Genomewide Association Studies — Illuminating Biologic Pathways

Joel N. Hirschhorn, M.D., Ph.D.

Human geneticists seek to understand the inherited basis of human biology and disease, aiming either to gain insights that could eventually improve treatment or to produce useful diagnostic or predictive tests. As recently as 2004, few genetic variants were known to reproducibly influence common polygenic diseases (including cancer, coronary artery disease, and diabetes) or quantitative phenotypes (including lipid levels and blood pressure). This relative ignorance limited potential insights into the pathophysiology of common diseases.

The completion of the human genome sequence in 2005 and the provision of an initial catalogue of human genetic variation and a haplotype map (known as the HapMap), together with rapid improvements in genotyping technology and analysis, have permitted genomewide association studies to be undertaken in a large number of samples.[1] In the first and current implementation of this approach, the great majority of genetic variants with population frequencies of 5% or more could be tested directly or indirectly for association with disease risk or quantitative traits — thus providing a potential path to gene discovery for polygenic diseases and traits.

Before the initiation of genomewide association studies, there was considerable and healthy skepticism about their likely success. For example, in 2005, two friends and well-known geneticists, Francis Collins and Thomas Gelehrter, made a public bet: Gelehrter predicted that no more than three new common variants would be reproducibly associated with common diseases by the time the American Society of Human Genetics (ASHG) held its meeting in the autumn of 2008.

During the past 2 years, however, genomewide association studies have identified more than 250 genetic loci in which common genetic variants occur that are reproducibly associated with polygenic traits.[1-4] This explosion represents one of the most prolific periods of discovery in human genetics, with most new loci identified in genomewide association studies published during the past 18 months. The bet was settled: Collins was the clear winner, by a margin of more than 200 new associated variants.

New skeptics have now questioned the value of these recent discoveries. They cite the modest effect sizes of common variants, both individually and in combination, and argue that the small fraction of heritability that is explained by these variants precludes practical prediction or meaningful biologic insights. A second argument is articulated by Goldstein in his Perspective article in this issue of the *Journal* (pages 1696–1698); he predicts that genomewide association studies will not yield too few loci but rather too many. Extrapolating from recent discoveries, he builds a speculative mathematical model and infers that there will be tens of thousands of common variants influencing each disease and trait. Assuming that these variants will be evenly distributed across the genome, he concludes that every gene in the genome could theoretically be implicated, a scenario that would prohibit useful biologic insights.

I believe that the skeptics' arguments either misconstrue the primary goal of genomewide association studies or are contradicted by their findings. The main goal of these studies is not prediction of individual risk but rather discovery of biologic pathways underlying polygenic diseases and traits. It is already clear that the genes being identified expose relevant biology. Genomewide association studies have "rediscovered" many genes that have been shown by decades of work to be important. Of the 23 loci found to be associated with lipid levels, 11 implicate genes encoding apolipoproteins, lipases, and other key proteins in lipid metabolism.[2] Studies of other diseases and traits have highlighted equally relevant genes.[1,3,4] Nearly one fifth of the approximately 90 loci that were found to be associated with type 2 diabetes, lipid levels, obesity, or height include a gene that is mutated in a corresponding single-gene disorder.[2,4] The number of such overlaps is overwhelmingly greater than what would be expected by chance. Furthermore, genomewide association studies have highlighted genes encoding the sites of action of drugs approved by the Food and Drug Administration, including thiazolidinediones and sulfonylureas (in studies of type 2 diabetes),[2] statins (lipid levels),[2]

and estrogens (bone density).[5] Each of the associated variants at a drug-target locus explains less than 1% of phenotypic variation in the population, demonstrating that small effect sizes do not preclude biologic importance.

Critically, genomewide association studies have also highlighted pathways whose relevance to a particular disease or trait was previously unsuspected. The genetic variants that are associated with age-related macular degeneration strongly implicate components of the complement system, the loci associated with Crohn's disease[3] point unambiguously to autophagy and interleukin-23–related pathways, and the height loci[4] include genes encoding chromatin proteins and hedgehog signaling. This clustering into biologic pathways is highly nonrandom (as has been demonstrated by Raychaudhuri and Daly). Already, efforts are under way to translate the new recognition of the role of autophagy in Crohn's disease into new therapeutic leads. As more pathways are highlighted and additional hypotheses emerge, new projects can be born.

Finally, many newly identified loci do not implicate genes with known functions. It is hardly surprising that we do not yet understand the biologic import of every recently associated locus: the associations sometimes do not point unambiguously to a particular gene, and even genes that are clearly implicated are often unannotated with respect to function. For these genes, greater effort will be required before we can generate hypotheses for future work, but by charting new paths, such efforts could eventually lead to the most novel and important insights.

With regard to prediction, the common variants described by genomewide association studies almost universally have modest predictive power, and for most diseases and traits, these variants in combination explain only a small fraction of heritability. However, the success of genomewide association studies is not tied to prediction. If we identify only new pathways underlying disease, these studies will have a tremendous impact.

Nevertheless, it remains likely that for some diseases, the loci that are highlighted in the studies will provide useful predictive information. For several diseases, associated variants already explain 10 to 20% or more of heritability, a magnitude that is similar to the proportion of risk explained by nongenetic tests in widespread clinical use (such as levels of low-density lipoprotein cholesterol or prostate-specific antigen). Furthermore, current estimates are a lower bound for the eventual predictive power of recently discovered loci, which have not been thoroughly examined for additional common and rare variation. Indeed, early experience suggests that multiple independent causal variants may be found at each locus, accounting for additional increments of heritability.[1] Genomewide association studies that are performed in larger samples and that use genotyping platforms designed to test variants with a prevalence of less than 5% will increase the variation explained at these and other as-yet-undiscovered loci, as will studies taking into account interactions among genes and between genes and the environment. Ultimately, the usefulness of genetic information for prediction will depend not on the absolute fraction of heritability explained but rather on how much this additional information can shift the cost–benefit ratios of available clinical interventions. For diseases without potential therapies, even perfect prediction might not be clinically useful. By contrast, for diseases with effective preventive measures that are too costly or for which the risk–benefit balance is nearly neutral, small increments in predictive power could help effectively target preventive efforts, with substantial clinical impact.

The biologic pictures being revealed by genomewide association studies are still quite incomplete. We should strive for as complete a catalogue of validated risk variants as possible, through additional genomewide association studies and complementary approaches (such as exon-based or genomewide sequencing in sufficiently large samples) as they become available.

New biologic insights do not guarantee a rapid translation into clinical practice; the latter will require great effort by basic, translational, and clinical researchers. The difficulty in translation is not unique to genetic discoveries: nearly a century and three Nobel Prizes separate the determination of the chemical composition of cholesterol from the development of statins. Each discovery of a biologically relevant locus is a potential first step in a translational journey, and some journeys will be shorter than others. With a more complete collection of relevant genes and pathways, we can hope to shorten the interval between biologic knowledge and improved patient care.

In response to the skeptics, I offer a new bet. I predict that by the 2012 ASHG meeting, genomewide association studies will have

yielded important new biologic insights for at least four common diseases or polygenic traits — and that efforts to develop new and improved treatments and preventive measures on the basis of these insights will be well under way.

Dr. Hirschhorn reports receiving consulting fees from Correlagen and Ipsen, having an equity interest in Correlagen, receiving lecture fees from Pfizer, and receiving grant support from Novartis. No other potential conflict of interest relevant to this article was reported.

This article (10.1056/NEJMp0808934) was published at NEJM.org on April 15, 2009.

Dr. Hirschhorn is an associate professor in the Program in Genomics and the Divisions of Genetics and Endocrinology, Children's Hospital, Boston; an associate professor of genetics at Harvard Medical School, Boston; and an associate member and coordinator of the Metabolism Initiative at the Broad Institute of Harvard and MIT, Cambridge, MA.

1. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008;322:881-8.

2. Mohlke KL, Boehnke M, Abecasis GR. Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. Hum Mol Genet 2008;17:R102-R108.

3. Lettre G, Rioux JD. Autoimmune diseases: insights from genome-wide association studies. Hum Mol Genet 2008;17:R116-R121.

4. Hirschhorn JN, Lettre G. Progress in genome-wide association studies of human height. Horm Res (in press).

5. Styrkarsdottir U, Halldorsson BV, Gretarsdottir S, et al. Multiple genetic loci for bone mineral density and fractures. N Engl J Med 2008;358:2355-65.

Genetic Risk Prediction — Are We There Yet?

Peter Kraft, Ph.D., and David J. Hunter, M.B., B.S., Sc.D., M.P.H.

A major goal of the Human Genome Project was to facilitate the identification of inherited genetic variants that increase or decrease the risk of complex diseases. The completion of the International HapMap Project and the development of new methods for genotyping individual DNA samples at 500,000 or more loci have led to a wave of discoveries through genomewide association studies. These analyses have identified common genetic variants that are associated with the risk of more than 40 diseases and human phenotypes. Several companies have begun offering direct-to-consumer testing that uses the same single-nucleotide polymorphism chips that are used in genomewide association studies. These companies claim that such testing should be made available to consumers who are interested in their personal level of risk for the relevant diseases. Now, "risk tests" for specific diseases such as breast cancer are also being marketed to physicians and consumers.[1]

The availability of highly predictive and reasonably affordable tests of genetic predisposition to important diseases would have major clinical, social, and economic ramifications. But the great majority of the newly identified risk-marker alleles confer very small relative risks, ranging from 1.1 to 1.5,[2] even though such analyses meet stringent statistical criteria (i.e., the identification of associations with disease that have very small P values and hence are unlikely to be false positives). However, even when alleles that are associated with a modest increase in risk are combined, they generally have low discriminatory and predictive ability.[3]

One argument in favor of using the available genetic predictors is that some information must be better than no information, and we should not let the perfect be the enemy of the good by refusing to make use of our knowledge until it is more complete. Why not begin testing for common genetic variants whose associations with susceptibility to disease have been established?

The answer lies in the stability of the current risk estimates. Genetic variants conferring the highest relative risks are almost certainly overrepresented in the first wave of findings from genomewide association studies, since considerations of statistical power predict that they will be identified first. However, a striking fact about these first findings is that they collectively explain only a very small proportion of the underlying genetic contribution to most studied diseases. (Some exceptions exist — notably, age-related macular degeneration, for which a few alleles explain a substantial fraction of the genetic contribution.) Several lines of evidence support this overall conclusion.

First, the relative risks that are found to be conferred by common risk genotypes account for only a small proportion of the sibling recurrence risk (or the risk that a sibling will also have the disease of interest). Second, in multivariate analyses of large epidemiologic data sets in which a family history of a disease is a risk factor, the inclusion of data regarding which subjects carry the known associated variants only minimally reduces the risk asso-