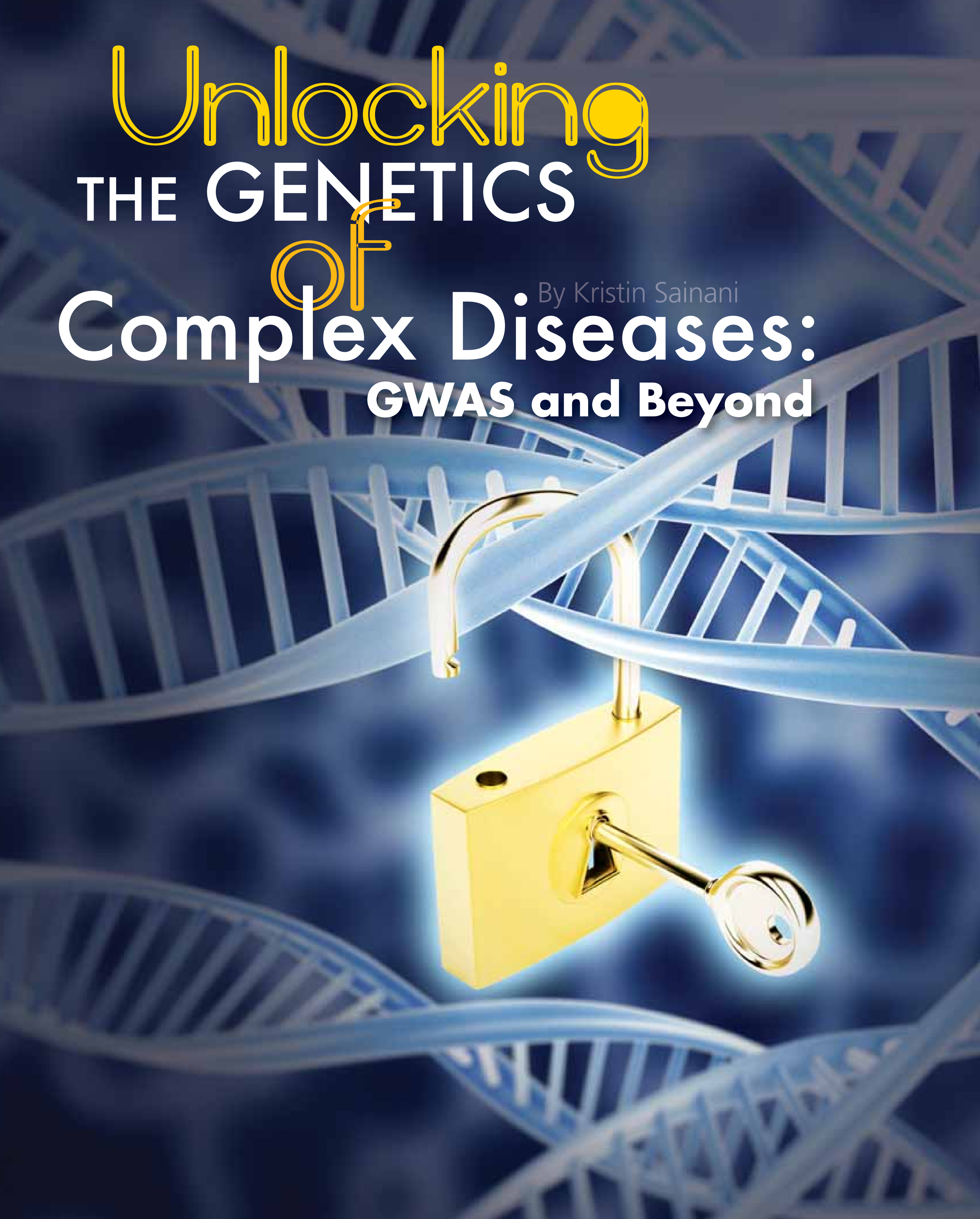
Unlocking THE GENETICS of Complex Diseases: GWAS and Beyond

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.

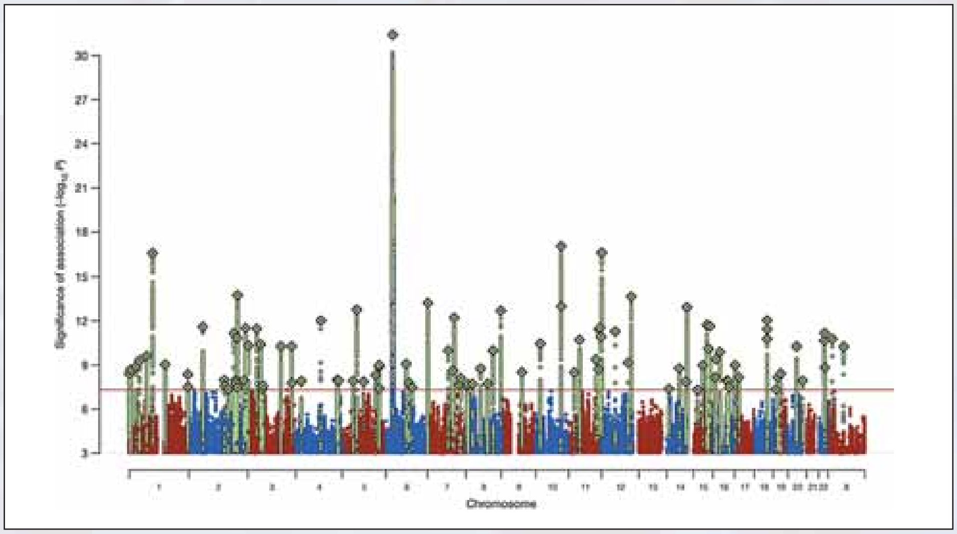
Given the messy genetic reality of complex diseases, GWAS have delivered exactly what they are capable of delivering: not answers, but an enormous number of clues. To date, GWAS have reliably linked thousands of genetic variants to hundreds of complex diseases or traits. "Many variants have now been 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.

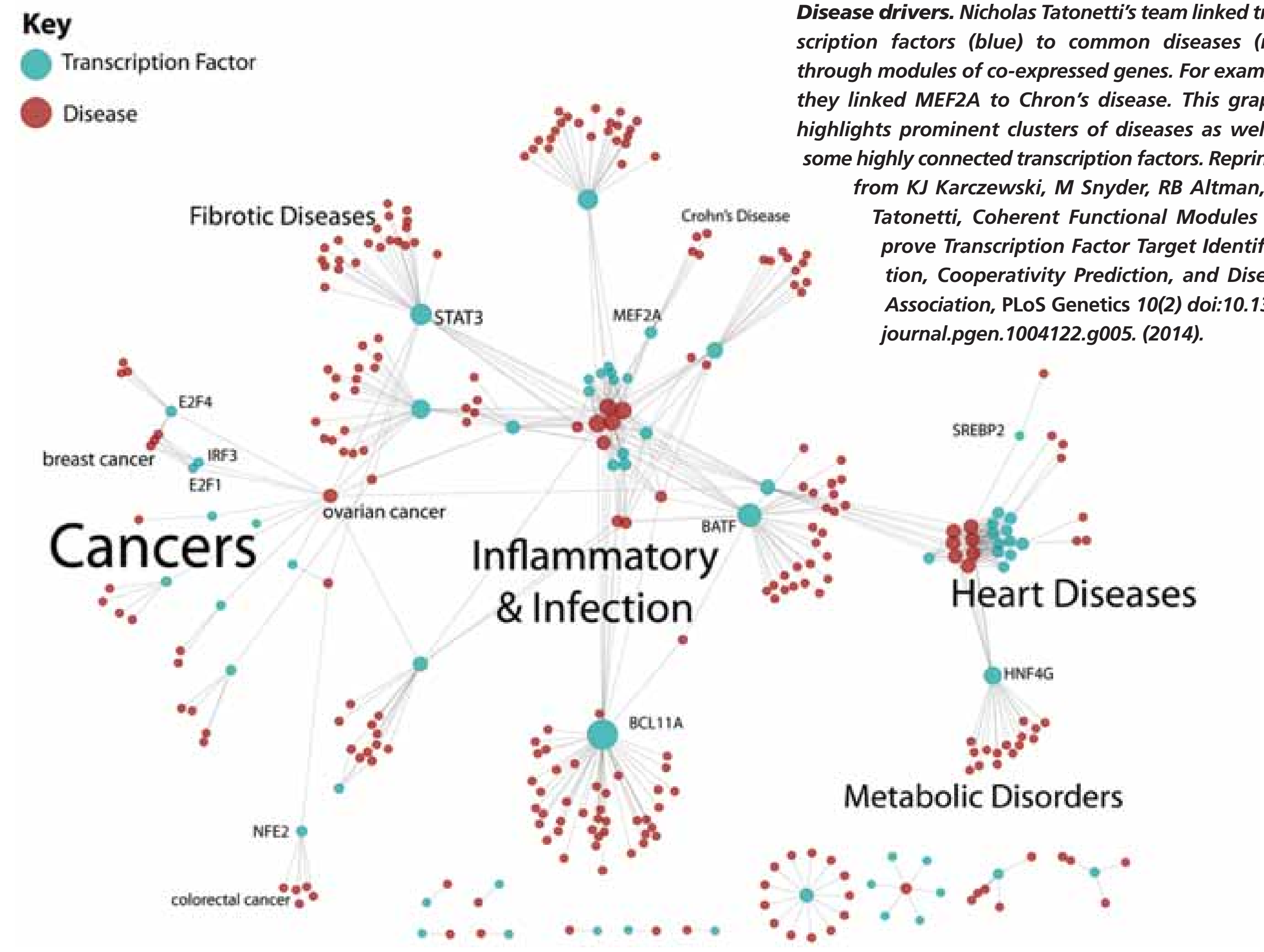
To link GWAS hits into pathways, some researchers are hunting down the transcription factors (TFs) that initiate the expression of clusters of genes in concert and thus may play a role in complex disease. Data from high-throughput experiments called "ChIP-Seq" can yield valuable information about where 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).*

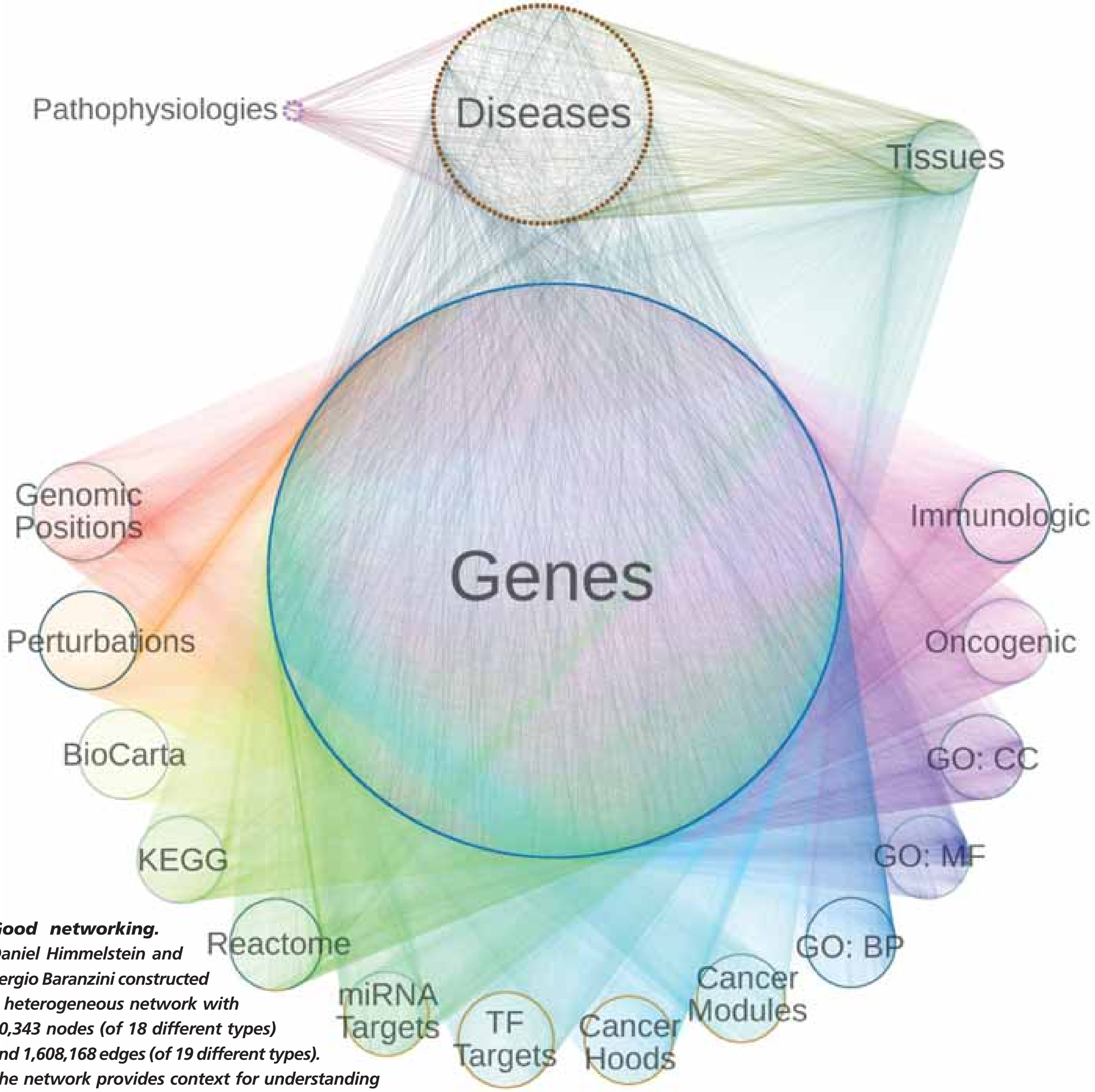
tems approach even further by incorporating non-molecular data, including environmental and clinical data," Tatonetti says.

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-authoriships 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.

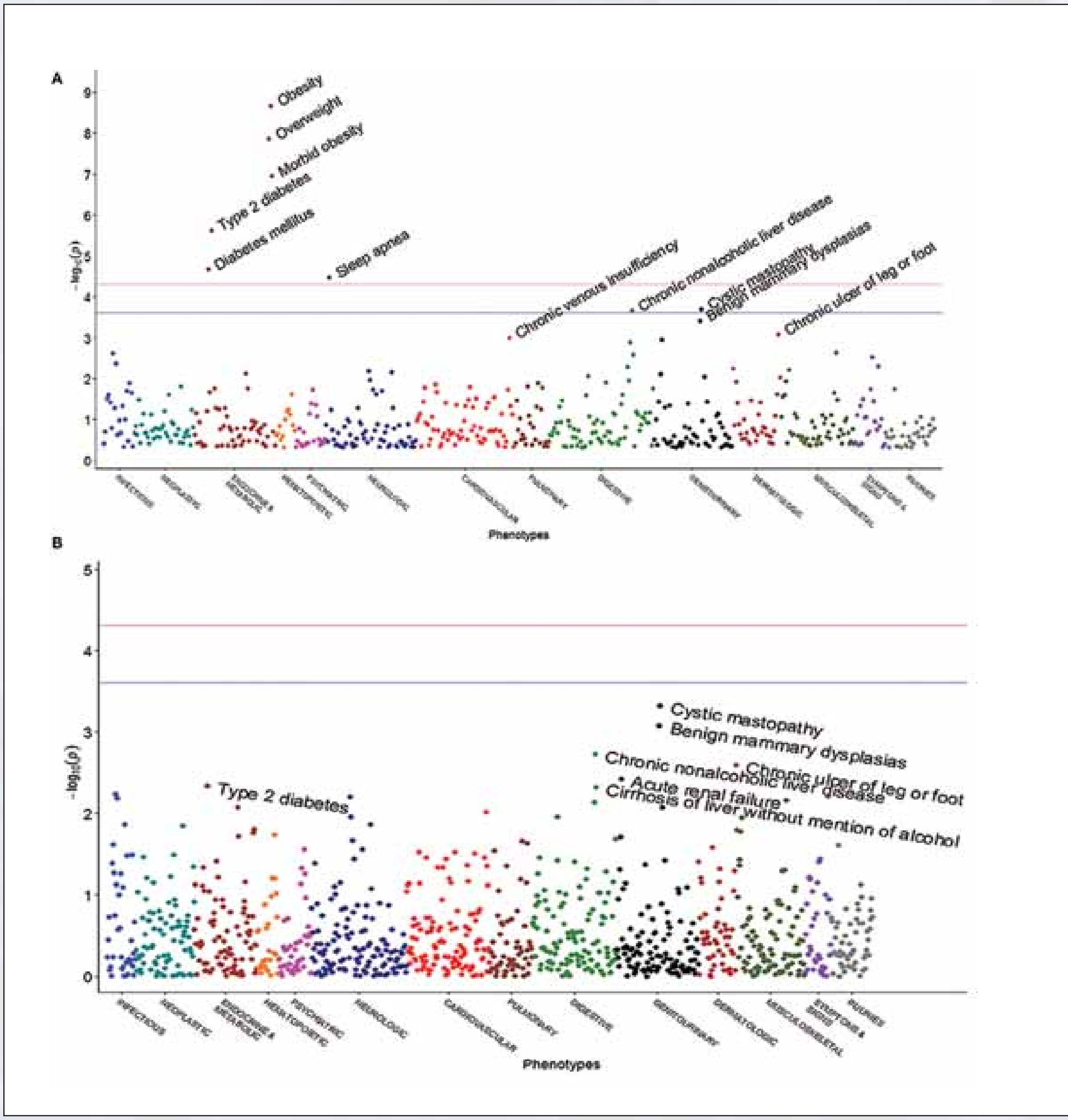
PheWAS: REVERSING GWAS

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.” □