

COMPUTING GENE INTERACTIONS: Functional and Statistical Approaches Converge

By Chandra Shekhar, PhD

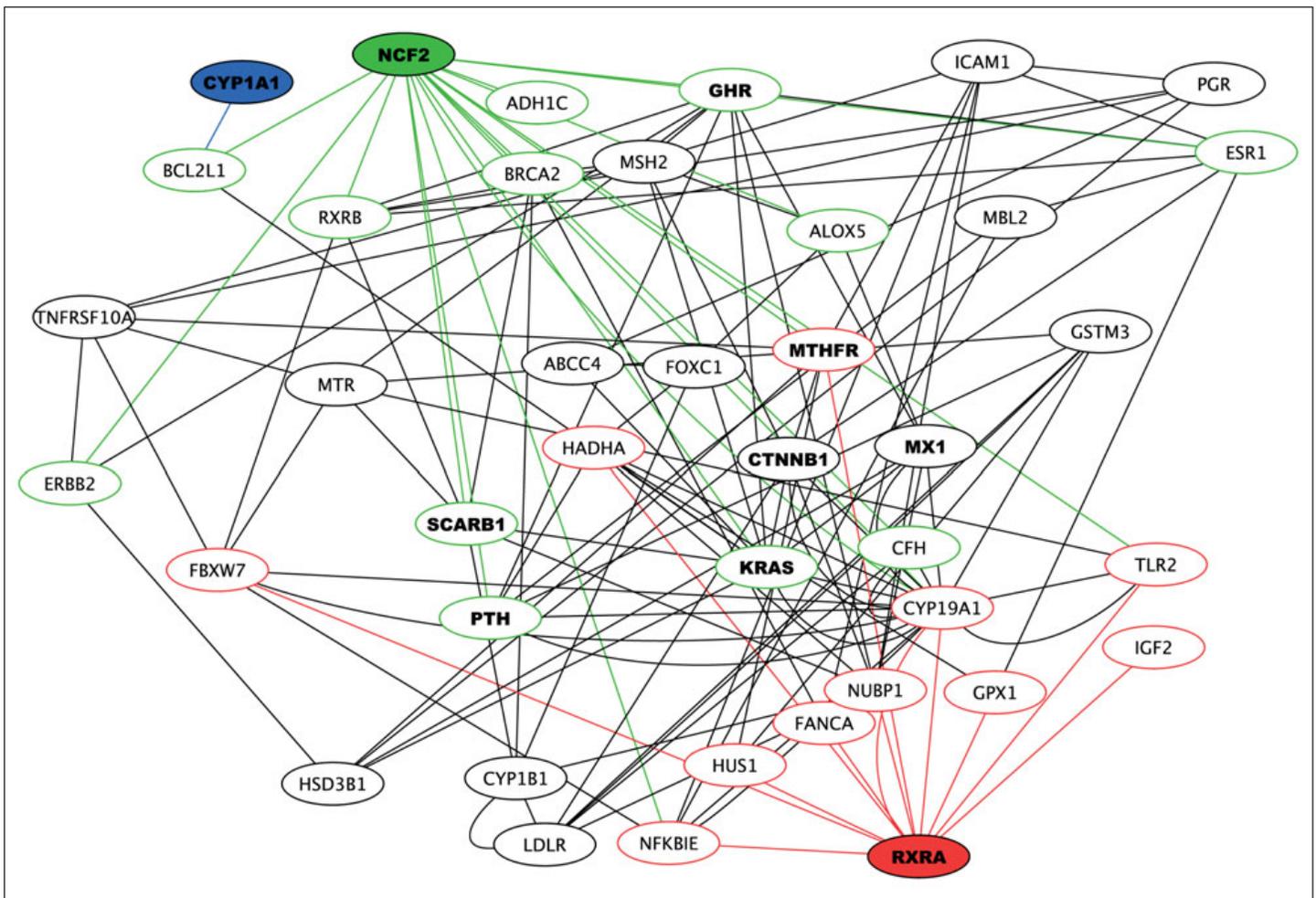
When people work together, some individuals may hinder team performance—essentially masking the abilities of other members—while others may boost the group’s performance beyond the sum of individual efforts. Genes interact in a similar way. This phenomenon, termed epistasis, could underlie the problem of “missing heritability”—the fact that individual genetic variation accounts for only a tiny fraction of the risk of complex diseases such as type 2 diabetes, high blood pressure, and multiple sclerosis.

Studying epistasis poses formidable conceptual and computational challenges, but recent advances promise to help unlock its potential to explain the inheritance of complex disease.

The conceptual challenge springs from two different fundamental interpretations of the term, one functional and the other statistical. In 1909, zoologist William Bateson defined epistasis as the phenotypic masking of one gene by another. For instance, a gene that causes albinism masks, or is epistatic to, genes that deter-

mine skin color. This definition evolved into approaches to understanding how genes interact functionally. “However, just because genes interact structurally or functionally doesn’t mean those interactions will be important for explaining disease,” says Patrick Phillips, PhD, a geneticist at the University of Oregon in Eugene who is developing a unified approach to epistasis.

In 1918, statistician Ronald Fisher redefined epistasis as what happens when a quantitative phenotype can’t be explained



Network of interactions associated with immune response to smallpox vaccination from a genetic study of 136 volunteers that assayed 1536 SNPs. McKinney’s group used a nonparametric method called Evaporative Cooling to identify the SNPs most relevant to smallpox vaccine response. The most important SNP turned out to be one in retinoid X receptor alpha (RXRA), a nuclear receptor gene that mediates vitamin A and D signaling. This gene’s strong genetic effect stems from its connectivity; it would not stand out in analyses that focus on individual gene effects. Courtesy of Brett McKinney. Reprinted with permission from N.A. Davis, J.E. Crowe, Jr., N.M. Pajewski, and B.A. McKinney. Surfing a genetic association interaction network to identify modulators of antibody response to smallpox vaccine. *Genes and Immunity*, 2010. Link: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3001955/>

by adding the effects of different gene variants. This view has evolved into a statistical concept of genetic disease risk. But this too can lead to results that do not fully explain disease. “Sometimes you can’t tell if the statistical interactions you see are functionally important,” Phillips says.

Despite this conceptual challenge, both approaches to characterizing epistasis have made significant progress in recent years. Statistical geneticists initially worked mainly with linear models and other parametric methods. When applied to genetic studies, however, these classic methods typically require too many parameters to be estimated from relatively sparse data. (For instance, a complete model to predict the effect of a pair of genes, each with two possible variants, has nine parameters: the overall mean effect, additive and dominance effects at each genetic location, or locus, and four parameters for the interaction effects.) Further, the methods prioritize main effects over interactions. “Modeling interactions is an afterthought in the traditional statistical paradigm,” says **Jason Moore, PhD**, of the Dartmouth Medical School in Lebanon, N.H.

“The problem then is that you completely miss them in the absence of main effects.”

To address these issues, Moore and his team have developed Multifactor Dimensionality Reduction (MDR), now one of the most widely used algorithms for epistasis. Instead of treating each genetic locus as a separate feature, MDR constructs a new attribute by pooling together genotypes from multiple loci. A multilocus genotype is labeled high-risk for a disease phenotype if its ratio of cases to controls exceeds a threshold; otherwise it’s labeled low-risk. This turns the high-dimensional problem of assessing multilocus interactions into a more tractable one of classifying a single attribute as low or high risk. Further, the new attribute automatically embodies interactions. Using this approach in a recently published bladder cancer study, Moore and colleagues identified a pair of interacting DNA repair enzyme genes whose combination made a better predictor of the disease than smoking. “If we had looked at the polymorphisms one at a time, we would have missed this strong genetic effect,” he says.

Some researchers have combined non-parametric methods with different relative strengths to make a stronger algorithm to detect epistasis. Random forests (RF) is a

powerful classification tree approach to finding patterns in data, but, as with classical parametric methods, tends to be more responsive to main effects than interactions when used in genetic studies. In contrast, Recursive Elimination of Features-F (Relief-F) is an algorithm that detects interactions even when main effects are absent, but struggles to handle noisy variables. **Brett McKinney, PhD**, of the University of Tulsa’s Institute for Bioinformatics and Computational Biology in Oklahoma has developed an approach called Evaporative Cooling that balances interactions (from Relief-F) and

“Although we are still a ways from bridging the gap between the functional and statistical approaches, it’s exciting to see these two ways of looking at the problem coalescing into one,” Phillips says.

main effects (from RF) in a statistical thermodynamics framework. When the system is optimized, the emerging network of interactions and associations captures the underlying genetic architecture. “The worst genes evaporate,” says McKinney. “The ones left are the ‘cooler’ genes more relevant to the phenotype.”

McKinney applied this filtering method to data from a recent genetic study of the strength of immune response to vaccination for smallpox, a disease that continues to pose a bioterror threat despite being eradicated. He then used a criterion called eigenvector centrality to find important SNP nodes associated with smallpox vaccine antibody titer. The central hub turned out to be a nuclear receptor gene involved in the metabolism of vitamins A and D. “The gene’s main effect was unremarkable—and, for that matter, its interactions were also quite small individually—but it had a major influence on the immune response phenotype because it interacted with so many other genes,” says McKinney. And, indeed, *in vitro* experiments by other groups have shown that boosting the effect of this gene (by adding small concentrations of its key product) increases antibody production in the immune system’s B cells.

When working on genome-wide data, even the cleverest algorithm will struggle to separate real epistasis from the millions of possible interactions. This is where prior biological information comes in handy, says **Marylyn Ritchie, PhD**, of Pennsylvania State University, one of Moore’s collaborators on the MDR algorithm. Ritchie and her team have developed Biofilter, a system that uses knowledge from public repositories of pathways and gene groupings such as the Genetic Association Database to produce multi-SNP models for genetic studies. “We reduce the search space by focusing on the interactions that make biological sense,” says Ritchie. Using this approach in a 2011 multiple sclerosis (MS) study, her team was the first to find MS-associated SNPs in a neurological pathway. “In contrast, all past genetic associations of MS have been autoimmune,” she notes.

Some important genetic interactions, however, have no obvious biological basis. And to find them, some researchers have used a functional approach. In a mammoth study of genetic interaction in the yeast *S. cerevisiae* published last year, a team led by University of Toronto researchers examined 5.4 million colonies of the organism, each with a different pair of genes knocked out. A gene pair was said to interact if the corresponding double mutant was fitter or less fit than would be expected from independently acting genes. This criterion identified about 170,000 interactions, many between genes in completely different pathways. “We were able to capture far-reaching relationships that often didn’t represent physical connections,” says **Chad Myers, PhD**, of the University of Minnesota in Minneapolis. This type of functional interaction network could be very useful in drug design, says Myers; for instance, a potential drug could reduce unforeseen effects by avoiding targeting hubs in the network.

With statistical geneticists like Ritchie employing biological criteria and functional geneticists like Myers adopting quantitative methods, the field of epistasis is rapidly evolving. “Understanding epistasis is key to inferring the genetic basis of complex traits,” says Phillips. “Although we are still a ways from bridging the gap between the functional and statistical approaches, it’s exciting to see these two ways of looking at the problem coalescing into one.” □