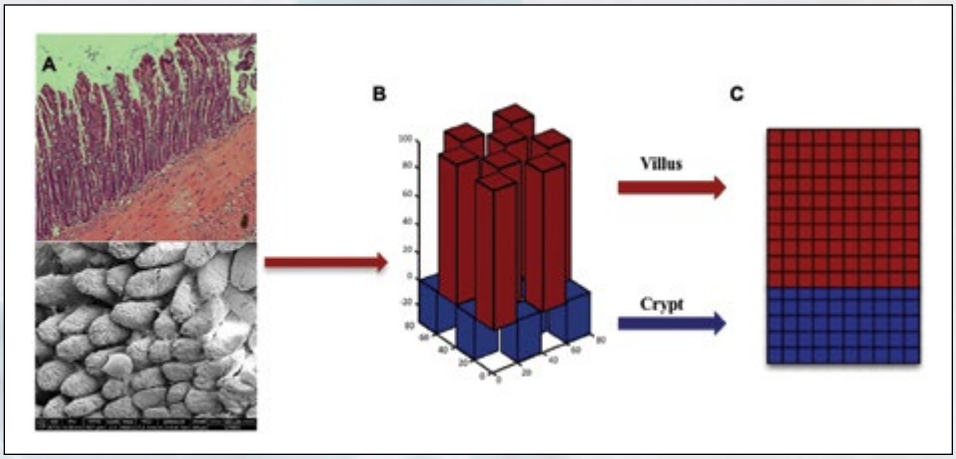
Just as birds arrange into a flock, "Tissue forms a certain structure because of the cells' interactions," An says. "My emphasis on the models is their ability to generate tissue architecture." This is important because histology—the study of tissue slices under a microscope—is a primary means by which physicians diagnose and characterize disease.

Consider for example ulcerative colitis, a disease in which the gut surface becomes unusually sensitive, leading to dysregulated inflammation and painful ulcers in the digestive tract. Some cases are treated by removing the colon and folding a piece of the small intestine to form a stool-collection pouch. The problem is, "the pouch can become inflamed and make people sick," An says.

An's team suspected that the resulting condition, called pouchitis, is caused by inflammatory signaling from high bacterial counts in the accumulating stool. If so, clues to detect the transition from normal to pathological could appear as shifts in the gut's tissue architecture. In March 2014, An and colleagues published a *PLoS Computational Biology* paper that describes their Spatially Explicit General-Purpose Model of Enteric Tissue (SEGMEnT). The model incorporates existing knowledge of how gut epithelial cells behave and respond to inflammation.

An's team has since harnessed a supercomputing version of this knowledge-based model to characterize the clinical trajectories of individual patients. As reported in March 2015 in *PLoS ONE*, the researchers calibrate the model with data from a clinical trial on patients with pouchitis to see if certain features of their model have predictive power—to determine, for instance, at what point physicians should consider putting patients on antibiotics to hold off development of pouchitis.

An's overarching goal is to develop models that describe how an individual will behave over time and explain how a particular trajectory could be changed. "In medicine it's not sufficient to just prognose and diagnose. We want to be able to control what's going to happen to you," An says. "Models like this can provide that answer." □



Gary An's agent-based models reflect the physical form of the intestinal tissue they are modeling. Panel A shows a histological cross section of ileal tissue (top) and a scanning electron microscopy image of the mucosal surface of ileum (bottom), while panel B shows the topology used in An's model, with crypts and villi represented by a matrix of rectangular prisms. Each individual crypt or villus is then "unwrapped" onto a 2-dimensional grid (Panel C), on which signaling interactions, morphogen diffusion and physical cellular actions take place. Reprinted from C Cockrell, S Christley, G An, Investigation of Inflammation and Tissue Patterning in the Gut Using a Spatially Explicit General-Purpose Model of Enteric Tissue (SEGMEnT), PLoS Comp Biol *doi:10.1371/journal.pcbi.1003507 (2014).*

BY XIAOQIAN JIANG, PhD, SHUANG WANG, PhD,
STEPHANIE FEUDJIO FEUPE, LUCILA OHNO-MACHADO, PhD

PRIVACY-PROTECTING ANALYSIS OF DISTRIBUTED BIG DATA

A practical solution for sharing patient data while maintaining privacy protections.

Large clinical data research networks (e.g., PCORnet, HMORnet, ESPnet) have been established to accelerate scientific discovery and improve health. However, a big barrier to making full use of clinical data is the public's concern that researchers' access to demographics, diagnostic codes, genome sequences, etc., can pose risks for individual privacy, with potential implications for employment, security, and life and disability insurance. The current practice of "de-identifying" records before sharing them has limitations, including the likelihood of re-identification [1]. A better approach is to protect the privacy of patients involved in the study by controlling access to patient-level data in a way that respects their preferences while also facilitating research. This can be accomplished using customized distributed protocols that perform specific data analyses while storing and exchanging aggregated patient data through a trusted authority (TA)—an entity that can be trusted not to snoop into the data of the various parties.

Many existing algorithms for privacy-preserving distributed data analysis provide feasible but impractical solutions because they either involve very heavy computation (e.g., homomorphic encryption—computing on encrypted data) or introduce noise (e.g., methods based on differential privacy [2]). Other algorithms may have reasonable performance in a two-party setting but do not scale well to multiple parties.

One pragmatic and efficient framework for constructing accurate multivariate predictive models without ever exchanging patient-level data is Secure Multiparty Computing (SMC) with a TA. For example, biomedical computing centers with private HIPAA compliant clouds, such as iDASH (Integrating Data for Analysis, Anonymization and Sharing) [4] can offer themselves as a TA with whom authorized researchers can collaborate on distributed data analysis.

In this framework, local parties compute intermediary partial results (e.g., sufficient statistics, kernel matrices, etc.) and leverage the TA to combine partial results and coordinate iterative computation. This combination of partial results may be as simple as calculating a global average using partial averages and counts received from the parties, or may require the decomposition of algorithms in a way that allows combination of partial functions by the TA—for example by developing a distributed version of the Newton-Raphson algorithm [3]. Using a hub-and-spoke structure, local parties need only exchange information (at an aggregated level) with the TA through secure channels (such as through Secure Sockets Layer (SSL) in HTTPS) and always keep their patient-level data private.

This strategy has proven effective for a large family of data analysis models (including various generalized linear models and survival models) using different sets of patient data distributed across different servers [3] as well as using different sets of variables from the same patients distributed across different servers [5]—for example when patient phenotypes are hosted at a medical center and their genomes are hosted at a sequencing facility.

In addition to being privacy-protecting, this framework is efficient because the computation (1) is essentially parallelized (similar to the well known Map-Reduce architecture); (2) is amenable to optimization strategies such as prioritizing memory consumption or communication overhead; (3) creates a central point for building models, posing queries, and monitoring activities; (4) limits the need for communication between parties; and, perhaps most importantly, (5) ensures the reproducibility of experiments. □

REFERENCES:

1. Vaidya J, Shafiq B, Jiang X, *et al.* Identifying inference attacks against healthcare data repositories. In: *AMIA Summits Transl Sci Proc*. San Francisco, CA: 2013. 1–5.
2. Dwork C. Differential privacy. *Int Colloq Autom Lang Program* 2006;4052:1–12.
3. Wu Y, Jiang X, Kim J, *et al.* Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *J Am Med Inform Assoc* 2012;2012:758–64. doi:10.1136/amiajnl-2012-000862
4. Ohno-Machado L, Bafna V, Boxwala A a, *et al.* iDASH. Integrating data for analysis, anonymization, and sharing. *J Am Med Informatics Assoc* 2012;19:196–201.
5. Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L, VERTIcal Grid lOgistic regression (VERTIGO), *J Am Med Inform Assoc* 2015;doi: 10.1093/jamia/ocv146.

DETAILS

Lucila Ohno-Machado is a professor of medicine and chair of the department of biomedical bioinformatics, and Xiaoqian Jiang and Shuang Wang are assistant professors in the biomedical informatics department at the University of California, San Diego, School of Medicine. Stephanie Feupe is a graduate student working in Ohno-Machado's lab.

DIVERSE DISCIPLINES, ONE COMMUNITY
BiomedicalComputation REVIEW
Published by the Mobilize Center, an NIH Big Data to Knowledge Center of Excellence

Nonprofit Org.
U.S. Postage Paid
Permit No. 28
Palo Alto, CA

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

SeeingScience

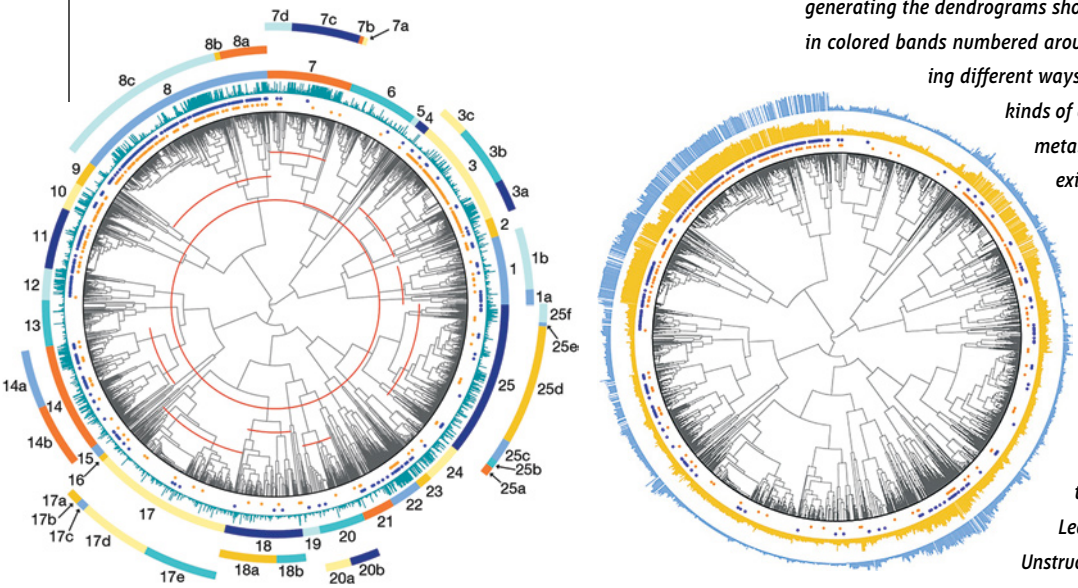
BY KATHARINE MILLER

AUTOMATING LITERATURE SURVEILLANCE

Today, if researchers want to study complex relationships among genes, diseases and drugs, they have to hope that human curators have read the scientific literature, extracted the relevant information, and put it in a database. "It would be a lot more efficient if computers could perform that surveillance of the literature for us," says **Beth Percha**, a graduate student working with **Russ Altman, PhD**, at Stanford University.

In recent work, Percha and Altman made steps toward that goal, effectively extracting drug-gene relationships from the literature and clustering them in ways that proved meaningful (see dendrogram caption). Percha is also applying the same method to other situations such as gene-disease and disease-drug relationships. Ultimately, she'd like to be able predict drug-drug interactions based on drug-gene relationships automatically extracted from the literature.

"The dendrogram is pretty and it's a good sanity check because it reproduces knowledge we already have," Percha says. "But what's exciting is to be able to discover new relationships from the literature quickly, cheaply and without a ton of human effort." □



For 3,514 drug-gene pairs that co-occur at least five times in Medline sentences, each represented as a black dot at the edge of the black circle, Percha and Altman used a novel algorithm that recognizes when two such pairs share a similar relationship. They then used a clustering algorithm to connect drug-gene pairs that act similarly, generating the dendrograms shown here. The clusters revealed 25 "themes"(shown in colored bands numbered around the outside of the circle at far left), representing different ways that drugs interact with genes, such as by various kinds of activation (13-14), inhibition (8, 11) or an effect on metabolism (3). These concurred with information from existing knowledgebases including DrugBank (blue dots) and PharmGKB (orange dots) while also discovering many new relationships that likely should be included in those knowledgebases, as shown in the smaller dendrogram at near left (blue spikes predict drug-target relationships that should be in DrugBank, and orange spikes predict relationships that should be in PharmGKB because mutations in the gene likely impact a person's response to the drug). Reprinted from B Percha, RB Altman, Learning the Structure of Biomedical Relationships from Unstructured Text, PLoS Comp Biol, *11(7):e1004216 (2015).*