European Journal of Human Genetics (2009) 17, 1205–1206

www.nature.com/ejhg

NEWS AND COMMENTARY

Back to the Genome

From genotypes to genometypes: putting the genome back in genome-wide association studies

JH Moore

European Journal of Human Genetics (2009) **17,** 1205–1206; doi:10.1038/ejhg.2009.39; published online 11 March 2009

Human genetics has a long history of benefiting from technological advances that have made it possible to measure genomic variation. Research over the last 5 years has focused on genome-wide association studies (GWAS), which, for the first time, allow us to measure most of the relevant single-nucleotide polymorphisms (SNPs).[1,2] Research over the next 5 years will likely focus on measuring the entire genomic sequence in multiple subjects that can be used in application areas like the human microbiome project.[3] Although these technology-driven approaches are thought of as 'genomic' because they measure information from across the genome, they are still primarily approached analytically one SNP at a time. That is, the relationship between interindividual variation in the genome and variation in a given biomedical trait is assessed for each SNP independently of all the other measured SNPs and available measurements of human ecology. There are several reasons for this. First, parametric statistical approaches that form the foundation of statistical genetics and epidemiology are based on the generalized linear model that has much higher power to detect independent main effects than complex interactions among multiple risk factors. As a result, there is a statistical culture of ignoring interactions because interactions are often not detected using methods such as linear regression. Second, there are a number of practical barriers to routine analysis of multiple genetic and environmental risk factors. Powerful machine learning methods and fast parallel computers are needed to detect nonlinear interactions in high-dimensional genome-wide datasets. As a result, the special expertize in computer science, software engineering and computer hardware that are needed to implement these methods are often out of reach for the typical geneticist or epidemiologist. Finally, successful detection of nonlinear interactions still requires experimental validation and biological inference, which is much easier if only a single risk factor is considered. The high-throughput experimental methods for perturbing multiple genetic and environmental factors in a model organism or cell line are not yet available.

Now that many of the technical and quality control issues for GWAS have been addressed, it is time to return to thinking about the complex mapping relationship between genotype and phenotype. We desperately need biostatistical and bioinformatics methods that confront and embrace the full complexity of human health and disease. There are signs that the tide is turning. For example, the scientific content of the 2008 meeting of the International Genetic Epidemiology Society had a major emphasis on the use of knowledge about biochemical pathways and gene networks as an integrated part of genetic association analysis including GWAS. This is a recognition that the agnostic statistical paradigm that specifically ignores this type of information will only be useful for uncovering part of the genetic architecture of any given complex trait. In addition to statistical genetics and genetic epidemiology, there is a major paradigm shift happening in bioinformatics. It was evident from the recent 2009 Pacific Symposium on Biocomputing that much more emphasis is being placed on developing algorithms and software for the analysis of systems and networks rather than single biological molecules. As such, the paper by Emily *et al*[4] showing how biological networks can be used to guide a GWAS analysis is particularly timely.

It is generally recognized that epistasis or gene–gene interaction plays an important role in the genetic architecture of human health.[5] Detecting and characterizing gene–gene interactions in GWAS is computationally challenging because of the extreme combinatorial nature of the problem.[6] In fact, there are not enough computers in the world to exhaustively enumerate all the three-way, four-way and five-way combinations of SNPs in a GWAS. As such, we need creative alternatives to the brute-force combinatorial approach. One idea is to use our knowledge of protein–protein interactions to help guide a GWAS analysis of epistasis.[7] The idea is that two or more genes with protein products that physically interact are more likely to exhibit a statistical interaction that can be detected in a human population. The paper by Emily *et al*[4] in this issue specifically tests this hypothesis using GWAS data from the Wellcome Trust Case Control Consortium for several different common human diseases. This paper shows how protein–protein interactions from the STRING database can be used to prioritize SNPs for interaction analysis, thus significantly reducing the total number of SNP pairs that need to be evaluated. This effectively reduces computational analysis time and the total number of tests that need to be performed, thus reducing the potential number of false-positives. Under the assumption that genes with protein–protein interactions are more likely to exhibit statistical interactions, this approach is expected to be more powerful than the brute-force approach of exploring all possible combinations.

The paper by Emily *et al*[4] is an example of how the knowledge of pathways and networks can be used to enhance GWAS analysis. There are several other recent

examples as well that support the idea that this is a growing trend. Bush *et al*[8] propose a Biofilter approach that uses knowledge from public databases such as STRING to reduce the number of SNPs that need to be evaluated for interactions. In a slightly different approach to the same problem, Askland *et al*[9] showed that biological pathways with ensembles of significant SNPs from GWAS are more likely to replicate across studies than individual SNPs. These studies support the idea that our knowledge of biology will play a very important role in our ability to embrace the complexity of the genetic architecture of human health and disease. For the genome-wide analysis or epistasis to become a reality, we need to develop the statistical and computational methods that can fully exploit the growing body of expert knowledge. For example, Greene *et al*[10] have proposed using stochastic search algorithms that are guided by earlier statistical knowledge or by biological knowledge such as protein–protein interactions. The methods presented by Emily *et al*[4] and others show great promise for moving us beyond chip-based technology, for example, towards the scientific focus on and motivation to embracing and studying the complexity of human biology. This represents an early step in the progression from considering single SNPs as risk factors to considering multiple interacting SNPs as risk factors to considering the entire genome as a risk factor. The latter end of this complexity spectrum suggests that our individual 'genometype' may ultimately prove the most useful for personalized medicine and personal genetics. If this is the case, it will necessarily alter our general approach to human genetics■

Dr JH Moore is at the 706 Rubin Bldg, HB7937, Dartmouth Medical School, One Medical Center Dr., Lebanon, NH 03756, USA.
Tel: +603 653 9939;
Fax: +603 653 9900;
E-mail: jason.h.moore@dartmouth.edu
Web: www.epistasis.org

References

1 Hirschhorn JN, Daly MJ: Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005; **6**: 95–108.
2 Wang WY, Barratt BJ, Clayton DG, Todd JA: Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005; **6**: 109–118.
3 Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggitt CM, Knight R, Gordon JI: The human microbiome project. *Nature* 2007; **449**: 804–810.
4 Emily M, Mailund T, Hein J *et al*: Using biological networks to search for interacting loci in genomewide association studies. *Eur J Hum Genet* 2009; **17**: 1231–1240.
5 Moore JH: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 2003; **56**: 73–82.
6 Moore JH, Ritchie MD: The challenges of whole-genome approaches to common diseases. *JAMA* 2004; **291**: 1642–1643.
7 Pattin KA, Moore JH: Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. *Hum Genet* 2008; **124**: 19–29.
8 Bush WS, Dudek SM, Ritchie MD: Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput* 2009; 368–379.
9 Askland K, Read C, Moore J: Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Hum Genet* 2009; **125**: 63–79.
10 Greene CS, White BC, Moore JH: Ant colony optimization for genome-wide genetic analysis. *Lect Notes Comput Sci* 2008; **5217**: 37–47.