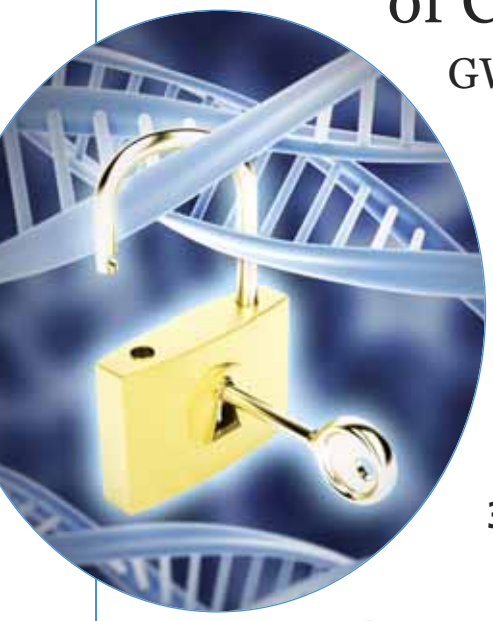
DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation REVIEW

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

NIH Launches A United Ecosystem FOR BIG DATA

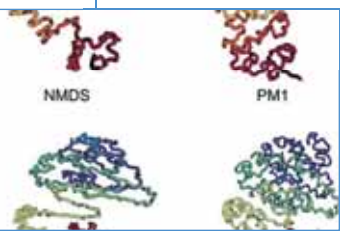
ContentsWinter 2014/2015

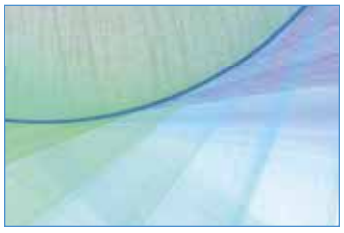


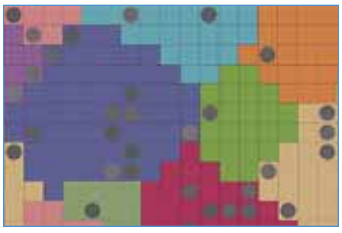
DEPARTMENTS

Cover and Page 13 Art: Created by Rachel Jones of Wink Design Studio using cloud image, © Elenabsl | Dreamstime.com, and circuitry background, © Sailorman | Dreamstime.com.

Page 22 Art: Created by Rachel Jones of Wink Design Studio using DNA background, © Jesper | Dreamstime.com, and padlock image, © Mile Atanasov | Dreamstime.com.







Winter 2014/2015
Volume 11, 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, Kristin Sainani, Katharine Miller, Sarah C.P. Williams

Community Contributors
Philip Bourne, PhD, Madeleine Udell, Stephen Boyd, PhD

Layout and Design
Wink Design Studio

Printing
Advanced 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 S271
Stanford, CA 94305-5444

Biomedical Computation Review is published by:

mobilize
The Mobilize Center, an NIH Big Data to Knowledge (BD2K) Center of Excellence
mobilize.stanford.edu

Publication is supported by NIH Big Data to Knowledge (BD2K) Research Grant U54EB020405. Information on the BD2K program can be found at http://bd2k.nih.gov/about_bd2k.html. 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, and

Daofen Chen, National Institute of Neurological Disorders and Stroke

GuestEditorial

BY PHILIP E. BOURNE, PhD, ASSOCIATE DIRECTOR FOR DATA SCIENCE,
NATIONAL INSTITUTES OF HEALTH

Building a Biomedical Data Ecosystem

This issue of the *Biomedical Computation Review* features the Centers of Excellence for Big Data Computing. These 12 Centers, funded by the NIH's Big Data to Knowledge Initiative (BD2K), have been established on the principle that we must be united in our efforts to accelerate the translational impact of big data on human health.

The Centers will become hubs in an emerging worldwide biomedical data ecosystem. The foundations for this ecosystem are already being built. Collaborations between international groups, federal agencies and the biomedical science community are forging the way forward with pilot projects and initiatives. These projects are designed to influence advancement through a set of central drivers: to inform policy decisions, to build infrastructure, and to expand the biomedical data science community.

Each of the Centers will investigate a different programmatic theme. However, most will also face similar challenges. Some of these challenges will be dealt with by consortium-wide consensus, while we envision that others will be addressed by each of the Centers in different ways. The individuality of each Center and the collaborations between them will allow us to identify best practices, effective strategies, and program models for solving common biomedical data science problems. These findings will be the foundation for future biomedical data science policy decisions.

The BD2K Centers consortium is also a pivotal element in NIH efforts to develop common infrastructure that supports data and software sharing and cloud computing efforts within the biomedical data science community. This infrastructure, which we call the Commons, will be piloted by the Centers and guided, in-part, by the outcome of another BD2K funded project, the Data Discovery Index Coordination Consortium (DDICC). The DDICC is a community-based effort to establish core principles for finding, accessing, and citing digital research objects (data, software, narrative etc.). The results of the BD2K Center cloud pilots and the DDICC efforts will provide a basis for widespread application of the Commons.

The Centers, the DDICC, and the Commons are part of an emerging ecosystem that fosters collaboration and sharing. This environment will facilitate the expansion of data science beyond the Centers and the DDICC to their collaborators, their colleagues, and their students. A major focus of BD2K efforts is on supporting training to help today's biomedical scientists incorporate data science into their research and to produce the next generation of data-centric biomedical researchers. Each Center has a training plan to support this mission and several of the Centers are planning to collaborate in their training efforts. The network of collaboration, sharing, and training provided by the Centers has great potential to accelerate the growth of a supportive and united biomedical data science community.

The Centers cover a broad swath of experimental data and metadata issues, and their research will provide use cases for essential problems to the biomedical community. Center directors and investigators showed great enthusiasm at a recent kick-off meeting, providing an early indication of the energy and commitment of this consortium. Sustainable growth of biomedical knowledge requires sharing. Through a united ecosystem we hope to improve productivity through faster discovery at reduced cost. Success will not only be measured by the Center's individual projects, but through other laboratories becoming part of the ecosystem and sharing their digital research objects. □

BY JOY P. KU, PhD

Welcome to the New *Biomedical Computation Review*

For nearly ten years, this magazine has been published by Simbios (under principal investigator [PI] Russ Altman) as part of the National Institutes of Health's National Center for Biomedical Computing (NCBC) program. With the end of that program last summer, the magazine faced an uncertain future. But it has gained new life with the support of the Mobilize Center (under PI Scott Delp) as part of BD2K.

Expect to see similar wide-ranging stories that cover the gamut of biomedical computing. But also watch for some new big data–focused stories and columns as well.

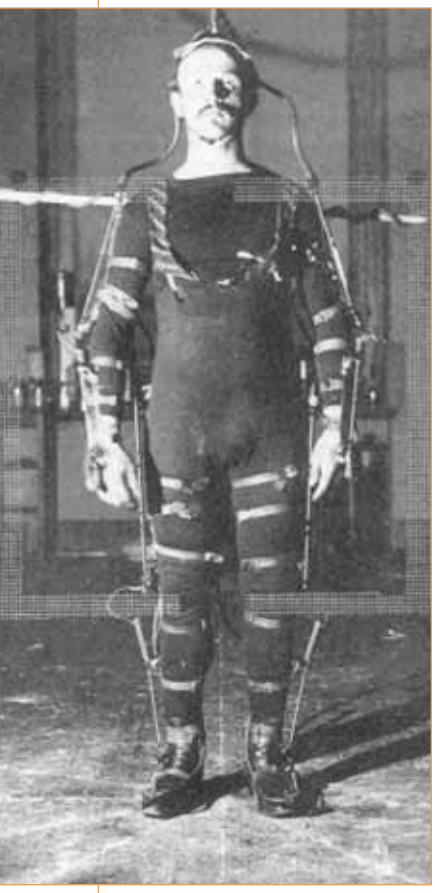
Happy Reading.

BY JOY P. KU, PhD, DIRECTOR OF COMMUNICATIONS & ENGAGEMENT FOR THE MOBILIZE CENTER

Introducing: The Mobilize Center

In 1891, when researchers first analyzed the mechanics of walking in three dimensions, the process was painstaking and cumbersome. It took six to eight hours just to prepare the subject for data collection. Today, such data are routinely collected for various therapeutic and research purposes using digital motion capture technology. At the same time, databases are filling with movement data collected by wearable fitness devices and smartphone apps. The availability of all these data raises the question: How can researchers best make use of it?

Enter the Mobilize Center, a new Big Data to Knowledge (BD2K) Center of Excellence funded by the National Institutes of Health (NIH). "Over the past decades, we've created human models and simulations that embed the mechanics and physiology of movement," says **Jennifer Hicks, PhD**, Director of Data Science for the Center. "Combining these mechanistic models with advanced methods from statistics and machine learning opens up huge opportunities for understanding all these data," she says.



(left) The earliest 3-D motion analysis required Geissler tubes attached to the subject to emit light, a tuning fork to produce bursts of electrical current to the tubes, and thick rubber straps for electrical isolation. (right) Today's gait analysis is much more straightforward to set up, requiring reflective markers to be attached to the subject. Reprinted with permission from Richard Baker, The History of Gait Analysis Before the Advent of Modern Computers,* Gait and Posture *September 2007. Motion capture photo courtesy of: Eric F. Chehab and Matthew R. Titchenal, Stanford BioMotion Lab.

For example, in collaboration with **Michael Schwartz, PhD**, and his colleagues from Gillette Children's Specialty Healthcare, the Center will use statistical approaches to predict the outcome of surgery in children with cerebral palsy (CP)—a neurological disorder that affects motor control—using Gillette's vast database of information about how children with CP move.

A few studies have used these data to understand why an individual walks with a certain gait. However, Hicks says, "We've only scratched the surface in terms of the types of questions we can answer."

To dig deeper, the Center will need to figure out how to make more effective use of time-series data, such as the levels of muscle strength measured every few months for several years. In the past, researchers uncertain of how to make use of such data might simply select a peak or average value from the time period and discard the rest. The data science tools the Center is developing "will help us figure out ways to better condense these time-series signals or include more of them in our models," Hicks says. That additional information, combined with a mechanistic understanding of how humans move, could identify new factors that better predict whether a child will benefit from a given surgery and thereby improve outcomes.

The Mobilize Center is also teaming up with industry partners, such as Azumio, to take advantage of the boom in physical activity monitoring. Azumio, a leader in health and fitness apps on mobile devices, has tons of data from some 70 million users. Indeed, says **Bojan Bernard, PhD**, Azumio's CEO, "The distance walked in one day by users of our Argus app is equal to two trips to the moon and back."

There is a big gap, though, between having all that data and having validated tools that actually encourage people to move more and in ways that reduce injuries. The Mobilize Center, with its experts in data science, biomechanics, and behavioral science, will address that gap. Using anonymized data from Azumio and other mobile applications and sensors, they will shed light on how various factors—environmental, social, biomechanics knowledge—impact behavior, particularly physical activity. Ultimately, that knowledge could be used to create more effective interventions.

"We're very interested in collaborating with researchers to learn what's hidden inside the data," says Bernard. "In the long term, our goal is to use this information to improve human health. Working with the Mobilize Center is an important step towards this goal." □

Big DataHighlight

BY KATHARINE MILLER

Disease Trajectories, Danish Style

In the first large-scale study of its kind, researchers in Denmark computed the disease trajectories for the country's entire population—approximately 6.2 million people—over the course of 15 years, revealing some surprising and interesting patterns.

"The advantage in Denmark is that we can follow people with this lifelong perspective," says **Soren Brunak**, professor of systems biology at the Technical University of Denmark, and lead researcher on the paper published in *Nature Communications* in January 2014.

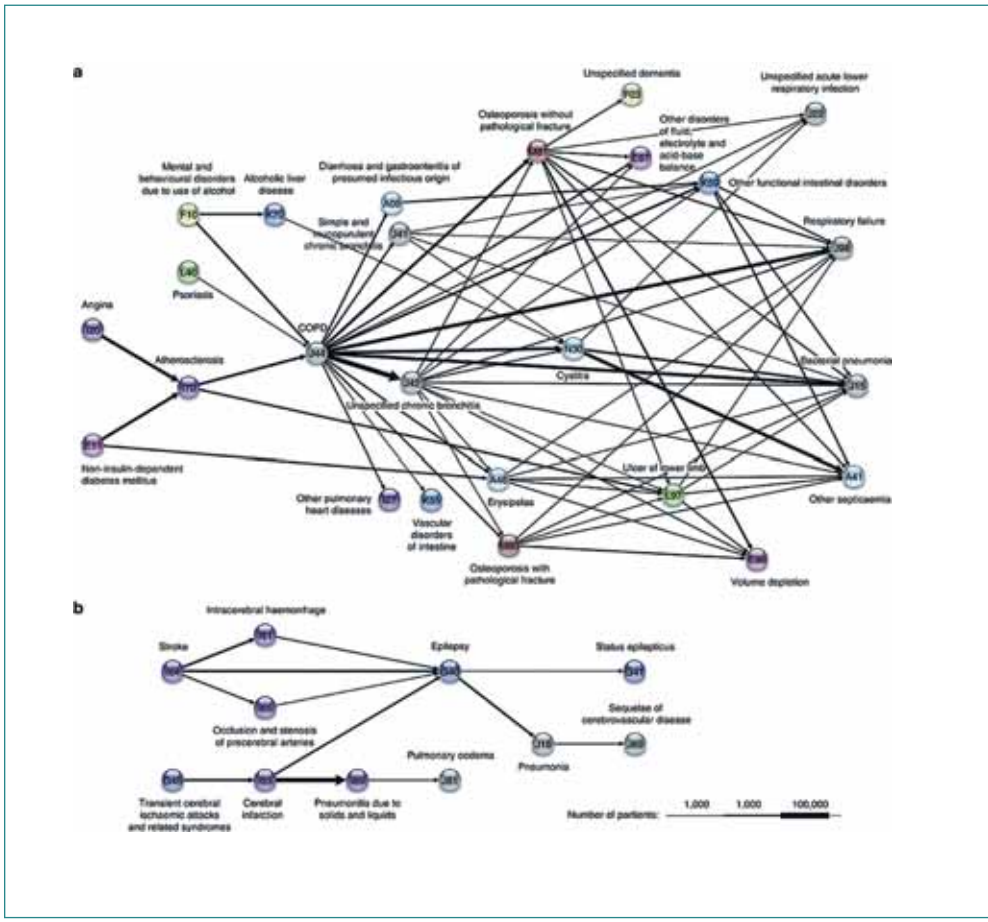
Since 1968, every Danish medical record has been linked to the individual's social security number. Even when people move or visit different physicians or hospitals, the data tracks them. The medical records also document, in chronological order for each patient, every diagnosis. And the dataset covers thousands of diseases.

To determine disease trajectories, Soren Brunak and his colleagues relied on the diagnostic codes from the Danish dataset and looked for pairs of diagnoses with positive relative risk—the probability that one diagnosis would follow another. They then went through two levels of condensation to get at the patterns in the data. First, they condensed the 6.2 million patient trajectories into the 1200 most common pairs of positive relative risk. Next, in a non-hypothesis driven way, the researchers condensed the 1200 trajectories into five trajectory networks. "We took the data, clustered it, and saw what popped up," Brunak says. The five largest networks covered diabetes, cardiovascular disease, cerebrovascular diseases, prostate disease and chronic obstructive pulmonary disease (COPD).

Naturally, many of the strongest trajectories were known by physicians, Brunak says, but there were a few surprises. "Gout, for example, showed up in a very convincing way in the cardiovascular disease landscape," he says, a connection that wasn't previously confirmed.

Brunak is now working on making the networks predictive. By digging into the details of a specific network and layering it with biomarker data (which is also available for many patients) or lifestyle data (such as income or school performance, which are also tied to the social security number), Brunak and his colleagues hope to discover ways to predict individual disease paths. For example, the researchers might look at whether the data predict which diabetic patients will develop renal failure or blindness or some other complication. "What is there in the trajectory or genetic makeup that leads them down a different route?" Brunak says.

The team is also interested in understanding inverse comorbidities—where patients with one disease are less likely to get another disease. Having schizophrenia, for example, is associated with reduced risk of certain types of cancer. Is that the result of medications having a protective effect or do the cellular networks driving the disease provide protection? "It's important to view these trajectories as ways to discover what you don't see as well as what you do," Brunak says. □



Danish disease trajectories clustered into networks such as these for (a) COPD, showing five diseases preceding the COPD diagnosis as well as many diagnoses occurring after that; and (b) cerebrovascular diseases, with epilepsy as a key diagnosis. Reprinted with permission from AB Jensen, et al., Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients,* Nature Communications, *DOI: 10.1038/ncomms5022 (2014).

CANCER'S HETEROGENEITY: Modeling Tumors' Diversity

By Alexander Gelfand

Cancer might spring from a single cell gone awry, but tumors are not monolithic collections of clones. Far from it: They contain many different types of cancer cells, all with their own mutations, proliferation rates, metastatic capacities, and drug responses.

This diversity pushes the limits of current diagnostic and treatment capabilities. A biopsy might miss a crucial subpopulation of tumor cells; and a treatment that works for one set of cells might be ineffective against another set within the same tumor. Moreover, greater heterogeneity is associated with worse outcomes for several types of cancer.

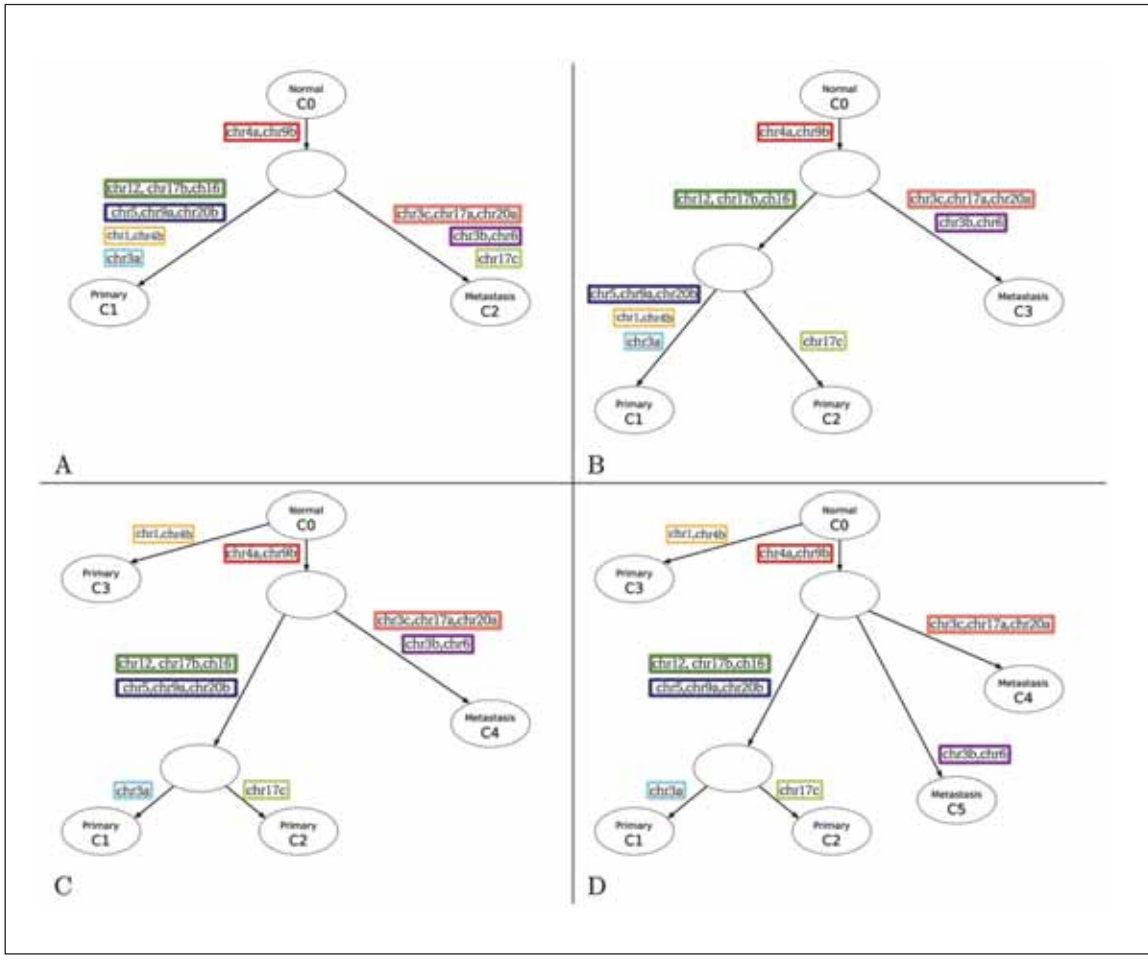
"We want to better understand how to treat tumors more effectively," says **Kornelia Polyak, MD, PhD**, a breast cancer researcher at the Dana-Farber Cancer Institute. And that's going to require learning a lot more about heterogeneity—including how it affects the way a tumor will respond to treatment, and how treatment itself may change the tumor.

To get a handle on cancer's heterogeneity, some researchers are using computational modeling and simulation. They aim to illuminate how the variety of cell types in a tumor influences cancer progression, and to predict the most effective course of treatment for a given tumor. To that end, they are using an assortment of tools almost as diverse as cancer itself, drawing upon fields ranging from machine learning to digital circuit design.

Defining the Clones

Ignoring the presence of even small clonal subpopulations within a tumor can allow them to flourish; so defining the number and nature of clones from limited tissue samples is extremely important, though difficult. Statistician **Daniela Witten, PhD**, and her colleagues at the University of Washington succeeded in doing just that—in work published in *PLoS Computational Biology* in July 2014—by applying some very sophisticated statistical techniques to some very rich next-generation sequencing data.

Witten and her collaborators began with multiple tissue samples from a 44-year-old breast-cancer patient. Some were from the patient's primary and metastatic tumors, others from her healthy breast tissue. Using DNA sequencing, the team identified normal alleles found in both the healthy and cancerous samples, and abnormal ones (mutations) found only in the cancerous samples. But knowing which mutant alleles are present in the cancerous tissue is not enough. Clinicians need to know the genotypes of the cancerous clones. So Witten



After using statistical methods to predict the genotypes of clones in breast cancer samples, Witten and her colleagues honed in on the possible number of clones by reconstructing possible phylogenetic trees for 3, 4, 5 or 6 clones (A-D here). Nodes correspond to inferred clonal populations, with C0 corresponding to the normal clone; edges are annotated with mutations that occur between the parent and child clones. Mutations are grouped into a colored box if they occur on the same branch in all four phylogenies. The team concluded that the three-clone model (A) was too simple to explain the data while four clones (B) explained the data quite well. Adding further clones (C and D) increases the level of detail in the trees but provides little additional predictive value, and can exaggerate the significance of minor fluctuations in the data. Reprinted from H Zare, J Wang, A Hu et al., Inferring Clonal Composition from Multiple Sections of a Breast Cancer,* PLoS Comp Biol, *July 2014, doi:10.1371/journal.pcbi.1003703.g006.

used an approach called statistical machine learning to determine probabilistic estimates of the clone genotypes as well as estimates of the frequencies of those clones in the different tumor subsections. The researchers then modeled different numbers of possible clones (e.g., 3, 4, 5, 6) to see which one would best explain the genetic variety observed in the sequencing data.

In the end, the model worked best with four clones that mapped the clone genotypes and frequencies to the tumor subsections in a way that corresponded both to the physical anatomy of the tumors, and to phylogenetic trees that described how the clones could have evolved from normal tissue, accumulating mutations and branching off from one another over time. That kind of evolutionary insight matters, since knowing how and when one clone gives rise to another could potentially help inform treatment decisions, such as when to introduce a particular anti-cancer drug. "We know that cancer changes fast, and we are really just trying to get a handle on how it's changing, and why," Witten says.

"We know that cancer changes fast, and we are really just trying to get a handle on how it's changing, and why," Witten says.

What Doesn't Kill Them Makes Them Stronger

Getting a handle on how and why cancer changes over time was in fact the primary goal of a series of computer simulations conducted by **Eleftheria Tzamali, PhD**, and her colleagues in the Computational Medicine Laboratory at the Institute of Computer Science, part of the Foundation for Research and Technology–Hellas, in Greece. In particular, Tzamali wanted to understand how different cell types combine with microenvironmental factors to influence the morphology and progression of tumors. For example, invasive behavior in glioma (a type of brain cancer) and breast-cancer cells has been tied to low oxygen levels—though it has also been observed regardless of oxygen level. Many treatments specifically target proliferative tumor cells by modifying their vasculature and starving them of oxygen, and Tzamali wondered how that might affect a tumor that had more than one kind of cell in it.

To answer that question, Tzamali designed her simulations to match the physical structure of gliomas, which are known to contain subpopulations of proliferative cells concentrated toward the center and invasive cells toward the edges. She ran separate simulations using two different kinds of invasive cells: one that is activated by low oxygen levels (i.e., hypoxia), and one that is invasive regardless of how much (or how little) oxygen it gets. And she varied the availability of oxygen to the tumors to mimic different levels of vascularization.

In work published in *PLoS One* in August 2014, Tzamali's simulations recapitulated the physical structure of a real glioma, with proliferative cells clustered in a compact core and invasive cells forming long, finger-like extensions along the rim. They also showed that changing the oxygen levels in a tumor can have unintended, and potentially undesirable, consequences. For example, establishing normal oxygen levels at the outset prevented the hypoxia-driven invasive tumor cells from establishing dominance; but it also accelerated the rise to dominance of the other, more generally aggressive cell type. In general, Tzamali says, the model suggests that drugs that target tumor vasculature might simply favor one phenotype in relation to another, or at best change the rate at which one clone overtakes its competitors.

In general, Tzamali says, her team's model suggests that drugs that target tumor vasculature might simply favor one phenotype in relation to another, or at best change the rate at which one clone overtakes its competitors.

Just as significantly, the model also predicted that invasive cells would appear at the outer edges of the simulated tumors at densities too low for an MRI scan to detect. Since MRIs are commonly used to diagnose gliomas, this argues for performing multiple physical biopsies well beyond a tumor's core in order to find any lurking killers, and against relying too much on drugs that only focus on the proliferative cells that tend to cluster closer in.

In the future, Tzamali and her colleagues plan to ramp up the complexity of their model, integrate more experimental data—and validate their results with lab animals.

Complexity and Chemo

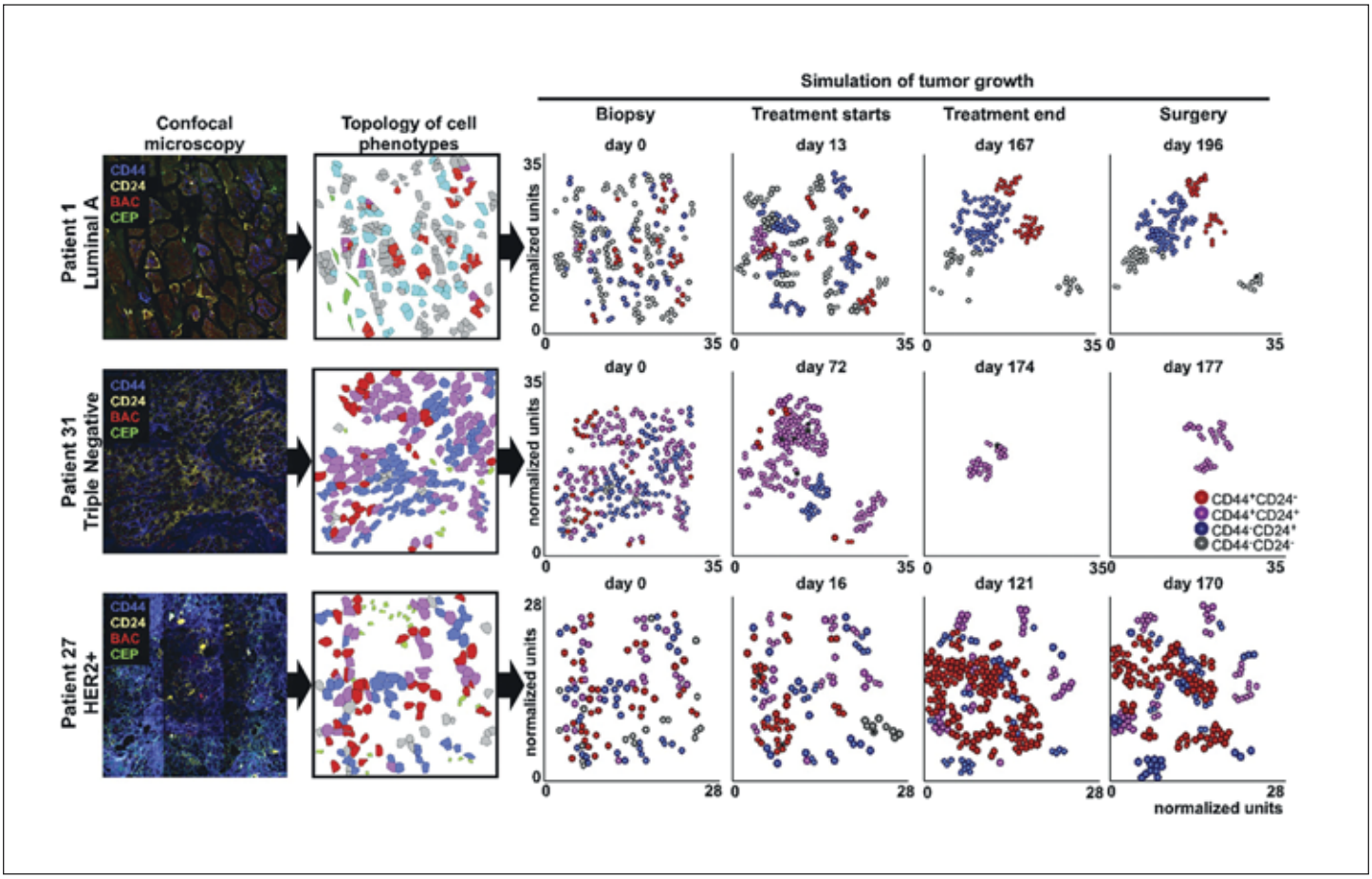
How tumors respond to chemotherapy also reveals the importance of cellular heterogeneity, according to a study by Polyak and an international team of researchers published in *Cell Reports* in February 2014. The team tracked changes in genotype (chromosome copy number), phenotype (four types with different proliferation and migration traits), and spatial coordinates for tens of thousands of tumor cells from 47 different breast-cancer patients who were given chemotherapy to reduce tumor size prior to surgery. They found that the patients with less pre-treatment diversity were more likely to have the tumor completely disappear, leaving nothing behind for the post-chemo analysis. For tumors that showed no or only partial response to chemotherapy, there was very little change in intra-tumor genetic diversity, but the frequencies of the different phenotypes changed, with the more-proliferative types typically declining—something that could happen because chemotherapy tends to target proliferative cells in particular, or because the cells are actually switching between one phenotype and another. Cells of similar phenotype also tended to cluster together after treatment, even when they were genetically different.

To understand what might be driving these growth patterns and evolutionary dynamics, Polyak's colleague, the compu-

tational biologist **Franziska Michor, PhD**, developed an agent-based model to simulate the proliferation and death of each patient's tumor cells before, during, and after treatment. Because the researchers knew the actual phenotypes, proliferation rates, and spatial coordinates of the cells at the beginning and end of treatment, they could vary parameters for each patient-specific simulation to see if the model could account for what had actually occurred.

Intriguingly, they found that proliferation alone could not produce the phenotypic clustering they observed in the patients' tumors. To do that, the model also had to include phenotype switching. Permitting cellular migration, as occurs in metastatic breast cancer, also increased the amount of phenotype switching that was required to explain the patient data. "We knew those were possibilities," Polyak says of their findings, "but it wasn't really expected."

After phenotype switching and migration were added to the model, it proved capable of predicting changes in tumor-cell distributions on a patient-by-patient basis, giving the researchers some idea of what was driving changes at the level of basic biology. In the future, Polyak hopes that such a model could be used to predict the likelihood that a specific individual would respond well to a specific drug, giving doctors the power to run *in silico* clinical trials for their patients and helping them develop better treatment strategies in general.



Michor and her colleagues built these computer simulations of tumor growth using data from actual patients. They show changes in cellular phenotype and topology during chemotherapy. Reprinted from V Almendro et al., Inference of Tumor Evolution during Chemotherapy by Computational Modeling and In Situ Analysis of Genetic and Phenotypic Cellular Diversity,* Cell Reports *6, 514–527, (2014).

Divining the Logic of Cancer

Many researchers aim to tailor treatments to address cancer's troublesome heterogeneity. That's certainly what **David Basanta, PhD**, and **Aniruddha Datta, PhD**, are after, albeit through very different means.

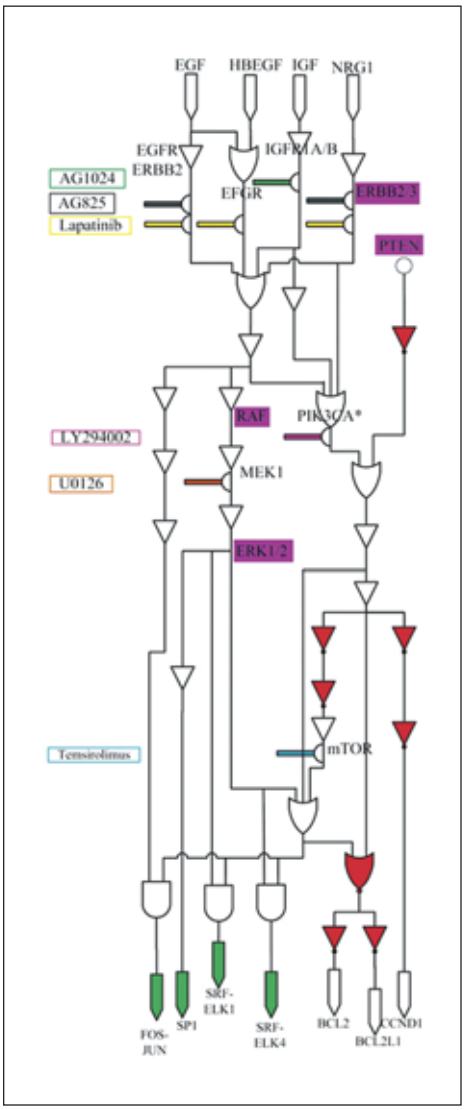
Datta, who directs the Center for Bioinformatics and Genomic Systems Engineering at Texas A&M University, has a background in control-systems engineering, and a fondness for modeling cancer with the kinds of Boolean networks that are used to study digital circuits. For example, with the help of his colleague, the computational biologist **Michael Bittner, PhD**, Datta has represented the mitogen-activated protein kinase (MAPK) signal transduction network, which plays a key role in cell growth, as a series of Boolean logic gates (AND, OR, etc.).

The MAPK network also possesses a number of pathways that, when dysregulated, can lead to cancer. Normally, cell proliferation results when growth factors bind to a cell, triggering a signal transduction cascade that eventually activates the genes involved in proliferation. But mutations in the genes within the cascade can cause those cell-proliferation genes to be switched on permanently, or turn off the genes that would normally inhibit cell proliferation. Both kinds of mutations can be represented in a circuit by something that engineers call a "stuck-at fault," in which a particular logic gate is stuck in the on or off position. The effects of anti-cancer drugs that target these mutations can be superimposed on the same circuit.

In a real patient with a heterogeneous tumor, however, each clonal subpopulation will require its own representative circuit. The trick, according to Datta, is to determine exactly where all of the various faults are occurring; sort out the relative influence that each circuit exerts on the others and on the tumor as a whole; and predict the efficacy of different drug combinations given

all of the above.

Toward that end, Datta and doctoral student **Anwoy Kumar Mohanty** used statistical methods to develop a multilevel, hierarchical model of cancer tissue that can accommodate a number of possible networks as well as multiple drugs. They also developed an algorithm that uses probabilistic techniques (including Bayesian ones) to estimate the frequencies of a tumor's various subpopulations and their relative influence on overall tumor behavior. The work, which is described in a paper that appeared in *IEEE Transactions on Biomedical Engineering* this past March, is still in the early stages: so far, the algorithm has only been validated using synthetic data and experimental data derived from normal, healthy cells. But Datta plans to begin testing the model against three different cancer cell lines within the next year. And he envisions a day when a patient's biopsy could be algorithmically analyzed to produce a drug regimen tailored to his or her particular tumor. "That's my goal," he says. "That's the finish line."



Datta and his colleagues constructed this Boolean network of the MAPK transduction network. Fault locations (i.e., genes stuck in the "on" or "off" position) are shown in purple boxes. Target locations of inhibitory drugs are indicated by solid colored rectangles, with the names of their molecular targets (PTEN, MEK1) printed nearby. The names of the drugs themselves (Lapatinib, U0126) appear in color-coded boxes on the left-hand side of the diagram. Reprinted with permission from A Mohanty and A Datta, A Model for Cancer Heterogeneity,* IEEE Transactions on Biomedical Engineering, *61(3) (2014).

From Genetic Mutations to Genetic Algorithms

It's a finish line that Basanta hopes to cross as well. Basanta and his fellow modelers in the Cancer Evolutionary Dynamics research group at the Moffitt Cancer Center in Tampa, Florida, work with biologists and clinicians to understand the evolutionary dynamics of cancer progression and treatment resistance. Recently, Basanta, **Arturo Araujo, PhD**, **Jill Gallaher, PhD**, and other modelers in the Mathematical Oncology department used differential equations to model a heterogeneous, metastatic prostate cancer tumor located in the bone—a tumor whose cells could possess any permutation of three different mutations, for up to eight possible cancer cell phenotypes with unique proliferation rates and drug responses. The model, which is described in a paper that appeared last August in *Clinical and Experimental Metastasis*, includes two sets of equations: one representing the tumor and its various phenotypes, which grow and respond differently to five distinct therapies (e.g., hormone deprivation therapy, chemotherapy, experimental therapies targeting specific pathways); and one representing the bone microenvironment. The two equations are coupled so that the tumor affects the bone microenvironment, and vice versa.

Basanta and Araujo fed the model with data drawn from the literature, and from lab experiments conducted by a team of Moffitt biologists led by **Conor Lynch, PhD**. And they looked to clinicians at the center for information on how particular treatments were applied to real patients–and how those treatments affected cancer cell growth rates and bone behavior. Once fully parameterized, the model was able to simulate a virtual patient; and each simulation could include a different tumor with its own particular mix of mutations and cell types.

The outputs of each patient-specific simulation were then fed into a genetic algorithm that produced 1,000 successive generations of treatment options using the five drugs in the model. With each generation, the algorithm dropped the worst performing treatments and kept the best, until it was left with only those that kept the cancer at bay the longest. For each patient, the algorithm came up with an optimal single-drug regimen, and another that used more than one therapy in a particular sequence. And it did it all pretty quickly: for a virtual patient with the most heterogeneous tumor possible, the algorithm arrived at the current real-world standard of care–continuous hormone deprivation therapy—in less than 15 minutes. It also yielded some surprising results. In a couple of cases, for example, the algorithm recommended sequential courses of only two of the five therapies. And as Basanta notes, the first therapy didn't even have to kill as many cancer cells as possible; it just had to prime the tumor for the second one.

"It really defies logic," Araujo says. "You'd think that if you threw everything you had against the tumor, it would work."

"It really defies logic," Araujo says. "You'd think that if you threw everything you had against the tumor, it would work."

But, he explains, the tumor is evolving and developing resistance according to a genetic algorithm of its own. "How can you tackle that? How can you keep one step ahead of what the tumor's going to do? The only way is to simulate how the tumor might evolve and anticipate all of its moves using the same genetic algorithm."

Like Datta, Basanta and Araujo hope for a day when a patient can walk into the clinic, have his tumor sampled and analyzed, and receive a personalized treatment regimen based on its specific composition. "A lot of these decisions are being made in the dark," Araujo says of the current approach to treating tumors that may contain multitudes. "But mathematics says there is a better way." □

STEM CELL (RE)PROGRAMMING: Computing New Recipes

By Sarah C.P. Williams

Most scientists seeking to turn back adult cells' developmental clocks rely on go-to recipes that—when followed just right—will yield stem cells. A dash of one reprogramming factor, a sprinkle of another, and let the mixture stew. Likewise, when researchers want stem cells to remain stem cells or, alternatively, when they want them coaxed down a particular developmental pathway, they have cocktails they turn to. Most of these recipes were concocted using trial and error over the past few years, and then they've been passed between labs. Whether they're the best ways to derive or control stem cells, or the most efficient, is unclear.

Now, by harnessing the power of big data, modeling, and computational biology, scientists are starting to write new—and potentially better—protocols for creating and maintaining stem cells, based on a better understanding of how large networks of genes and proteins interact to influence cellular development and differentiation.

"It takes time for stem cell researchers to embrace these kinds of systems level methods," says **Avi Ma'ayan, PhD**, of the Icahn School of Medicine at Mount Sinai. But as these approaches have started yielding results, he says, there's much more interest in giving them a shot.

Reasoning a Better Reprogramming Recipe

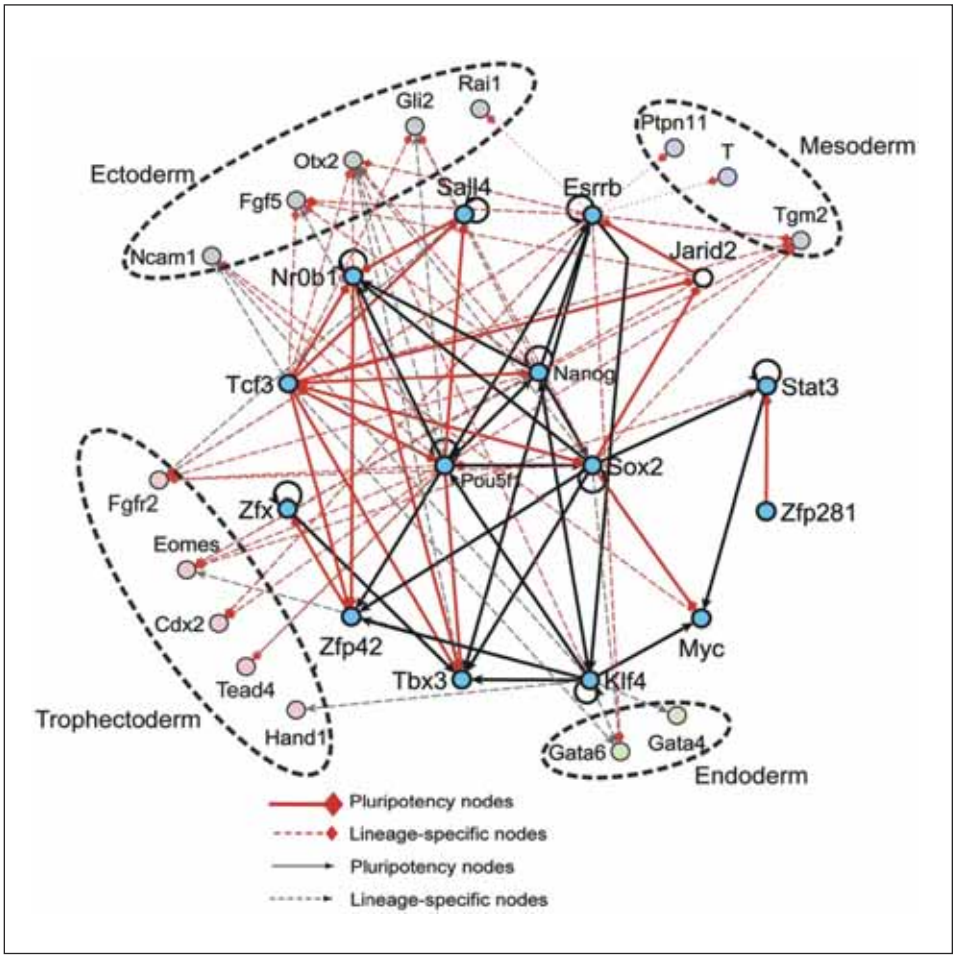
At the Hebrew University of Jerusalem, one team of researchers was getting frustrated by the low yield of the stem cell reprogramming methods they were using to coax adult cells into a pluripotent state—able to become any cell in the body.

"The problem was that this program was very inefficient; only a small number of cells became stem cells, and then you'd have to use single cell technologies to capture these pluripotent cells," says **Yosef Buganim, PhD**, of the Hebrew University of Jerusalem. Moreover, once those induced pluripotent stem cells (iPSCs) were isolated, the quality of them varied. Only about 20 percent of the mouse iPSCs, Buganim says, had the capability to develop into a whole mouse—the true test for a stem cell.

Buganim's team was using a well-known mixture of transcription factors, dubbed OSKM for its four main ingredients: Oct4, Sox2, Klf4, and Myc. The scientists began to wonder whether the OSKM factors were turning on not only the genetic programs that led to pluripotency, but other programs that were contrary to this goal. So, using a combination of lab work and bioinformatics, they started figuring out how OSKM influence 48 other transcription factors that were turned on during the reprogramming process.

The final network of genes they uncovered revealed just what they wanted: four transcription factors turned on by OSKM which could, themselves, induce pluripotency without turning on other, unwanted genetic programs. "It would take you a hundred years if you just tried culturing cells with all these different combinations of factors," Buganim says. But with bioinformatics, they could analyze the gene expression patterns much more quickly.

When Buganim's group used the new mixture—SNEL for Sall4, Nanog, Esrrb, and Lin28—on adult mouse cells, they were able to generate higher quality iPSCs than ever before. Eighty percent could generate a whole mouse that could live for more than year, Buganim says. The results were



Ma'ayan and his collaborators connected 15 known pluripotency regulators to 15 lineage markers in this network, which shows how various combinations push the cell toward four different fates (circled in dotted lines). Reprinted from Xu H, Ang Y-S, Sevilla A, Lemischka IR, Ma'ayan A (2014) Construction and Validation of a Regulatory Network for Pluripotency and Self-Renewal of Mouse Embryonic Stem Cells.* PLoS Comput Biol *10(8): e1003777. doi:10.1371/journal.pcbi.1003777.

published in *Cell Stem Cell* in September.

Now, his team is focused on uncovering how even more genes interact with the pluripotency program. Rather than analyzing just 48 genes already suspected to play a role, they're using big data to look at all the genes in the cells as they reprogram from adult to iPSC.

"We have the technology to probe the transcriptome of the entire cell, and that makes the bioinformatics analysis that much more important," says Buganim. "When you're talking about 20,000 genes instead of 48, it would be a nightmare to analyze by hand."

Knocking Down Barriers to Reprogramming

With the growing use of bioinformatics in biology labs, Buganim's group isn't the only one using computational modeling approaches to work out better reprogramming recipes. **Aaron Diaz, PhD**, an applied mathematician at the University of California, San Francisco, recently used a massive library of short hairpin RNA (shRNA) to selectively block genes in cells as they were being reprogrammed toward pluripotency. Diaz and colleagues then analyzed the results of this genome-wide screen, using systems biology and bioinformatics approaches, and discovered key pathways regulating the transition to pluripotency.

Each shRNA—from a library of shRNAs targeting more than 19,000 genes in human cells—was packaged inside a viral particle that had a unique barcode added. Then, each of the hundreds of thousands of unique viruses were added to human fibroblasts. Using the classic OSKM technique, Diaz and colleagues then coaxed the cells to become iPSCs. If a cell contained a shRNA that blocked a gene necessary for reprogramming, it would fail to turn into an iPSC. Using high-throughput next-generation sequencing, the researchers could then determine which shRNAs were present in cells that became iPSCs, and which shRNAs were enriched in cells that failed to reprogram.

Next, they turned to bioinformatics to analyze these results and filter out off-target effects—false positives, essentially. More than a thousand genes originally appeared in the screen as influencing reprogramming, but the analysis honed the list down to about 20 that, if blocked, enhanced reprogramming. They then clustered genes into functional modules and studied their interactions, using a combination of tandem knockdown experiments and a novel software tool they developed: HiTSelect. The proteins they ended up validating were involved in a range of different cellular processes including transcription, chromatin regulation, vesicle-mediated transport and cell adhesion. The work was reported in a July 2014 *Cell* paper. Now, Diaz says, they are able to add shRNAs—or other molecules that selectively block these barrier genes—to the OSKM cocktail to help lift these blocks on reprogramming.

"I think it's going to be more and more mandatory for biologists to have some background in computation," says Diaz. "With advances in single-cell sequencing, for example, it is becoming routine for us to generate hundreds of genome-wide profiles per experiment. There's just no way you can analyze that many datasets without modern data science approaches."

Computationally Informed Differentiation

Once iPSCs are generated, researchers want to know how to engineer the fate of these cells—either sending them down a pathway to become brain, blood, bone, or any other type of somatic cell, or keeping them dividing as stem cells. Looking deeply at broader cell networks by analyzing the expression levels of many genes can definitely help, says Ma'ayan. "There's a lot of excitement around using bioinformatics to improve differentiation protocols," he says.

In an August *PLoS Computational Biology* paper, Ma'ayan and collaborators described a new model of how 15 different transcription factors and 15 lineage markers interact with each other to influence the differentiation of stem cells. Other researchers, Ma'ayan says, have generated a plethora of data on these transcription factors—using techniques ranging from cDNA microarrays, RNA-seq, chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq), mass spectrometry proteomics and phospho-proteomics, and RNAi screens. But the data from all these approaches has never been integrated before.

"Finding the data is not very hard," Ma'ayan says, "but putting it together is a real challenge."

So his group, using a database program they developed called ESCAPE—for Embryonic Stem Cell Atlas of Pluripotency Evidence—took on this challenge. They manually collected and organized each piece of evidence to fit it into ESCAPE, then added it to their pile of evidence. The new network, a dense spider web of arrows between the 15 transcription factors and 15 lineage markers, shows how the increased expression of one factor can push a cell toward one of four different fates: ectoderm, mesoderm, trophoectoderm, and endoderm; this network was then validated experimentally in living cells by knocking down individual or combinations of factors and then measuring the changes in expression of the rest of the nodes in the network model.

"We didn't find anything earth shattering, but now we have a global framework to work from," Ma'ayan says. "In principle, if the model works and it becomes predictive and large enough, we can use it to improve differentiation protocols and reprogramming strategies."

Researchers agree that the future of stem cell research—and therapeutics based on stem cells—requires the ability to quickly and efficiently create personalized stem cells from a patient's own adult cells, and then coax these iPSCs into whatever healthy cell is needed by that patient. To meet that end, though, the field needs more predictable methods to direct stem cell fates. Computational models are helping to achieve this goal. □

"I think it's going to be more and more mandatory for biologists to have some background in computation," says Diaz.

"There's a lot of excitement around using bioinformatics to improve differentiation protocols," Ma'ayan says.

ASSEMBLING THE 3-D GENOME:
A Puzzle with Many Solutions

By Katharine Miller

As a result of experimental techniques developed about a decade ago, researchers now have data that can be used to reconstruct how the genome is arranged inside the nucleus. This 3-D structure likely plays a role in determining cellular function by affecting cells' ability to access, read and interpret genetic information.

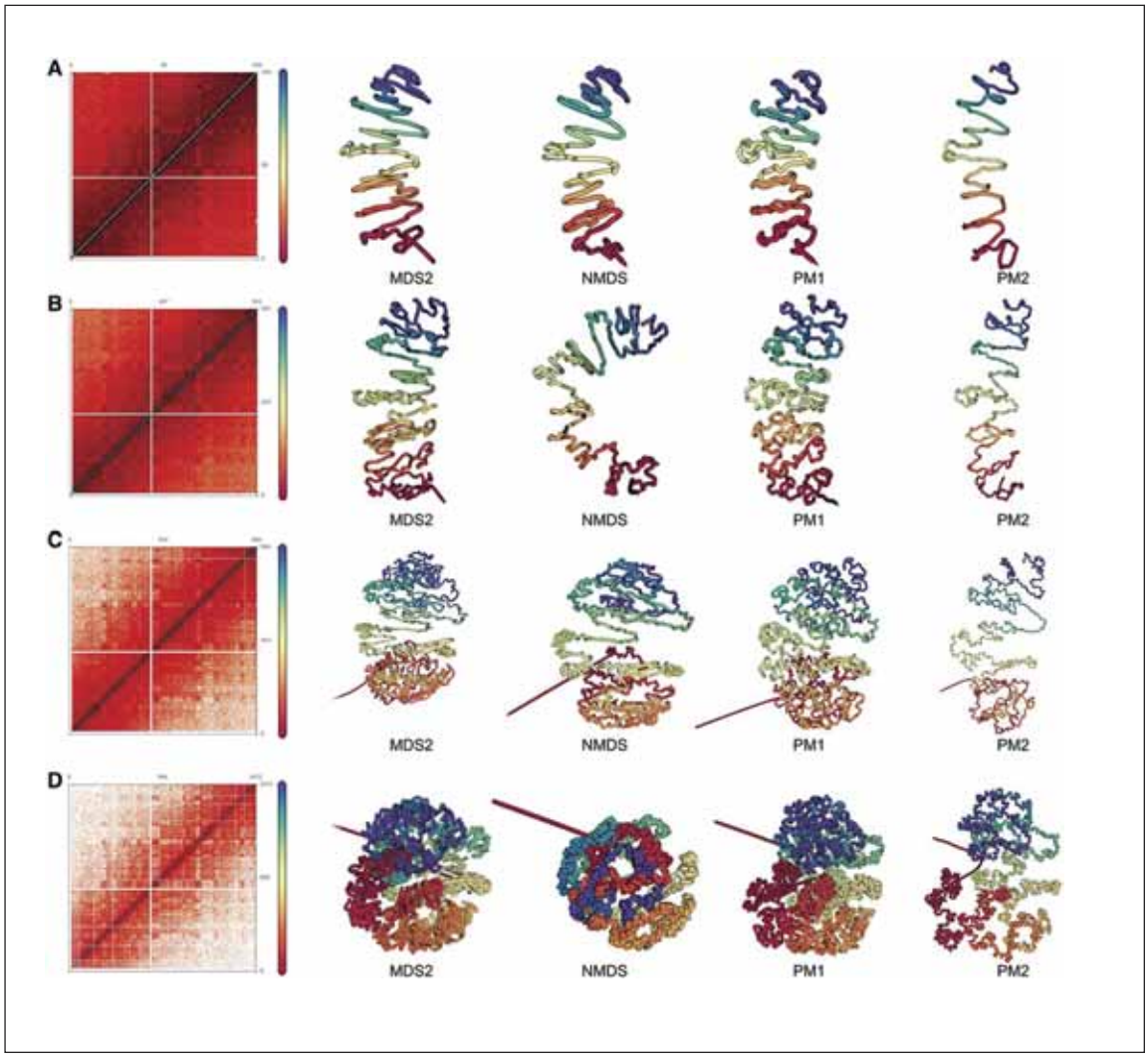
"We want to use 3-D genome reconstruction to understand the guiding principles of genome organization," says **Frank Alber, PhD**, associate professor of molecular and computational biology at the University of Southern California. "There is a lot to be learned. We are just at the beginning."

Experiments called chromosome conformation capture—of which there are now multiple types, including 3C, 4C, 5C and Hi-C—allow scientists to determine the frequency with which loci on the genome are in contact with one another—considering all possible interactions. These contact frequencies are derived from experiments that are done on 10 to 20 million cells at a time, and therefore do not represent the 3-D structure of any one cell. Using computational approaches, however, researchers have developed ways to assemble plausible 3-D structures.

One approach is to convert contact frequencies to Euclidean distances—using one of several mathematical or probabilistic options—and then optimize the distances to generate a consensus structure. Other researchers try to infer which contacts co-occur and then generate an ensemble of possible structures. Both methods—consensus and ensemble—generate structures that are essentially fictional: There is currently no way to know whether a structure generated by these computational methods actually occurs in nature. Yet insights can be gained from these methods nonetheless.

Converting Frequencies to Distance

Researchers have tried a variety of mathematical approaches to convert contact frequencies into distance. Initial efforts assumed that the frequency of intrachromosomal contacts could be directly mapped to Euclidean distance in 3-D, says **William Noble, PhD**, professor of genome sciences at the University of Washington. They plotted genomic distance as a function of contact count and then swapped genomic distance for a distance in 3-D (Euclidean distance) that was calibrated using imaging.



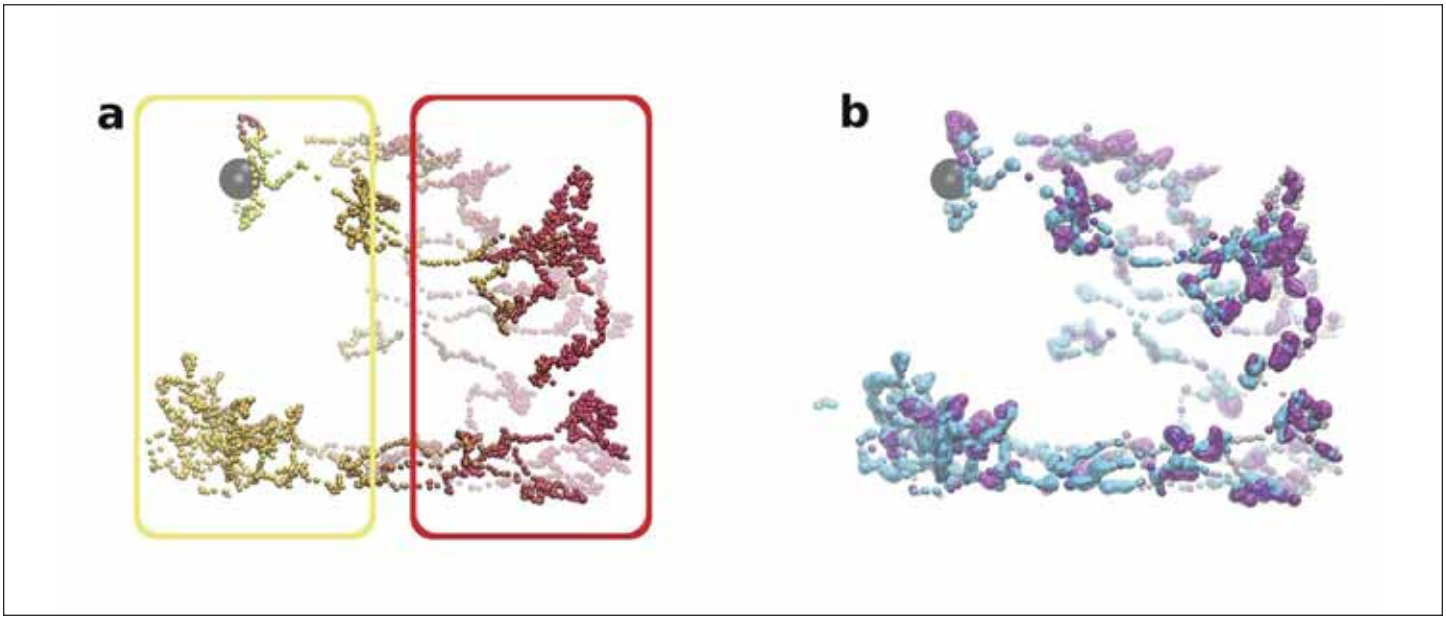
Noble and his colleagues compared two different variants of his Poisson method of predicting the 3-D structure of chromosome 1 with several different multidimensional scaling algorithms (MDS) at different resolutions: 1 Mb (A), 500 kb (B), 200 kb (C) and 100 kb (D). The second Poisson method was more stable in response to resolution changes than were the other methods. Reprinted from N Varoquaux, F Ay, WS Noble, and JP Vert, A statistical approach for inferring the 3D structure of the genome,* Bioinformatics *(2014) 30 (12): i26-i33.

"It's basically a ruler," Noble says. Other methods have determined the ruler differently. For example, a tool called ChromSDE fits a parametric curve to the data. And recently Noble used a model that assumes the contacts are generated according to a Poisson process where events occur randomly over a given time interval with a particular (Poisson) form of distribution.

In each case, researchers optimize the distances to converge on a single, consensus structure that is essentially an average of all possible structures that could exist in the millions of cells sampled. Noble says the consensus structure is useful for visualization and hypothesis generation, but he cautions: "It's risky to conclude anything from these models alone, and validations [using fluorescence *in situ* hybridization or FISH] are expensive and therefore sparse."

A recent paper published in *Nature Methods* by **Julien Mozziconacci, PhD**, Lecturer in Physics and Biology at Pierre & Marie Curie University, Paris, France, and his colleagues took a graph theoretic approach to the same problem. They graphed the contact frequencies from a set of Hi-C data as a single structure where the weights on the graph are the inverse of the contact frequency. "The higher the frequency, the closer the distance," Mozziconnaci says. For unconnected nodes, the graph assigned the shortest possible distance in order to fulfill the triangular inequality, i.e., given three points in space—A, B, and C—the sum of the distance between A and B and the distance between B and C is always greater than or equal to the distance between A and C. "If you don't have this property then you are not talking about distances," Mozziconnaci says.

Genome sequence data and related information, such as function and epigenomic data, are essentially one-dimensional. Annotating the 3-D structure with this information and viewing it in a 3-D genome browser allows novel observations, Mozziconacci says.



Unlike the other approaches described above, Mozziconacci's approach, called ShRec3D, does not include an iterative optimization step. "The matrix analysis directly gives the structure," he says. The triangular inequality is satisfied on the graph, but not necessarily in Euclidean 3-D space. "In the end, the structure is a view of the mind. There is no such structure in 3-D space."

Despite their fictional nature, one advantage of the various consensus approaches, Mozziconacci says, is that they can be integrated with a 3-D genome browser. Genome sequence data and related information, such as function and epigenomic data, are essentially one-dimensional. Annotating the 3-D structure with this information and viewing

ShRec3D allows researchers to superimpose existing information onto reconstructed 3-D chromosome structures. For example, chromatin might be partitioned into compartments as shown in (a), where yellow indicates gene-rich, GC-rich regions, on the left and red indicates gene-poor, AT-rich regions, on the right. Or researchers might display, on the 3-D structure, linear information such as shown in (b), where cyan regions harbor a high level of acetylation; pink regions harbor a high level of tri-methylation; and purple regions harbor both modifications. Reprinted with permission from A Lesne, J Riposo, P Roger, A Cournac, J Mozziconacci, 3-D genome reconstruction from chromosomal contacts,* Nature Methods *(2014) doi:10.1038/nmeth.3104.

it in a 3-D genome browser allows novel observations, Mozziconacci says. For example, a researcher might display where a particular transcription factor lies in 3-D space relative to other loci with which it is known to interact. Mozziconacci looks forward to a time when different techniques, such as sequencing and microscopy, are brought together in a unified model. "People get very excited about getting the crystal-like structure of the genome, but we need to assess the structure-function relationship," Mozziconacci says. "I don't think we've seen many insights on the function side yet. That's still to be discovered."

Ensemble Methods

If structural heterogeneity in the genome reflects functional variations among cells, consensus approaches might not provide the full picture, Alber says. "It's unlikely that the genome falls into a single optimum structure."

So he and his colleagues use large Hi-C datasets to generate a range of possible 3-D genome structures. "We deconvolute the Hi-C data into a population of individual structures that, as a whole, are statistically consistent with the data." The aim is to figure out which contacts are most likely to co-occur. Simply embedding the data in 3-D limits the interaction among two regions. Alber also considers the cooperativity principle—if two regions are interacting then perhaps neighboring interactions are more likely.

The frequency of each contact is then accurately reproduced in an ensemble. So if we infer from Hi-C experiments that A contacts B in 15 percent of cells, A will contact B in 15 percent of the ensemble. "This is an approximation of the true population," Alber says. "We don't know what the true population is, and the data are incomplete, but we integrate additional information to get a better approximation."

Once there's an ensemble of tens of thousands of structures, there remains the question of what biology you learn from it. There's a need for new structural biology tools that can mine the structures in the population to find patterns of co-occurrence (when A contacts B does it also tend to contact F?) and relate them to function, Alber says.

From Fiction to Reality: Single-cell Hi-C

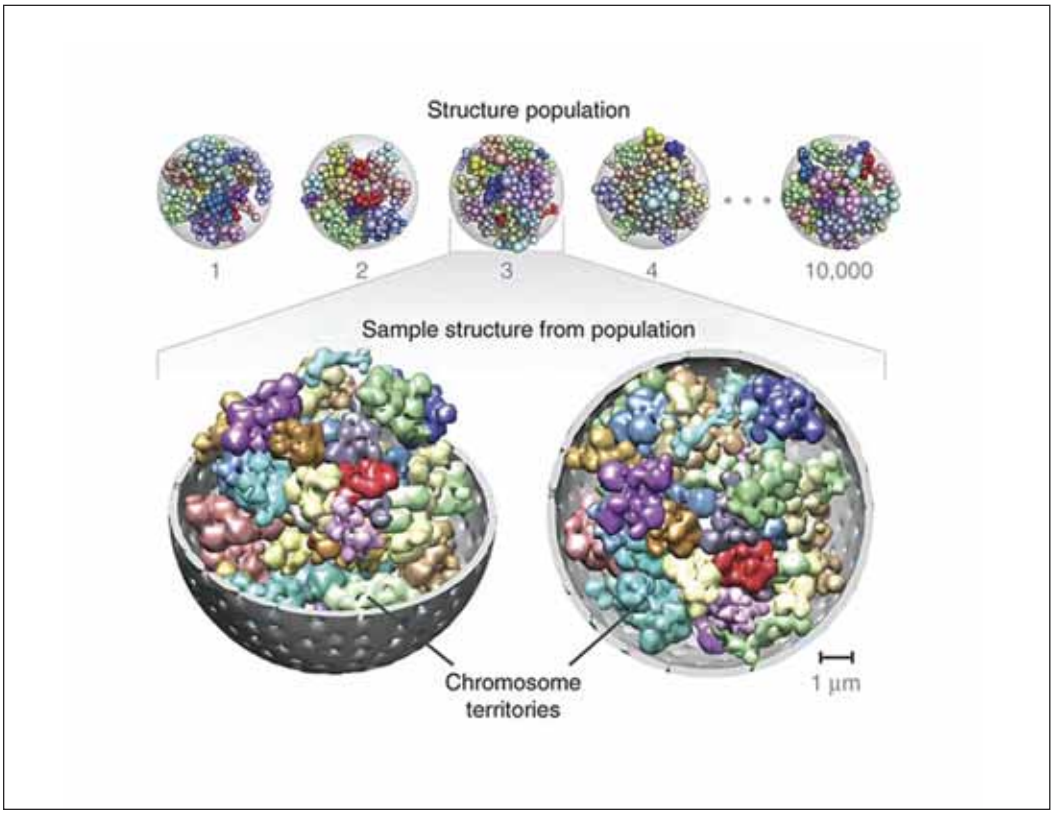
The consensus approach tends to average out the real differences among heterogeneous structures, Alber says. Take the situation where A contacts B half of the time and A contacts C half of the time. In fact, in half the structures these pairs are actually interacting, but the consensus will put these both at a specific distance from each other based on frequency.

But the ensemble approach may suffer from a different kind of unreliability, says **Mark Segal, PhD**, professor of biostatistics at the University of California, San Francisco. The ensemble may not correspond to actual variation in a cell population, he says. "It might, but it's unfounded. It could all be algorithmic as opposed to correlating with anything biological."

One thing all agree on: Single-cell assays have the potential to be more informative. "That's what's coming next," Segal says.

One thing all agree on: Single-cell assays have the potential to be more informative. "That's what's coming next," Segal says.

Single-cell Hi-C is done, as the name suggests, on a single cell. It still suffers from the same low efficiency that troubles Hi-C generally—it may only identify 1000 loci in any one cell. But if done on hundreds of thousands of cells, it could produce an ensemble that could then be linked computationally to current ensembles—lending them a basis in reality. □



Alber's ensemble approach produces a population of more than 10,000 genome structures. A schematic view of the calculated structure population is shown on top. A randomly selected sample from the population is magnified at the bottom. All 46 chromosome territories are shown. Homologous pairs share the same color. The nuclear envelope is displayed in gray. For visualization purposes, the spheres are blurred in the magnified structure because the use of 2 × 428 spheres to represent the genome makes the territories appear more discrete than they actually are. Reprinted with permission from R Kalhor, H Tjong, N Jayathilaka, F Alber, and L Chen, Genome architectures revealed by tethered chromosome conformation capture and population-based modeling,* Nature Biotechnology *30(90–98) (2012).

NIH Launches A United Ecosystem FOR BIG DATA

By Katherine Miller

Francis Collins, MD, PhD, Director of the National Institutes of Health (NIH), says he used to feel "data envy" toward the field of physics. In those days, "no one would have predicted that biology would emerge as the biggest challenge in terms of data. But that is now the case."

Last year, under Collins' leadership, the NIH stepped up to take on that challenge by announcing the Big Data to Knowledge (BD2K) program, a wide-ranging plan to enhance biomedical researchers' ability to make effective use of big data.

"The goal is to begin establishing an ecosystem that supports tools, data and best practices for this new expanded way of doing biomedical research," says **Philip Bourne, PhD**, NIH Associate Director for Data Science.

The first step toward achieving that goal was the announcement last September that the NIH is establishing 12 BD2K Centers of Excellence, granting each center approximately $2 million a year for four years ($24 million/year total).[1]

At the time of this writing, the Centers are just getting started. But interviews with the principal investigators (PIs) about their data science goals reveal the Centers' potential to alter the landscape of big data science in biomedicine.

Imagine: All health information across an individual's lifetime accessed through a single system and layered with data about lifestyle and environmental exposures; wearable sensors continuously tracking patients' health status and allowing remote interventions that are both effective and reliable; physicians predicting, with a few tests and a few clicks of a mouse, what treatments are appropriate for a specific patient based on his or her unique genetic make-up; and the cooperative analysis of neuroimages worldwide to generate an exponential increase in our understanding of the brain. These imaginings are all part of the future envisioned by the BD2K Centers and supported by the NIH.

But before these visions can become a reality, the Centers have their work cut out for them. They need to find ways to make effective use of the big data that continues to flow from biomedical labs and high-throughput experiments, including patient records, genomics and other –omics data, imaging data, and data from mobile devices and wearable sensors. The Centers are charged with integrating vast amounts of data, connecting data to knowledge, and developing new statistical and analytical approaches that work well with big data. And let's not forget the nuts and bolts of standardizing data, collecting better metadata and building pipelines to bring all of this to bench biologists.

Wisely, NIH has directed that the data science goals be achieved and validated in a biomedical research milieu. "For methods to be broadly applicable, they need to be developed in the context of a particular question," says **Scott Delp, PhD**, professor of biomedical engineering and PI of the BD2K-funded **Mobilize Center**. For example, several centers will determine whether wearable sensor data, collected for 24 hours seven days a week, can be used to improve patient health by motivating exercise, detecting when former smokers relapse, or reducing hospital admissions for congestive heart failure.

By taming big data, the BD2K ecosystem will enable a deeper understanding of the human organism while at the same time motivating major improvements in the practice of medicine.

"My view," says **Santosh Kumar, PhD**, associate professor of computer science at the University of Memphis and PI for the **Center of Excellence for Mobile Sensor Data-to-Knowledge** (MD2K), "is that after four years, it becomes possible for any researcher to use all of these tools collectively to get a holistic view of the person they are studying." Biomedical discovery will no longer be siloed in any particular data source but instead available to anyone with a computer. "With that, the true power of BD2K will be realized," Kumar says.

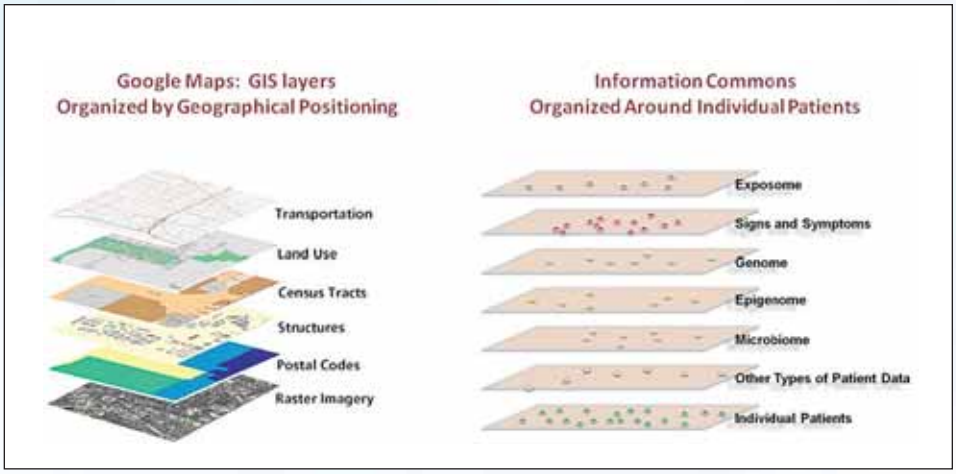
By taming big data, the BD2K ecosystem will enable a deeper understanding of the human organism while at the same time motivating major improvements in the practice of medicine.

PATIENTS AND BIG DATA: *Going Holistic*

Building a Patient-Centered Coordinate System

To gain a holistic view of patients, physicians and researchers need health data to be better organized and more readily accessed. The new BD2K-funded Patient-Centered Information Commons (PIC) envisions a Google maps-like layering of data, with patients as the essential coordinates.

The idea sprang from a National Acad-



PIC proposes developing a Patient-Centered Information Commons (right panel) that is analogous to a layered geographic information system (left panel). This integrated system would include clinical information from electronic medical records as well as genomics, proteomics, and environmental context, thus enabling a much more comprehensive characterization of disease states and health states. Reprinted with permission from **Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease.** ***Washington, DC: The National Academies Press, 2011.***

[1] Additional BD2K initiatives announced at the same time bring the total to $32 million per year for four years. Watch this magazine for future articles about these projects.

emies of Science report "Toward Precision Medicine," says PIC PI **Isaac Kohane, MD, PhD**, professor of pediatrics at Harvard Medical School. The system would include clinical information from electronic medical records as well as genomics, proteomics, and environmental context—the so-called exposome.

"If we could bring all these disparate data together across a lifetime, we'd have a much better sense of what informs disease state and health state," Kohane says. "That's the information commons. And our goal is to make that a pragmatic reality so others can explore it."

The challenges are numerous. Lacking a national health identifier, it's difficult to determine what medical data belong with a single patient, let alone how to layer that with non-health sources of data, such as shopping behavior, local pollen counts, pollution indices, or social media information, Kohane says. Medical record numbers uniquely identify patients within particular healthcare environments, but distinguishing one John Smith from another across a claims database (without medical record numbers) is another matter. A few identifying characteristics may allow probabilistic assignments of data to specific people, but that becomes non-trivial as the data get farther and farther apart, Kohane says. "All these data are heterogeneous, sparse, biased and noisy," he says. "So how you actually clean them up in a way that can use our standard tool kits against them is an open question."

PIC plans to start by building a virtual sandbox with data uploaded to a secure cloud service. "That will allow us to start playing the games of finding common coordinate systems across the data types," he says.

PIC's first focus will be neurodevelopment, for which several large hospitals have committed genomic and clinical data. The center will layer the data together as a way to not only achieve the center's goals but also answer questions for the neurodevelopment community. The success of a center like this, Kohane says, "lies not just in developing a widely adopted architecture, but in using the architecture to do interesting things."

Exploiting Mobile and Wearable Sensor Data

A variety of sensors on wristbands and inside smart phones allow the collection of health data around the clock, potentially enabling a holistic view of patient health in ways never before imagined. Several BD2K centers are exploring this potential. At the Mobilize Center, for example, researchers envision using wearable sensor data to encourage healthy physical activity in people at risk for obesity or to warn runners of impending injury.

At the same time, the team at the MD2K Center of Excellence will investigate whether data from multiple types of mobile sensors can help clinicians monitor—and eventually treat—people with various kinds of chronic illness. The MD2K team will gather immense amounts of data from a population of smokers for two weeks using wristbands and chest bands to track activity levels; mobile phones to track not only movement but also location; and Google glass to track what is in an individual's field of vision. In addition, they will strap these devices as well as radiofrequency sensors to a separate population of heart disease patients. It's a huge amount of data that MD2K will be collecting 24 hours a day at a rate of tens of kilobytes per second, Kumar says.

Several BD2K centers will research the efficacy of using mobile sensors such as mobile phones, wrist sensors and Google glass to monitor health status on a continuous basis and to create more effective interventions. The 24/7 nature of these devices have the potential to radically change the way medicine is done. RisQ mobile smoking detection app and wristband image reprinted from Abhinav Parate, Meng-Chieh Chiu, Chaniel Chadowitz, Deepak Ganesan, Evangelos Kalogerakis, RisQ: Recognizing Smoking Gestures with Inertial Sensors on a Wristband,* Proceedings of the 12th International Conference on Mobile Systems, Applications and Services *(MobiSys 2014). Google glass image by Mikepanhu, creative commons license.

At first, the researchers will just be trying to convert sensor data into markers of health state, behavior, or environmental exposure. For example, can wrist sensors reveal arm movements indicative of smoking as compared to eating? Can chest sensors signal stress levels in ways that relate to an urge to begin smoking again? Can GPS or Google glass data signal proximity to social cues (such as being in a bar, or in close proximity to cigarettes in a store) that might prompt smoking? And for heart patients, can radiofrequency sensors yield valuable information about fluid accumulation in the lungs that might suggest an adverse health event and be used to reduce hospital readmissions?

Once MD2K finds markers of health state, the team will apply machine-learning approaches to discover associations among the various markers. "The challenge is that any one sensor by itself has some information but not necessarily enough," Kumar says. And data quality is-

sues abound. "We need to be able to figure out when changes in the data are due to something we want to infer or something else entirely." MD2K's goal is to make extracting health markers from sensor data feasible as well as reliable enough to trigger an appropriate intervention.

DISEASES, DRUGS, AND THERAPIES: *Integrating Data and Knowledge to Improve Patient Care*

Large public datasets such as GenBank, the Protein Data Bank (PDB) and many others are already widely used by biomedical researchers. But numerous valuable data resources remain dispersed and isolated at institutions around the world. The BD2K Centers will develop various approaches for connecting multiple data types and knowledge resources with one another. Thus, just as PIC is planning to bring together disparate data to gain a much more holistic understanding of patient health, many of the other BD2K Centers plan to integrate multiple types of data and knowledge to achieve a more comprehensive understanding of the human organism. Though the approaches the Centers take to this task may differ, they all have the potential to generate insights that could lead to new drugs, targeted drug regimens, or personalized therapies or surgical interventions.

Data Integration and Cellular Signaling

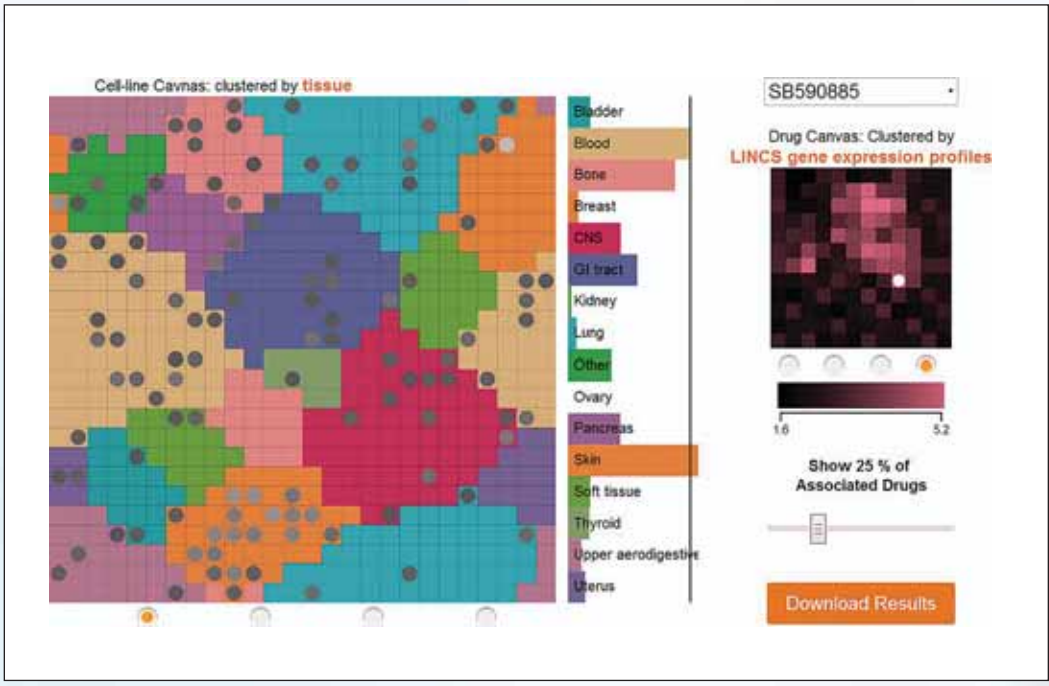
For bureaucratic and scientific reasons, the pharmaceutical development process is notoriously slow. Many researchers believe drug discovery would be more efficient if we had a better understanding of the relationships between diseases, the drugs that treat them, and the pathways the drugs target in different cells and tissues. Gaining that understanding requires gathering and integrating lots of different kinds of data and knowledge.

That thinking lies, at least in part, behind BD2K support for the **Data Coordination and Integration Center for LINCS** (BD2K-LINCS DCIC). LINCS is the Library of Integrated Network-based Cellular Signatures (LINCS), a group of NIH-sponsored centers, each of which is tasked with characterizing how various cells, tissues and networks respond to disruption by drugs or genetic perturbations. The centers are producing a variety of different data types, including gene expression, epigenetic changes, proteomics, and images.

BD2K-LINCS DCIC, under the leadership of **Avi Ma'ayan, PhD**, associate professor of pharmacology and systems therapeutics at the Icahn School of Medicine at Mt. Sinai, is charged with integrating these diverse data. His team will pull all of the LINCS datasets together, standardize them with ontologies, and also integrate them with data from elsewhere. "We're organizing the data into networks based mostly on correlations between diseases, side effects, genes and drugs and bringing it all together," Ma'ayan says.

For example, the team might have a matrix of merged LINCS experiments where each column is a different experiment—a drug treatment for a single cell or tissue type, say—and the rows are the gene expression responses. This matrix might then be converted into a gene-gene similarity network based on similarities among the drugs' effects or among groups of genes with a correlated response. Then researchers can look at the overlap between these networks and other networks—for example a network of known drug side effects. "If there is some relationship [between the known side effects and gene expression], it can be very powerful," Ma'ayan says. "Now you can take new drugs and predict their side effects ahead of time." Performing this same operation over numerous different cell types and drugs becomes a vast integration task with enormous potential to learn new things.

Ma'ayan hopes the lessons of LINCS data integration will be broadly applicable to the larger BD2K effort. "We want to bring the BD2K effort into LINCS and make LINCS part of the BD2K effort."



Ma'ayan and his colleagues have already developed a drug/cell-line browser for LINCS. Users can select a dataset and then visualize the effect of more than 100 drugs on various different cell lines by tissue, mutation, gene expression profile, and drug sensitivity. Shown here: a visualization of cancer cell lines and their sensitivity to the drug SB590885, a Raf inhibitor, with the top 25 percent of the most sensitive cell lines highlighted with circles. The vertical bar graph shows that skin is most sensitive to the drug, which is consistent with the known role of B-Raf in many melanomas. Reprinted with permission from Q Duan et al., Drug/Cell-line Browser: interactive canvas visualization of cancer drug/cell-line viability assay datasets,* Bioinformatics *30 (22): 3289–3290 (2014). The browser is freely available at http://www.maayanlab.net/LINCS/DCB/

Combining Knowledge with Data for Breast Cancer Pharmacogenomics

To understand why standard chemotherapy works for some breast cancer patients and not others, researchers at Mayo Clinic in Rochester, Minnesota, are sequencing and comparing the genomes of patients' tu-

morous and normal tissue. But they'd like to evaluate their patients' tumors in light of what's already known about the various treatment options and the mutations they find. For example, how have those mutations been annotated in another context? This can be difficult because bench biologists typically use algorithms suited to analyzing spreadsheets, whereas huge public datasets have been analyzed using graph or network approaches, says **Saurabh Sinha, PhD**, associate professor of computer science at the University of Illinois, Urbana-Champaign (UIUC). "There's been very little work doing both at the same time."

Bridging that gap is a major goal of the new BD2K center called **KnowEnG, a Scalable Knowledge Engine for Large-Scale Genomic Data** under PI **Jiawei Han, PhD**, professor of computer science, also at UIUC. KnowEnG, when implemented as a cloud-based resource, will enable users to perform spreadsheet analysis in the context of the existing network of genomics data—without downloading the network. Individuals will be able to explore how their data fits in with large networks of public domain datasets, such as the STRING database, which describes protein-protein interactions, and Genemania, which is like a Google search engine for genes, where researchers input a set of genes and retrieve genes that are related in some way, drawing from all available genomics information.

Various KnowEnG team members already have working systems for analyzing large graphs, analyzing spreadsheets in scalable ways, and putting genomics datasets in scalable structures. "The initial one to two years of work will be about connecting these pieces together," says Sinha, who is in charge of the KnowEnG data science core.

Mayo Clinic will use KnowEnG to analyze breast cancer pharmacogenomics data in hopes that community knowledge will shed some light and generate testable hypotheses to improve chemotherapy outcomes. KnowEnG will also be applied to projects as diverse as exploring the relationship between gene expression and human behavior and predicting which bacterial strains are likely to produce novel antibiotic agents. "We're looking forward to having the KnowEnG framework tested on the frontlines," Sinha says.

Combining Data and Mechanics to Enhance Mobility

To evaluate surgical or therapeutic treatments for mobility problems resulting from conditions such as running injuries, osteoarthritis or cerebral palsy (a neurological disorder that affects movement and muscle coordination), the medical profession relies largely on trial and error. The Mobilize Center seeks to establish a different paradigm: the use of big data to optimize treatment.

To succeed, they need to find ways to ensure that imperfect data can yield useful information. Motion-capture data to study human mobility is often collected and recorded using different protocols at multiple labs and hospitals around the world. These data can be incomplete, noisy, imprecise, and heterogeneous, and integrating them with notoriously unreliable health records for the same individuals makes things even messier. Throw in the variable of change over time—from periodic clinic visits, for example—and it's a wonder researchers don't just throw up their hands. Fortunately, scientists—including those associated with the Mobilize Center—are becoming quite adept at handling data problems without tossing the baby out with the bathwater. "Part of the big challenge is that if you're trying to gain insight you can't expect to rely on perfect data," Delp says.

The Center will address data imperfection by building on a system called DeepDive developed by **Chris Ré, PhD**, assistant professor of computer science at Stanford and a data science core lead for the Center. Using statistical inference techniques, DeepDive can not only integrate diverse data types but also take imprecision into account and deliver probabilities that an assertion is true. Meanwhile, **Trevor Hastie, PhD**, professor of statistics at Stanford will lead a second data science effort to extract insight from time-varying mobility data spanning from seconds in duration to years.

But even after managing the data imperfection problems, what's left is still just data without any of the advantages of accumulated expert knowledge. The typical big data project looks at all the data and makes inferences based on statistics, says Delp. Sometimes this yields wonderful, insightful surprises, but it can also yield meaningless correlations among bizarre variables no one pays attention to, he says. To learn more from statistical techniques, he and his team at the Mobilize Center will bring statistical learning together with mechanistic understanding—knowledge of how something works based on the fundamentals of physics and biology. "By combining these approaches, you simplify your big data problem and gain insights that are meaningful to the biomedical researchers or clinicians," Delp says.

To learn more from statistical techniques, Delp and his team at the Mobilize Center will bring statistical learning together with mechanistic understanding—knowledge of how something works based on the fundamentals of physics and biology. "By combining these approaches, you simplify your big data problem and gain insights that are meaningful to the biomedical researchers or clinicians," Delp says.

For example, statistical learning across a large dataset of children with cerebral palsy might identify 23 variables that predict the outcome of a particular surgery intended to improve the patient's ability to walk. But a person with a mechanistic understanding of cerebral palsy gait might be able to select the three variables that can be easily measured and will give surgeons most of what they need to know. "That is so much more powerful to a clinician, when you are getting to the essence of how things work," Delp says. "Finding ways to combine mechanistic understanding with statistical methods is one of the tools the Mobilize Center will develop."

Extracting and Predicting Phenotypes

Researchers would like to be able to look at big data resources—such as electronic health records or collections of brain images—to easily determine a patient's disease status as well as predict how illnesses such

as breast cancer or Alzheimer's disease will progress. But "a lot of phenotypes are tricky to interpret or predict," says **Mark Craven, PhD**, professor of biostatistics and medical informatics at the University of Wisconsin, Madison. For example, it can take several months to design algorithms to extract a single phenotype—type 2 diabetes, say—from electronic medical records, just to identify a cohort of cases and controls to study. And even with a three-million-voxel brain image, it's hard to predict whether a patient will progress to Alzheimer's disease.

The new BD2K-funded **Center for Predictive Computational Phenotyping** (CPCP) under Craven's leadership is hoping to improve the methods for extracting and predicting phenotypes from electronic health records (EHRs), images or other large datasets such as transcriptomic or epigenomic data—as well as combinations of these different data types. "One of the interesting challenges is how you can leverage all of the different data sources," Craven says.

Like the Mobilize Center, CPCP will wrangle with datasets that are sparse, incomplete or untrustworthy, as EHRs typically are. Trying to identify something that should have been explicitly recorded in these records (such as a diagnosis) is surprisingly challenging. But it's even harder to extract information that is not explicitly measured, such as disease duration, risk factors for complications, or the effectiveness of a particular treatment. So CPCP will work on developing improved and streamlined approaches for extracting information from electronic health records, with an initial focus on such illnesses as heart attacks, asthma, and VTE (venous thromboembolism, a type of blood clot). For example, since reduced blood volume is a risk factor for VTE but is not directly recorded in the EHR, they will try to identify a constellation of other information that could be used to infer reduced blood volume.

Through its "Value of Information" lab, CPCP is also interested in using the data they already have to predict, in an optimal way, what information would add the greatest value and produce the greatest increase in predictive power, essential functionality for this information-rich age. Such a capability could help decide the minimum set of tests needed to arrive at a diagnosis for a patient, or which additional experiments a researcher should do to best understand a system. In the context of computational phenotyping, it would also help researchers extract the most predictive information from an EHR or image.

Accurately modeling causation in a biological system is challenging. The sheer number of variables can raise millions of chicken and egg questions about what aspect of a system caused another, not to mention whether a hidden variable (the rooster next door?) has a causal influence.

Finding Causation

Accurately modeling causation in a biological system is challenging. The sheer number of variables can raise millions of chicken and egg questions about what aspect of a system caused another, not to mention whether a hidden variable (the rooster next door?) has a causal influence.

The new BD2K-funded **Center for Causal Modeling and Discovery** (CCMD) **of Biomedical Knowledge from Big Data** is dedicated to deriving causal insight from the huge numbers of variables often present in biomedical data. For example, they will look for patterns that suggest causation among millions of variables involved in cancer signaling pathways, including genomics and gene expression data as well as data on cell function or dysfunction. Similarly, for chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis, they will hunt down causal factors in high-throughput data that includes gene expression, DNA methylation, and microRNA data as well as clinical records. "The algorithms can be applied to a wide variety of biomedical data," says **Greg Cooper, MD, PhD**, professor of biomedical informatics at the University of Pittsburgh and contact PI of CCMD. Having highly efficient algorithms to search for likely causal patterns in the vast biomedical data we have today could be a key step toward prevention and/or treatment for a huge range of diseases.

CCMD plans to provide "one-stop shopping" for these types of high quality causal discovery algorithms. Some algorithms search over different possible networks to determine which are most strongly supported by the data while accounting for prior knowledge and belief based on the scientific literature (using Bayesian methods). Others look for patterns of independencies and dependencies among the variables that suggest particular causal relationships.

By applying these methods to diverse biomedical problems and making the algorithms available through application programming interfaces (APIs), the Center will ensure that they are both broadly applicable and available to the research community. "These APIs are one key deliverable of our center," Cooper says.

POWER TO THE PEOPLE: *Distributing Algorithm Development*

In 2009, the ENIGMA Consortium (Enhanced Neuro-Imaging Genetics through Meta-Analysis) was launched to bring together researchers—and their genomic and imaging data—to get a better understanding of brain function and disease. The Consortium, which has now grown to over 300 researchers, has access to genome-wide, neuroimaging, and clinical data from more than 31,000 subjects worldwide. Using this impressive amount of data, they are studying ten major brain diseases, including schizophrenia, major depression, bipolar illness, attention deficit hyperactivity disorder, and autism. They can look for genes associated with these diseases, examine differences in how the brain reacts to different drugs, and trace how different parts of the brain are connected to one another in people with and without brain disease.

The beauty of ENIGMA is that the results are achieved without shipping data

around the world. Instead, small groups within the Consortium who wish to take a crack at a particular research question form an alliance to help each other out. They develop algorithms and distribute them to others in the Consortium, essentially enabling a meta-analysis across multiple centers. "They can get off the ground quickly with tens of thousands of data points by sending software out," says P**aul Thompson, PhD**, professor of neurology at the University of Southern California.

Thompson, who is the director of the ENIGMA Consortium, is now also the PI for the new BD2K-funded **Enigma Center for Worldwide Medicine, Imaging and Genomics**. The Center's work lies squarely in the area of developing analytical and statistical tools for big data, but it is funded as 20 sub-awards to researchers around the globe. "Data is nothing without people," Thompson says.

ENIGMA Center researchers will develop refined algorithms that can analyze brain maps, measures and signals, and relate them to genomic, environmental and epidemiological data as well as clinical outcomes.

"There are some alliances you can form that make it easier for everybody to do science," Thompson says. "We hope to see discoveries on a scale that hasn't been possible," he says.

The ENIGMA alliance studies brain scans and DNA at more than 185 sites around the world. They created working groups to pool and compare data from many neuroimaging centers in order to understand the effects on the brain of various conditions, including bipolar disorder, major depressive disorder (MDD), addiction and schizophrenia. The result is a data pool with tens of thousands of subjects. The institutions involved in the working groups are shown on this map from June 2013. Thompson PM et al., The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data,* Brain Imaging Behav. *2014. Epub 2014/01/09. doi: 10.1007/s11682-013-9269-5. PubMed PMID: 24399358.

DATA STANDARDS AND METADATA: *Sorting the Nuts and Bolts*

Basic tasks, such as storing and accessing data efficiently, may require little or no attention when researchers work with small datasets. But these foundational issues must be addressed head on when datasets become enormous. And several BD2K Centers are doing just that: creating data structures that allow for efficient storage of and access to big data; and developing ways to easily tag the data with annotations or "metadata" so they carry signatures of where they came from as well as how they've been used through time.

These are not simple problems. They are also not glamorous. "Most people are thinking about the great discoveries they are going to be making with the data and—make no mistake—we are too," says **David Haussler, PhD**, professor of biomolecular engineering at the University of California, Santa Cruz, and PI of the new BD2K **Center for Big Data in Translational Genomics** (CBDTG). "But we're also emphasizing the need to get the nuts and bolts right before making discoveries."

Establishing Data Standards

In cooperation with the Global Alliance for Genomics and Health (GA4GH), a nonprofit consortium of genomics researchers worldwide, CBDTG will develop and implement global standards for genomics data.

The Center's effort builds on work begun by the Thousand Genomes Project, which has already pioneered several novel file formats—BAM for storage of large files of DNA reads and VCF for storage of files called variants. GA4GH will make these formats ready for prime time and clinical use, as well as create an additional compressed format, called CRAM, that Haussler says will save millions of dollars in space costs for storing large genomics files. CBDTG will work with them to build ab-

stract data schemas so that the data can be stored efficiently and optimally accessed. "It would be hopelessly inefficient to paw through the coming massive amounts of genomics data to get the information you want if it is stored in the current file formats," Haussler says.

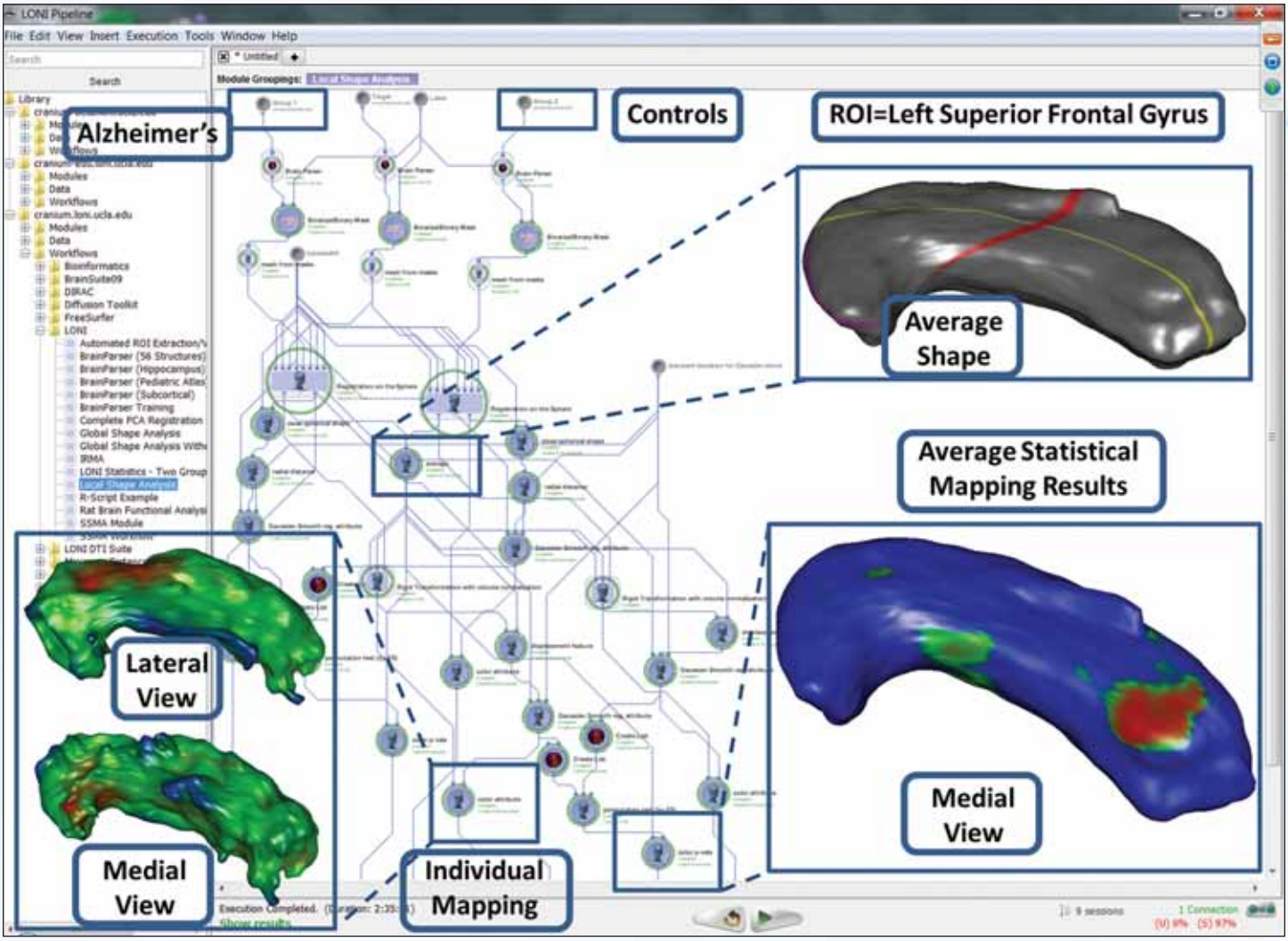
At the same time, Haussler's team is working with GA4GH on standards for representing genetic variation—not only single nucleotide changes but also rearrangements and duplications. "If we already knew all possible human variations, it would be a lookup problem. We'd have a name for each variation," Haussler says. "But that's not the case. Every individual's genome will reveal new variations."

CBDTG will also help the other BD2K centers deal with the federation of data across multiple locations. "We need something like URLs that identify the objects you're looking for," he says. "You shouldn't really care where it is." Haussler's center is working with Google, Amazon and Microsoft to address these problems at a fundamental level. One solution involves attaching a cryptographic signature to pieces of data that verify what they are, sort of like bar codes on a package. But there are lots of potential stumbling blocks. For example, there can be multiple copies of the same data, and one small thing can change in one copy. "It can become a nightmare in electronic librarianship," Haussler says. "We need rules of the road for how data are represented, verified, changed, stored back in, never lost, never compromised. It's great to be working with the biggest and best on that."

Making Annotation Happen

Ever since the 1980s when researchers were first required to post certain experimental datasets online, they have been required to also create metadata—basic information about how the data were produced. But scientists don't always do a thorough enough job of it.

"There's a real 'what's in it for me' problem to overcome," says **Mark Musen, MD, PhD**, professor of medicine at Stanford University and PI of the new BD2K **Center for Expanded Data Annotation and Retrieval** (CEDAR). And that threatens re-



One goal of the BD2K program is to provide computational tools in a format that can be easily used by biomedical researchers. This image shows an example of the pipeline workflow for a specific brain-mapping problem—local shape analysis—developed by Toga and his colleagues. The example here starts with the raw magnetic resonance imaging data for 2 cohorts (11 Alzheimer's disease patients and 10 age-matched normal controls), extracts a region of interest (left superior frontal gyrus [LSFG]) for each subject, and generates a 2-D shape manifold model of the regional boundary. Then the pipeline computes a mean LSFG shape using the normal subjects' LSFG shapes, co-registers the LSFG shapes of all subjects to the mean (atlas) LSFG shape, and maps the locations of the statistically significant differences of the 3-D displacement vector fields between the 2 cohorts. The insert images illustrate the mean LSFG shape (top-right), the LSFG for one subject (bottom-left), and the between-group statistical mapping results overlaid on the mean LSFG shape (bottom-right), red color indicates p-value < 0.01. Reprinted with permission from Dinov, ID et al., Applications of the Pipeline Environment for Visual Informatics and Genomics Computations,* BMC Bioinformatics, *12:304 (2011).

searchers' ability to rely on, replicate, or share one another's data. "Until we make it simple to annotate data in a clear way, we're going to have serious problems for science," Musen says.

The Center's team already has a few tricks up its sleeve for simplifying annotation, including technologies developed through earlier projects such as Protégé, an ontology development system, and BioPortal, an ontology repository developed by the National Center for Biomedical Ontology. CEDAR will leverage these two technologies to automatically create Web-based interfaces for filling out metadata templates. CEDAR will also create text analysis and predictive data tools to fill in portions of the templates automatically.

In addition, CEDAR will develop ways of allowing metadata to evolve as the data are re-analyzed or compared to other data. "Metadata are not a static description of an experiment," Musen says. "They are an evolving record of the conversations researchers have about experiments over time."

The CCMD team also plans to develop tools for annotation—but for application to causal models that are derived from analyzing biomedical data, rather than to the raw data. "Just as meta-information about data (metadata) can have significant value, so too can meta-information about the models derived from that data," says Cooper.

Whether simplifying metadata collection will incentivize researchers to do a better job remains to be seen. "I hope the overall ecosystem CEDAR creates will show researchers that the authoring of metadata need not be onerous, and that science has much to gain from first-rate annotations," Musen says.

Should CEDAR's researcher-driven strategy prove insufficient, **Andrew Su, PhD**, of the Scripps Research Institute has a different idea. As part of **The Heart of Data Science**, a new BD2K center based at the University of California, Los Angeles, under PI **Peipei Ping, PhD**, Su will enlist the help of citizen scientists who will extract information from the biomedical literature and annotate proteomic data for cardiovascular research. To evaluate the quality of these annotations, the crowdsourced work will be compared to the Reactome and Intact databases, which are curated by experts, says **Henning Hermjakob, PhD**, who will head up the center's data science core.

Ping's center will also showcase the value of metadata for discovery by expanding the Proteome Xchange, a consortium that aggregates proteomic metadata into a searchable centralized form. The Center will add more proteomics repositories to the Exchange as well as extend it to include other –omics data types, such as metabolomics. This program will not only motivate better metadata collection but also address "the absolutely nontrivial problem of finding which datasets are relevant to a particular research project," Hermjakob says.

OFFER IT TO THE WORLD: *User-Friendly Interfaces*

Thompson likens big data research to a lengthy relay race. The baton passes from labs into structured datasets; becomes integrated with other data; is subjected to analysis using novel tools and creative mathematics; and finally is presented in a user-friendly interface for others to use. That's the victory lap for the BD2K Centers: providing a means for other researchers or clinical personnel to use big data effectively.

"Great integrative work and scalable computing will amount to nil if the interface isn't immediately appealing to the biologist," says Sinha. The KnowEnG team's interface will allow biologists to identify genes that discriminate between samples as well as probe the literature for relevant information about those genes.

Similarly, CCMD will create a workstation for biomedical scientists so they can easily select datasets, apply causal discovery algorithms, see results graphically, and annotate and store them. And the PIs of other centers have similar plans.

But the new BD2K-funded **Big Data for Discovery Science** (BDDS) **Center** under PI **Art Toga, PhD**, provost professor and director of the Laboratory of Neuro Imaging at the University of Southern California, has interfaces as a focus. They're creating a smart pipeline that offers big data analysis tools for use by non-cognoscenti. It offers drag and drop glyphs in a graphical environment that is layered with expertise to guide users toward the appropriate tool for a given task. "The system is self-aware," says Toga. It will not only show publications about a particular tool, but also offer information about parameter settings based on past experience—acting as a sort of advisor on best practices. Toga says the pipeline has to make tools understandable within five to ten minutes, and also give users feedback when they make a mistake. "If we have to develop complicated user manuals, we've failed," he says.

The pipeline will also include novel ways to interrogate data casually, looking for relationships and providing instantaneous insights to drive hypothesis generation. For example, researchers could peruse large, integrated datasets to look for relationships between two cohorts that differ in only one particular feature. "We built a prototype of such a thing—with the computation attached to the database—and played with it and it was unbelievably useful," Toga says.

Though Toga will test the pipeline using neuroimaging, genomics and proteomics tools, nothing precludes its use in other scientific domains. "The workflow is agnostic as to the tool type," Toga says.

A VIRTUOUS CYCLE

Data science methods can't be developed in a vacuum. "You have to think, what does the method do; what can you learn; what problem are you solving," Delp says. It's an approach designed to ensure the development of big science methods that work—and the first step in assuring that they'll also be useful to others.

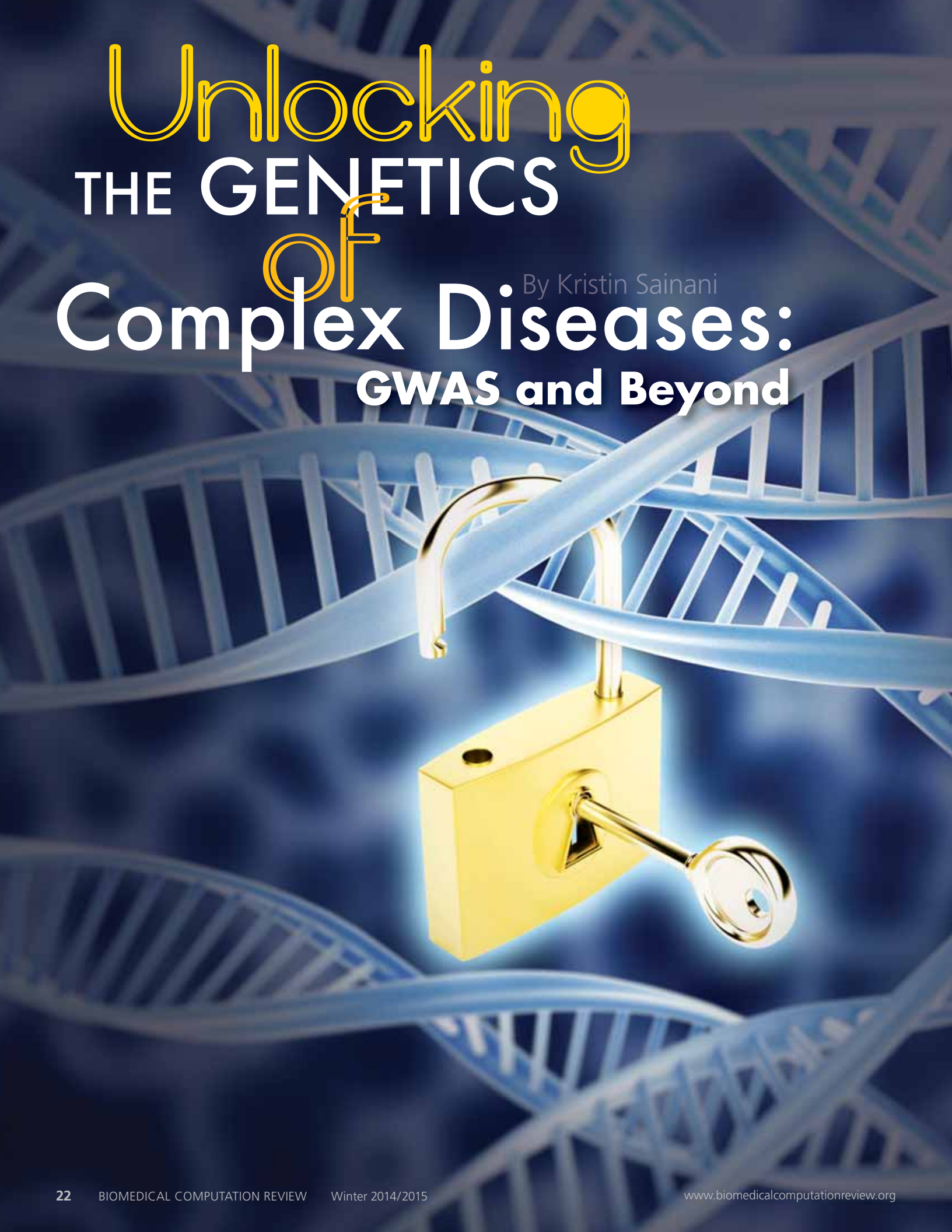
As Bourne puts it, "We view this as a virtuous cycle." The researchers are motivated by the biomedical research that gets done, and in the process of doing that work they generate data, use data, and develop tools that all get "virtuously" shared with others to provide further motivation. "Sharing the data and the software across the centers and to other investigators and beyond is key," he says.

So too is cooperation: Where the Centers can work synergistically with one another and the NIH, they will do so. For example, both the Mobilize Center and MD2K will be exploring the effectiveness of using wearable sensor data to change unhealthy behaviors. Ontologies and annotations, the focus of CEDAR, play a role in a number of the Centers, including CCMD and the Heart of Data Science. And nearly all the Centers have to find ways to deal with the sparsity, noisiness, and heterogeneity that so often characterizes big data. By tackling these challenges together, Collins says: "The whole is going to be a lot greater than the sum of its parts." □

Unlocking the GENETICS of Complex Diseases:

GWAS and Beyond

By Kristin Sainani



Some diseases, such as cystic fibrosis, have a beautiful simplicity: A genetic misspelling cripples a protein, which profoundly and predictably alters the body. Find the faulty gene for these so-called "Mendelian" diseases and you instantly reveal the biological story of the disease. Scientists have long hoped that, with the right genetic tools, solving the biology of more complex yet highly heritable diseases would be similarly elegant and definitive.

Indeed, the first genome-wide association study (GWAS), published in *Science* in 2005, generated enormous excitement. Comparing 116,204 genetic markers (single nucleotide polymorphisms, or SNPs) between just 96 cases and 50 controls, researchers discovered a genetic variant that was strongly related to age-related macular degeneration, increasing risk seven-fold for those carrying two copies.

Despite such successes, initial enthusiasm soon flagged as numerous GWAS revealed a far more nuanced and messy genetic landscape than anticipated: Hundreds of genes are involved in most complex diseases, and most raise the risk of disease just a small amount—on the order of 10 to 30 percent. Collectively, these genes explain only a fraction of disease heritability, what some have coined the "missing heritability" problem. By 2009, critics lamented that we had wasted hundreds of millions of dollars obtaining "surprisingly little new information."

These criticisms cast a long shadow over GWAS, but they were largely unfounded. "I've been quite bemused and surprised by the strange criticism of GWAS. Because it's telling us something about the state of nature. And we shouldn't be apologetic for that. That's just the way it is," says **Peter Visscher, PhD**, professor and chair of quantitative genetics at the University of Queensland.

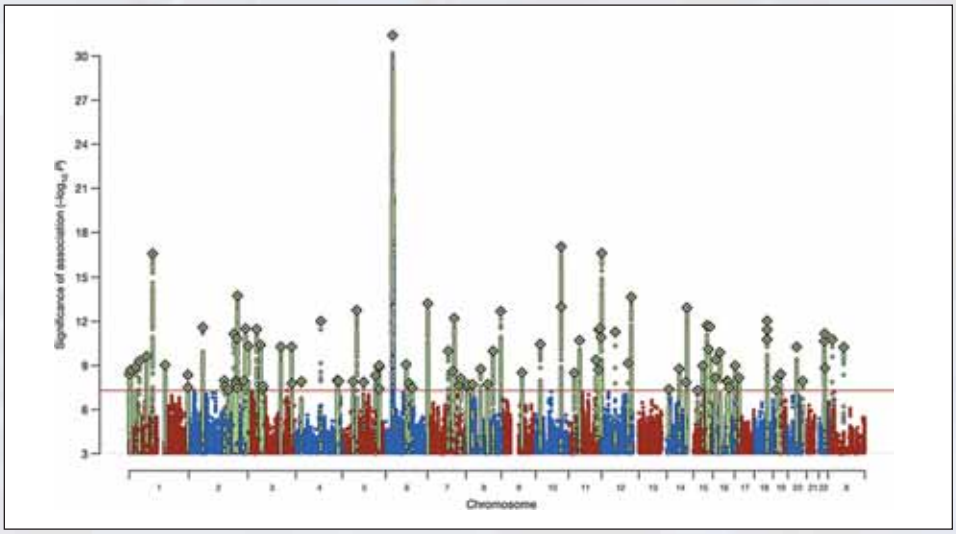
Given the messy genetic reality of complex diseases, GWAS have delivered exactly what they are capable of delivering: not answers, but an enormous number of clues. To date, GWAS have reliably linked thousands of genetic variants to hundreds of complex diseases or traits. "Many variants have now been found for diseases where maybe five years ago there was absolutely nothing," Visscher says. "Schizophrenia is a good example. Before 2009, there was no gene or variant that was robustly associated with the risk of schizophrenia. Now there are more than 100 loci that biologists are starting to follow up on."

Ever-larger GWAS will continue to contribute knowledge about complex diseases; and biologists will continue to pursue these leads. But the bigger breakthroughs may come from tweaking or building on GWAS as well as taking complementary approaches. A few such alternatives are beginning to pay dividends, including systems biology approaches; prioritizing GWAS hits; hunting for rare genetic variants; reversing GWAS; and exploiting the overlap between diseases, including between complex and Mendelian diseases.

"I think we're much closer than we have ever been—and probably closer than we realize—to understanding how genome variation affects disease risk," says **Nancy Cox, PhD**, professor of medicine and of human genetics at the University of Chicago. "It's an incredibly exciting time to be in genetics."

TAKING A SYSTEMS APPROACH

Complex disease genes likely exert their effects through small perturbations in biolog-



GWAS hits. *A GWAS comparing tens of thousands schizophrenia cases and controls turned up 108 genetic loci associated with schizophrenia (loci above the line have achieved genome-wide statistical significance). Many variants are located next to genes that operate in the brain or immune system—suggesting a possible link between the immune system and schizophrenia. Reprinted by permission from Macmillan Publishers Ltd: Schizophrenia Working Group of the Psychiatric Genomics Consortium, Biological Insights from 108 Schizophrenia-associated Genetic Loci,* Nature *511(7510):412-3 (2014).*

ical pathways. Rather than turning proteins on or off, for example, they may subtly alter the amount of proteins produced. Indeed, many studies have found that GWAS hits are substantially enriched in variants that affect gene expression, says **Tuuli Lappalainen, PhD**, assistant professor of systems biology at Columbia University and group leader at the New York Genome Center. Case-in-point: The strongest obesity-related GWAS hit—found in the FTO locus—was originally thought to affect FTO protein; but recent studies show that it exerts its effects by regulating a more distant gene, IRX3.

Thus, researchers need to consider how GWAS variants fit together into the larger biological picture, rather than focusing on them one at a time. If researchers can link GWAS hits together into pathways, they get immediate insight into the underlying biology; it also gives them a place to look for additional disease-related variants.

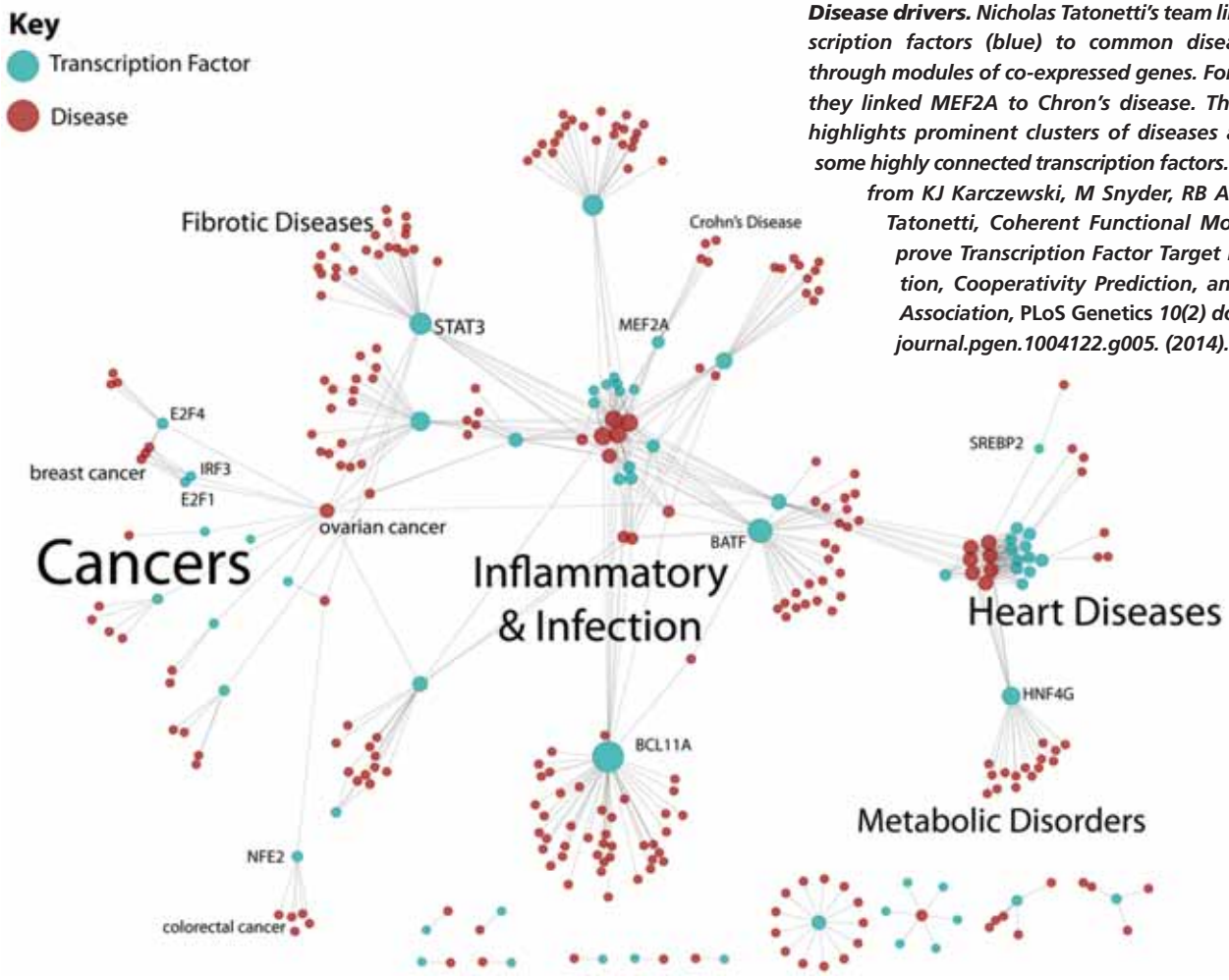
To link GWAS hits into pathways, some researchers are hunting down the transcription factors (TFs) that initiate the expression of clusters of genes in concert and thus may play a role in complex disease. Data from high-throughput experiments called "ChIP-Seq" can yield valuable information about where TFs bind to the genome, but because TFs bind to many genes beyond their primary targets, ChIP-Seq datasets alone are not enough. So **Nicholas Tatonetti, PhD**, assistant professor of biomedical informatics at Columbia University, decided to integrate ChIP-Seq data from ENCODE (ENCyclopedia Of DNA Elements) with other sources of information.

They turned to a paper by **Jesse M. Engreitz**, who did the work while a student in Russ Altman's lab at Stanford University. Engreitz used a statistical technique called independent components analysis (ICA) to identify 423 "gene expression modules"—sets of genes that are highly co-expressed and likely represent functional units. ICA is best known for its ability to solve the "cocktail party problem," Tatonetti explains—using data recorded on multiple microphones in a noisy room, ICA can isolate one individual's voice. "It came to us that what Jesse was really doing was identifying transcription factor signals," Tatonetti says. "Just like the microphone records a mix of people's voices, the gene expression arrays are recording the mixed signals of the transcription factors."

By overlaying these 423 gene expression modules on the ChIP-Seq binding data from ENCODE, Tatonetti's team was able to connect specific transcription factors to specific modules. Then, using data from GWAS catalogs, they found that some of these gene sets were also enriched with GWAS hits for a particular disease. "So those two links allow us to go from transcription factors through the modules to disease," Tatonetti explains. The work turned up a number of known associations between transcription factors and diseases, confirming that the method works. It also identified 458 novel transcription factor–disease links.

In an independent study, the team validated one of these leads—between MEF2A and Chron's disease. Both MEF2A itself and MEF2A-controlled genes were more highly expressed in 59 patients with Chron's disease than in 42 controls. MEF2A had previously been implicated in heart disease, but never in Chron's disease. The study was published in *PLoS Genetics* in 2014.

"In the future, we hope to take the sys-



Disease drivers. *Nicholas Tatonetti's team linked transcription factors (blue) to common diseases (red) through modules of co-expressed genes. For example, they linked MEF2A to Chron's disease. This graphic highlights prominent clusters of diseases as well as some highly connected transcription factors. Reprinted from KJ Karczewski, M Snyder, RB Altman, NP Tatonetti, Coherent Functional Modules Improve Transcription Factor Target Identification, Cooperativity Prediction, and Disease Association,* PLoS Genetics *10(2) doi:10.1371/journal.pgen.1004122.g005. (2014).*

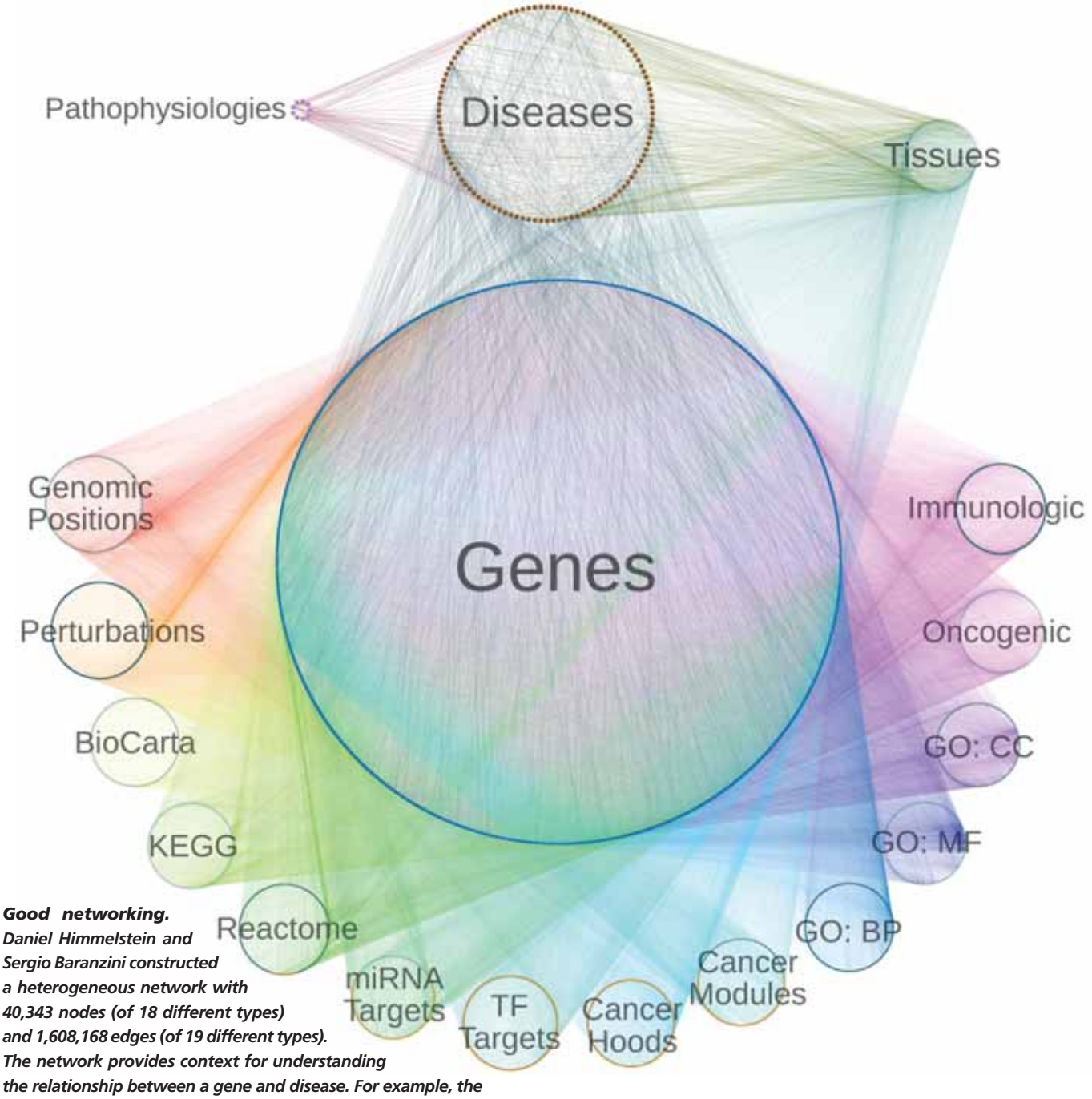
tems approach even further by incorporating non-molecular data, including environmental and clinical data," Tatonetti says.

PRIORITIZING GWAS HITS

Because GWAS researchers compare millions of SNPs between cases and controls, many SNPs may appear to be associated with the disease just by chance. To minimize false positives, researchers use a much more stringent significance cut-off than is traditionally required for statistical significance. But there's a cost to this rigor: many SNPs that are truly related to the disease don't make the cut. Increasing GWAS sample sizes makes it easier to find these true hits, but at considerable expense.

To help differentiate the gems from the duds among the "second-tier" SNPs—those that showed some signal, but not enough to be declared robust hits—without running ever-larger GWAS, researchers are combining GWAS results with other sources of evidence. For example, if a GWAS study for schizophrenia identifies a gene that is also already known to be expressed in brain tissue, researchers might conclude that the hit actually relates to the disease.

Daniel Himmelstein, a doctoral student in biological and medical informatics at the University of California, San Francisco, estimates the probability that a gene is associated with a specific disease by drawing on multiple and varied high-throughput datasets—combining, for ex-



Good networking. *Daniel Himmelstein and Sergio Baranzini constructed a heterogeneous network with 40,343 nodes (of 18 different types) and 1,608,168 edges (of 19 different types). The network provides context for understanding the relationship between a gene and disease. For example, the network could indicate whether a candidate gene participates in the same pathways as known disease genes. Using the topology of the network, their approach learns the mechanisms underlying known gene-disease associations to predict novel associations. Courtesy of: Daniel Himmelstein and Sergio Baranzini, UCSF. Preprint available at bioRxiv, the preprint server for biology, http://dx.doi.org/10.1101/011569.*

ample, data on transcriptional signatures, protein interactions, and gene functions. "In the past, people have only focused on one aspect, such as protein-protein interaction," says Himmelstein, who works in Sergio Baranzini's lab. "We've tried to take it to the next level by integrating more types of data."

But this strategy faces challenges of its own: Algorithms that work for one type of data don't necessarily scale to complex networks with multiple entities and types of relationships (called heterogeneous networks). So Himmelstein and his colleagues adapted a tool from social network analysis—an algorithm called PathPredict that makes predictions about the future (or unknown) connectivity of pairs of objects based on past connectivity. Though originally used to predict future co-authorships among scientists, Himmelstein says, "We realized it could work to paint an understanding of how a disease was related to a gene by understanding the topology of connections between them."

The team validated the approach by hiding the then-largest multiple sclerosis (MS) GWAS from the algorithm. Using only the results from smaller multiple sclerosis GWAS, their method assigned high ranks to all 37 protein-coding genes that were in fact discovered by the masked study. The approach also gave high ranks to other genes as well. Of the top four newly identified MS susceptibily genes, three were successfully validated with independent data.

"We predicted not all but quite a bit of the larger GWAS. So if we don't have the funding to do a larger GWAS, we can use these types of techniques to build off the existing data," Himmelstein concludes. "It's cool that you can do so much without having to spend any more money or recruit any more patients." The team has applied the method to 29 complex diseases; and the ranked variants are publicly available at het.io. "So other people can use our results for prioritization," Himmelstein says.

CHASING RARE VARIANTS

Natural selection weeds out highly deleterious mutations from a population. Thus, the genetic changes with the biggest impact on disease risk tend to occur infrequently. GWAS chips only capture SNPs found in at least a few percent of the population and thus miss rare variants—precisely those that may offer the most exciting biological insights. Some scientists even believe that these neglected rare variants explain much of the "missing heritability" of complex diseases.

"It's not clear how much of the inter-individual variability in risk for disease is driven by rare variation," Cox says. "But when we can find that variation—really rare stuff with big effects—it often gives us a disproportionate understanding of the biology."

To find rare variants, scientists must compare entire gene sequences between cases and controls. In the past, this has meant looking at only a handful of genes at once. But with the advent of next-generation sequencing, scientists are beginning to look for rare variants in a more systematic, large-scale way—comparing entire genomes or exomes (protein-coding genes) in what some have called a "Rare Variant Association Study," or RVAS.

Because you need to sequence a lot of people's DNA to pick up rare events, sample size requirements for RVAS will likely be as big as for GWAS, says **Benjamin Neale, PhD**, assistant professor in the Analytic and Translational Genetics Unit at Massachusetts General Hospital, and an associated researcher at the Broad Institute. Given the cost of sequencing, most RVAS studies to date haven't been that large. Even so, moderate-size RVAS with clever designs have turned up high-impact results.

"It's not clear how much of the inter-individual variability in risk for disease is driven by rare variation," Cox says. "But when we can find that variation—really rare stuff with big effects—it often gives us disproportionate understanding of the biology."

As an alternative to RVAS, some researchers focus on *de novo* genetic mutations—changes found in a child but not in the parents. Autism researchers, for example, have identified numerous rare variants using this approach. "*De novo* mutations have a lot of clear advantages in analysis and interpretation," Neale says. On average, exomes contain just one *de novo* mutation, which greatly narrows down the potential genetic culprits. Also, they are easier to find because they haven't yet been weeded out by natural selection, Neale explains.

In a 2014 paper in *Nature*, researchers compared whole-exome sequences in 2517 children with autism to those of their parents and unaffected siblings. They identified *de novo* events in 353 genes that would likely disrupt the corresponding protein and thus have a high chance of being causative. In 145 additional genes, protein-altering *de novo* events occurred in more than one autism case, suggesting potential causation. The genes with the most frequent hits played roles in synaptic communication, ion channels, and in proteins known to be involved in fragile-X mental retardation and Down's syndrome, among others.

Related rare variants have also been identified in schizophrenia. In a 2014 paper in *Nature*, researchers sequenced the exomes of 2536 cases with schizophrenia and 2543 unrelated controls. Individuals with schizophrenia had a significantly higher rate of rare disruptive mutations in protein-coding genes that were loosely suspected to play a role in schizophrenia. Moreover, disruptive mutations in 28 genes related to synaptic activity appeared in 9 cases versus none in controls; and disruptive mutations in 26 genes involved in calcium ion channels were found in 12 cases versus only one in controls. Genes in these two gene sets appear to explain about one percent of schizophrenia cases. "So that's consistent with the idea that there are many rare variants scattered throughout the genome, some of which probably confer risk for schizophrenia," Neale says.

Focusing on numerical traits (e.g., biomarker levels) rather than binary ones (e.g., disease/no disease) also increases statistical

power to detect effects. In a 2014 paper in the *New England Journal of Medicine*, researchers from the Broad Institute sequenced whole exomes of 3734 individuals and correlated these with plasma triglyceride levels. They found that carriers (about 1 in 150 people) of rare loss-of-function mutations in the APOC3 gene had 39 percent lower triglyceride levels than non-carriers, as well as better cholesterol levels. Using existing data from 15 studies covering more than 100,000 people, they then showed that carriers also had a 40 percent reduced risk of heart disease. Thus, it appears that disrupting the APOC3 gene is protective against heart disease—and drug companies are now following up on this lead.

"Even if rare variants don't cause a huge proportion of cases, every gene you nail this way is absolutely priceless," Cox says. "It's a wedge into the biology that we wouldn't have otherwise." Common variants that regulate these same genes may also impact the risk of complex diseases, though to a lesser extent, she adds.

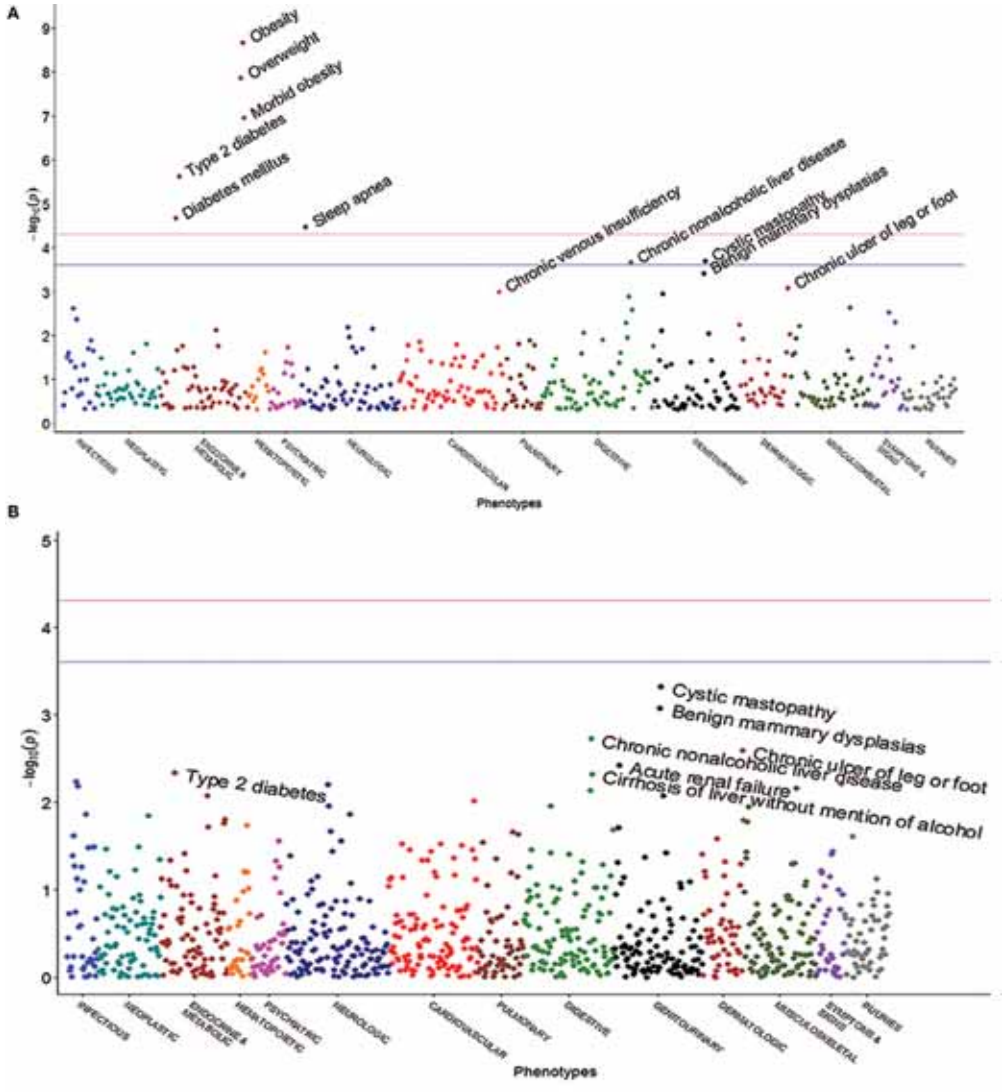
PheWAS: REVERSING GWAS

Scientists are also making inroads into complex disease genetics by focusing more on the phenotypic side of the equation. Scientists at Vanderbilt University created a new approach called a Phenome-Wide Association Study, or PheWAS. "PheWAS is essentially the inverse of a GWAS: You start with a given genetic variant and then you look at what diseases are associated with it," explains **Joshua Denny, MD**, associate professor in biomedical informatics and medicine at Vanderbilt University. PheWAS begins with genetic data on individuals for whom a rich phenotypic dataset is also available, such as in electronic medical records or a well-characterized cohort (such as the Framingham heart cohort).

Whereas GWAS consider only one disease at a time, PheWAS can look at multiple diseases and traits at once. Denny points to the FTO locus, which has been strongly associated with obesity. FTO was originally discovered in a GWAS for type II diabetes; it took additional GWAS to reveal that this locus only influences diabetes risk through its effect on weight. "When you do PheWAS on FTO, it's abundantly obvious that it's associated with obesity. You see type II diabetes and a whole host of other obesity-related phenotypes. So you wouldn't have had to run a number of subsequent GWAS to figure this out," Denny says.

With PheWAS, researchers can also look at the emergence of diseases over time, since electronic medical records contain long-term medical histories. "That gets you the power to think about the data longitudinally, which you can't do in most case-control studies," Denny says.

For example, in a 2013 paper in *Circulation*, Denny's team first performed a GWAS on 5272 genotyped patients from the eMERGE (Electronic Medical Records and Genomics) network who had previously had a normal electrocardiogram (ECG), and appeared free of heart disease at that time. They found 23 SNPs that were robustly associated with normal variation in the speed at which electrical pulses travel through the heart. In a subsequent PheWAS of 13,859 individuals in eMERGE, they linked two of these variants—in the genes SCN5A and



***Reverse GWAS.** In PheWAS, researchers scan the phenome rather than the genome. This PheWAS linked multiple phenotypes to the FTO locus. The pink line represents a more stringent cutoff for statistical significance; the blue line represents a less stringent cutoff. When researchers don't account for body mass index, many phenotypes are linked to FTO (A); however, adjustment for BMI greatly attenuates these associations (B), suggesting that FTO's effects are largely mediated through increased weight. Reprinted from RM Cronin, et al, Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index,* Frontiers in Genetics*, 05 August 2014 | doi: 10.3389/fgene.2014.00250.*

SCN10A—to reduced risks of atrial fibrillation and cardiac arrhythmias. Finally, they asked the question: What happened to the 5272 heart healthy individuals in the years (often decades) that followed their normal ECG? They found that those who carried one copy of the SCN10A variant were 20 to 30 percent less likely to go on to develop cardiac arrhythmias and atrial fibrillation; and those with two copies were 35 to 55 percent less likely. "The coolest thing about that study is that we had this ridiculously long prospective study that was just at our fingertips," Denny says.

In a paper published in *Nature Biotechnology* in 2013, Denny's team performed a large-scale PheWAS on individuals in the eMERGE dataset. They looked for links between more than 1000 clinical phenotypes and 3000 SNPs previously implicated in complex disease. The PheWAS replicated 210 GWAS findings, and also revealed 63 novel associations. In particular, they linked several genetic variants to skin conditions, including noncancerous skin growths (actinic and seborrheic keratosis) and nonmelanoma skin cancer. "We discovered a lot of stuff on skin phenotypes, probably because these have been understudied by GWAS," Denny says.

Among the most exciting findings, Denny's teams linked variants in the enzyme TERT, which helps maintain telomeres (the caps at the end of chromosomes that protect them from deterioration), to seborrheic keratosis, which produces waxy, wart-like growths. Unlike most genetic associations for skin phenotype, the effect did not appear to be mediated through sun sensitivity. Rather, variants that shorten your telomeres may speed up intrinsic skin aging, Denny explains.

PheWAS studies have also exposed numerous examples of pleiotropy—where the same gene influences multiple different clinical phenotypes. For example, variants at the 9p21.3 locus have been linked to heart attacks and blocked arteries; and Denny's team was one of the first to show that this locus is independently related to aneurysms and hemorrhoids. This finding gives clues to the genetics of all four diseases.

EXPLOITING PLEIOTROPY

Some researchers start from the assumption that there may be pleiotropy among complex diseases that share phenotypic characteristics, such as common symptoms or co-morbidities. By identifying these phenotypic overlaps, they hope to gain entrée into the underlying genetics. "For this approach to work, you have to have a big disease phenotype database," says **Rong Xu, PhD**, assistant professor of medical informatics at Case Western University. Xu is creating such a database by systematically mining the biomedical literature.

"It's a very difficult problem to extract fine-grained semantic relationships among diseases," Xu says. Her team uses natural language processing to parse the text in all abstracts in MEDLINE (22 million citations and more than 100 million sentences). Since this is a massive computing task, they use crowd computing to get it done quickly.

"It's a very difficult problem to extract fine-grained semantic relationships among diseases," Xu says.

Xu's team uses a semi-supervised pattern learning approach to extract disease-disease associations from the parsed text. For example, Xu may feed in the information that obesity is a risk factor for heart disease. The computer studies the language patterns that authors use to describe this relationship. Then the computer scans the corpus looking for similar language patterns between novel disease pairs—and infers a similar relationship.

In a 2013 paper in *Bioinformatics*, Xu's team used this approach to identify 121,359 disease pairs with overlapping symptoms; 99 percent of these relationships aren't captured in any other structured knowledge base. Her team is adding other disease-disease associations to the database, such as shared risk factors or treatments. And Xu has begun to leverage the database to predict disease genes and reposition drugs.

For example, she found that hypertension and type II diabetes have overlapping symptoms. When she pooled available GWAS results from both diseases, she turned up a novel candidate SNP that appears to be related to both diseases. The SNP showed only a weak signal in disease-specific GWAS, but a strong signal when the two diseases were pooled. "So this SNP may be underlying the mechanism of both hypertension and diabetes," Xu says.

Other researchers are exploiting phenotypic overlaps between Mendelian and complex diseases. It's well known that patients with Mendelian diseases are more prone to complex ones. Thus, Mendelian genes—and the pathways they are ensconced in—may harbor common variants that predispose to complex diseases. "It's essentially an approach to get to genetics of complex diseases using non-genetic (phenotypic) data," says **Andrey Rzhetsky, PhD**, professor of medicine and of human genetics at the University of Chicago.

In a 2013 paper in *Cell*, Rzhetsky's team looked for co-occurrences between 100 Mendelian and 100 complex diseases using more than 100 million electronic medical records from the United States and Denmark. They found 2909 associations, most of them novel. "What came out is that every complex disease has a unique set of companion Mendelian diseases, something like a bar code," Rzhetsky says. "This translates into a unique barcode of genes as well, because Mendelian diseases map to genes deterministically."

Their analysis revealed that schizophrenia, bipolar disorder, autism, and depression tend to co-occur with mutations in four genes associated with Mendelian diseases (Timothy syndrome, retinitis pigmentosa 18, and spinocerebellar ataxia). GWAS studies have identified common genetic variants in these same Mendelian genes that also predispose carriers to multiple neuropsychiatric disorders. This is just one of many examples where diverse approaches are converging on the same answers.

COMING TOGETHER

By themselves, GWAS findings are like disconnected pieces of a puzzle; they're essential—but, until they are connected to other information, or analyzed in new ways, there's no hope of seeing the bigger picture. Now, little by little, small glimpses of that picture are starting to emerge.

This year's American Society of Human Genetics meeting, for example, showcased a lot of really good science, Cox says. "It's all starting to come together," she says. "I think there's a palpable sense of excitement that things will finally start to break." □

Under TheHood

BY MADELEINE UDELL AND STEPHEN BOYD, PhD

Beyond Principal Components Analysis (PCA): Exploring Low Rank Models for Data Analysis

In many application areas, researchers seek to understand large collections of tabular data, for example, patient lab test results. The values in the table might be numerical (3.14), Boolean (yes, no), ordinal (never, sometimes, always), or categorical (A, B, O). As a practical matter, some entries in the table might also be missing.

To understand numerical data, a researcher might make a scatter plot; cluster the examples or the features; predict some of the values in the table based on others; remove (or simply identify) noisy or spurious values; or impute the values of missing entries. Many methods are available for any one of these specific tasks. By fitting a low rank model to the data, researchers can perform all of these computations simultaneously—even on large data sets containing heterogeneous values and many missing entries. Here, we describe what a low rank model is, give some examples of low rank models, and discuss how to pick a good low rank model for a particular application.

A low rank model approximates a table as the (matrix) product of two numerical matrices X and Y. Every example (e.g., patient) is represented by a row of X; every feature (e.g., lab test) is represented by a column of Y. The length of each of these rows and columns must be the same, and is called the rank of the model. A good low rank model compresses the information in the original data set using a rank that is much smaller than the number of rows or columns in the original table.

Principal Components Analysis (PCA), introduced by Karl Pearson in 1901, is a simple example of a low rank model. It finds a low rank model that minimizes the squared difference between the entries in the low rank model XY and those in the original data table.

PCA works well when the table consists only of numerical data with small, normal errors and has no missing entries. But often data does not fit these assumptions. In our lab test example, tests that have not been performed or survey questions left blank leave us with missing entries; malfunctioning sensors produce large, infrequent errors rather than small, normal errors. Moreover, PCA often returns a model that is difficult to interpret, and cannot be made to produce a model that captures our knowledge about the data, for example, it being nonnegative or sparse.

A number of methods have successfully extended PCA, each addressing one of these issues. These variations include nonnegative matrix factorization (which produces nonnegative factors), matrix completion (which handles missing data), robust PCA (which is less sensitive to noisy data), and sparse PCA (which produces factors with many zero entries).

A unified framework, which we call generalized low rank models, brings together the capabilities of these different techniques. It is able to simultaneously handle heterogeneous values, missing data, and prior beliefs about the factors. Even the well-known k-means clustering algorithm can be interpreted as a special case of a generalized low rank model. This framework makes it easier to use low rank models in everyday data analysis workflows. □

DETAILS

Madeleine Udell is a PhD candidate at Stanford University's Institute of Computational & Mathematical Engineering. She works with Stephen Boyd, PhD, professor of electrical engineering, with a focus on on convex optimization applications. The Boyd lab has developed and released a number of software packages for modeling and fitting generalized low rank models, available in different languages:

- **Julia (https://github.com/madeleineudell/LowRankModels.jl);**
- **Python (https://github.com/cehorn/GLRM); and**
- **Spark (http://git.io/glrmspark).**

The Julia and Spark packages are able to scale to datasets with billions of entries.

To learn how to use low rank models to produce scatter plots, cluster data, predict missing entries, and identify noisy or corrupted data, visit http://www.bcr.org/content/using-low-rank-models.

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 S271
Stanford, CA 94305-5444

seeing science

SeeingScience

BY KATHARINE MILLER

Joining the Atomic Ballet

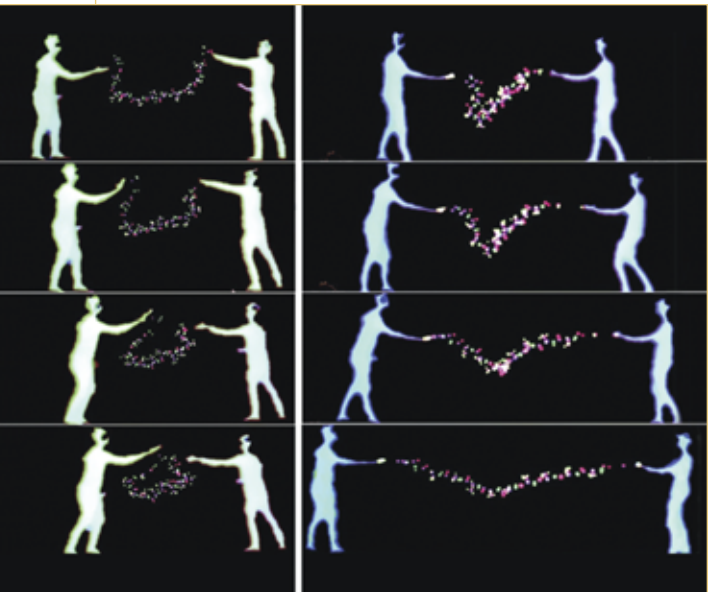
Colored atoms bounce off one another on a vast wall at the Stanford Art Gallery on the Stanford University campus. But when visitors approach, their energy avatars appear on the wall and the atoms react: Sparkly blue ones accumulate inside the avatars while others bounce off them, scattering and gathering in response to their tilting, dancing bodies.

This is Danceroom Spectroscopy (dS), an immersive, interactive simulation that **David Glowacki, PhD**, a Stanford visiting scholar and Royal Society research fellow, produced in collaboration with more than 40 scientists and artists.

Xbox 360 Kinect cameras capture gallery visitors' 3-D images and convert them into an energy landscape embedded in a quantum simulation of an atomic liquid. The system rotates through hundreds of different atomic simulation setups, each with different physical properties and visual effects. And the soundscape changes too, in response to participant-generated waves and ripples in the atomic bath. It's all running as close to real-time as it can (just a 17 millisecond delay), harnessing the power of more than 5000 GPU cores on a computer that, Glowacki says, "is pushing the limits of what interactive computing can do."

Taking advantage of that power, Glowacki recently used dS for more scientific purposes: a molecular dynamics (MD) simulation of the 10-alanine peptide embedded in 10,000 water molecules. The work was reported during a *Faraday Discussions* meeting in March 2014. In ordinary MD simulations, Glowacki says, "it can take a long time to simulate the rare events you really care about." Using hand manipulations of 10-alanine, experts and novices were able to accelerate the rare events by three to four orders of magnitude.

Someday soon, Glowacki hopes to investigate whether dS has any potential as a crowd-sourced platform for mapping conformations in proteins and other molecules—enabling the video game generation to help advance biomedical science. □



Using Danceroom Spectroscopy (dS), human avatars push or pull a simulated 10-alanine peptide molecule into various conformations. Above right: A dancer performs with dS. Photos by Paul Blakemore courtesy of Danceroom Spectroscopy.