Common Genetic Variation and Human Traits

David B. Goldstein, Ph.D.

The human genome has been cracked wide open in recent years and is spilling many of its secrets. More than 100 genome-wide association studies have been conducted for scores of human diseases, identifying hundreds of polymorphisms that are widely seen to influence disease risk. After many years in which the study of complex human traits was mired in false claims and methodologic inconsistencies, genomics has brought not only comprehensive representation of common variation but also welcome rigor in the interpretation of statistical evidence. Researchers now know how to properly account for most of the multiple hypothesis testing involved in mining the genome for associations, and most reported associations reflect real biologic causation. But do they matter?

Unfortunately, most common gene variants that are implicated by such studies are responsible for only a small fraction of the genetic variation that we know exists. This observation is particularly troubling because the studies are largely comprehensive in terms of common single-nucleotide polymorphisms (SNPs), the genomic markers that are genotyped and with which disease associations are tested. We're finding the biggest effects that exist for this class of genetic variant, and common variation is packing much less of a phenotypic punch than expected. Some experts emphasize that small effect sizes don't necessarily mean that a gene variant is of no interest or use. Effect size is a function of what a variant does: it may change only slightly a gene's expression or a protein's function. The gene's pathway, however, may be decisive for a particular condition, or pharmacologic action on the same protein may produce much larger effects in controlling disease. These arguments are reasonable, as far as they go, and there are supporting examples, such as a polymorphism of modest effect in *PPARG*, a gene that encodes a drug target for diabetes.

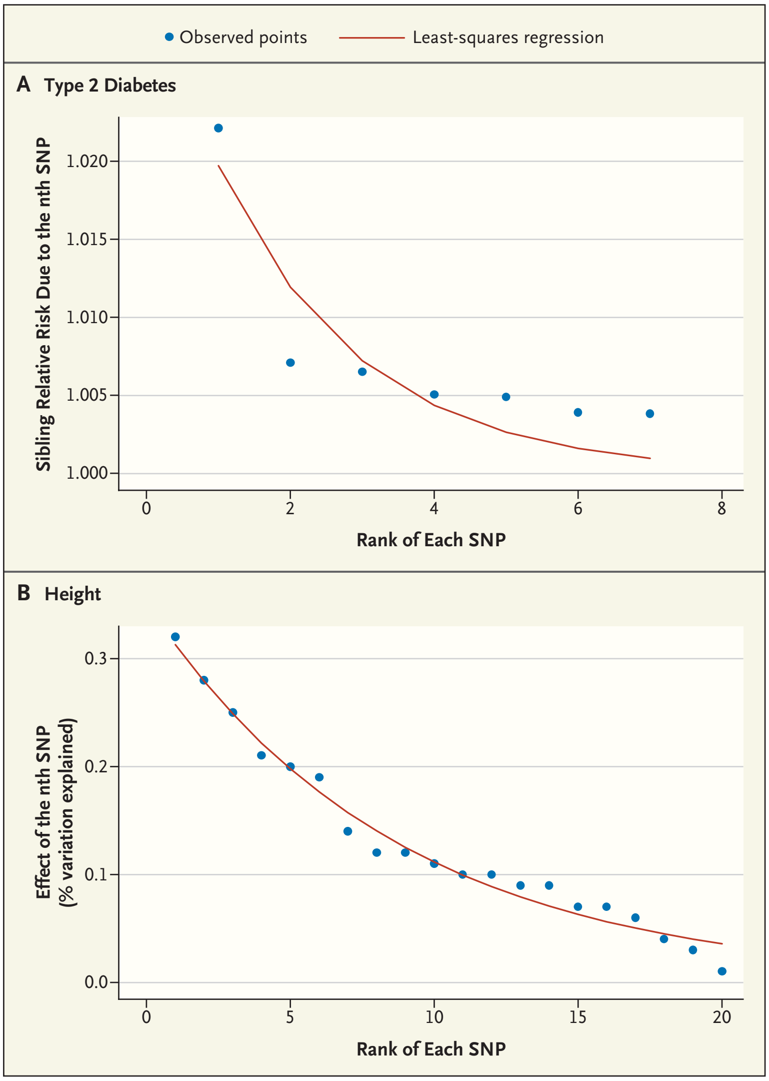
But the arguments hold only if common genetic variation implicates a manageable number of genes. If effect sizes were so small as to require a large chunk of the genome to explain the genetic component of a disorder, then no guidance would be provided: in pointing at everything, genetics would point at nothing. To assess whether effect sizes are too small in this sense, consider two examples of complex human traits — type 2 diabetes and height. In their recent review, Manolio et al.[1] described seven gene variants that influence the risk of type 2 diabetes. In addition to these variants, the one with the strongest effect on familial aggregation is in the *TCF7L2* gene.

One way to assess a variant's effect is by comparing the disease risk of the sibling of an affected person with that in the general population (sibling relative risk). The *TCF7L2* variant is associated with a sibling relative risk for type 2 diabetes of only about 1.02, whereas the overall risk of disease among siblings of affected persons is three times that in the general population. If the human genome carried scores of variants with such effects, they would collectively generate a substantial sibling relative risk. Unfortunately, we now know this is not the case: the contribution of common risk alleles to familial clustering falls off dramatically after *TCF7L2* and appears to become asymptotic at a level only marginally above 1 (see Panel A of the figure).[2] It seems likely, then, that an unreasonably large number of such variants would be required to account for the genetic component of diabetes risk, even if the sibling relative risk values overestimate the genetic component of disease.

A more quantitative evaluation is available for height, for which Weedon et al.[3] identified 20 polymorphisms. Using a replication sample set, they estimated that collectively, the variants they studied explain less than 3% of the population variation in height (see Panel B of the figure). To estimate the full distribution of effect sizes (including those of variants not yet discovered), one could assume an exponential distribution and estimate the parameters from the observed data. The predicted effect of the nth SNP is calculated as follows:

$$\text{Effect size of nth SNP} = k + a \times \text{Exp}[-bn],$$

in which $k = 0.0008$, $a = 0.35$, and $b = 0.1152$. To estimate the number of SNPs required to explain 80% of population variation in height (the most common estimate of height's heritability), this equation can be integrated and solved numerically. The answer is that approximately 93,000 SNPs are required to explain 80% of the population variation in height. In



Sibling Relative Risk for Each of 7 SNPs Associated with Type 2 Diabetes (Panel A) and Percentage of Variation Explained by Each of 20 SNPs Associated with Height (Panel B).

Panel A shows the contribution to a sibling relative risk of type 2 diabetes for each of seven SNPs, as estimated from data reported by Manolio et al.[1] with the use of formulas from Risch and Merikangas[2] and plotted against the rank order of the SNPs in terms of the magnitude of their contributions. Panel B shows the percentage of variation explained by each of 20 SNPs associated with height, as reported by Weedon et al.[3] For a quantitative trait, the natural measure of effect size is the proportion of variation in the trait that the SNP explains, which depends on both the allele frequency and the intergenotype differences. Effect sizes are shown as points as well as a fitted exponential function with the use of least-squares regression.

the fitted distribution, the constant term (0.0008) can be viewed as the predicted smallest effect size in the genome, given the 20 strongest effects already identified. The resulting integral can be considered valid only over the range of 1 to approximately 93,000, at which point all heritability would be explained.

I assume that all SNPs yet to be discovered have weaker effect sizes than the weakest so far found. Though the strongest SNP may have been found, many SNPs could remain unidentified in the range of the lower effects that have been determined. If such SNPs are accounted for, fewer SNPs will be required to explain a given proportion of variance. The sample sizes that have been studied for height, however, range from 14,000 to 34,000. At the lower sample size, the power of detection is 90% for the largest effect size; for effect sizes as small as 0.05%, the largest sample size provides a 10% chance of detection. Even if we conservatively assume that all remaining unidentified variants influencing height each explained as much as 0.05% of the variation, 1500 such variants would be required to explain the missing heritability. These calculations also assume that the effects of "height SNPs" are additive. If variants show meaningful interactions, a somewhat stronger genetic effect could emerge among variants with small individual effect sizes. But only dramatic departures from these assumptions would allow a manageable number of common SNPs to account for a sizable fraction of the heritability of height.

If common variants are responsible for most genetic components of type 2 diabetes, height, and similar traits, then genetics will provide relatively little guidance about the biology of these conditions, because most genes are "height genes" or "type 2 diabetes genes." It seems much more likely, however, that most genetic control is due to rarer variants, either single-site or structural, that are not represented in the current studies and that have considerably larger effects than common variants. Whether these "rarer" variants are only slightly below the threshold for detection on current platforms or substantially more rare remains to be seen. If, however, rarer variants are primarily responsible for the missing heritability, we may yet identify a manageable number of genes and pathways.

Either way, it's hard to have any enthusiasm for conducting genome scans with the use of ever larger cohorts after a study of the first several thousand subjects has identified the strongest determinants among common variants. These initial studies for a given common disease are worth doing, since common variants do appear to explain a sizable fraction of the heritability of certain conditions — notably, exfoliation glaucoma, macular degeneration, and Alzheimer's disease. Beyond studies of this size, however, we enter the flat or declining part of the effect-size distributions, where there are probably either no more common variants to discover or no more that are worth discovering.

By contrast, genome scans have not yet been performed in search of variants involved in many responses to drugs or infectious agents, even though there are examples in both categories of common polymorphisms whose effects dwarf those seen for type 2 diabetes and many other diseases. For example, when exposed to the anti-HIV drug abacavir, a hypersensitivity reaction develops in more than half the carriers of the HLA-B*5701 allele, whereas such a reaction occurs in less than 5% of patients without this allele.[4] Similarly, just three common variants are sufficient to explain 14% of the population variation in HIV-1 viral load.[5]

But with traits such as height or type 2 diabetes, it seems that an inordinate number of common SNPs would be needed to account for a sizable fraction of heritability. Indeed, it's possible that the way genome scans are being interpreted actually overestimates the contributions of common variants. Most variants that have been identified to date are markers, not causal variants, and are generally assumed to reflect the effects of some other, as-yet-unidentified common variant. Another possibility, however, is that some of the associations that are credited to common variants are actually synthetic associations involving multiple rare variants that occur, by chance, more frequently in association with one allele at a common SNP than with the other. In this case, as well, genome scans will overestimate the contribution of common variants.

The apparently modest effect of common variation on most human diseases and related traits probably reflects the efficiency of natural selection in prohibiting increases in disease-associated variants in the population. I believe attention should shift from genome scans of ever larger samples to studies of rarer variants of larger effect. Effectively searching the full human genome for rare variants will require not only sequencing capacity but also thoughtful selection of the most appropriate groups of individual genomes to resequence and thoughtful evaluation and prioritization of the many rare variants identified. There's no guarantee that associations with rare variants will point directly to causation. Nevertheless, the limited role of common variation in many highly heritable diseases argues strongly that there are many rare variants to be found, and it seems reasonable to hope that some of them will suggest novel therapeutic targets or help in the design of personalized prevention or treatment regimens.

These conclusions imply no criticism of the strikingly successful efforts to represent common variation and relate it to common diseases. Indeed, I share the view Hirschhorn presents in his Perspective article (pages 1699–1701) that the early skeptics have been proved wrong about genomewide association studies in most details: patterns of linkage disequilibrium are sufficiently consistent to allow efficient representation of common variation with the use of "tagging" SNPs, and secure associations between polymorphisms and diseases were rapidly and easily identified. But even though genomewide association studies have worked better and faster than expected, they have not explained as much of the genetic component of many diseases and conditions as was anticipated. We must therefore turn more sharply toward the study of rare variants.

No potential conflict of interest relevant to this article was reported.

This article (10.1056/NEJMp0806284) was published at NEJM.org on April 15, 2009.

Dr. Goldstein is director of the Center for Human Genome Variation, Institute for Genome Sciences and Policy, Duke University, Durham, NC.

1. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 2008; 118:1590-605.
2. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516-7.
3. Weedon MN, Lango H, Lindgren CM, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet 2008;40:575-83.
4. Mallal S, Phillips E, Carosi G, et al. HLA-B*5701 screening for hypersensitivity to abacavir. N Engl J Med 2008;358:568-79.
5. Fellay J, Shianna KV, Ge D, et al. A whole-genome association study of major determinants for host control of HIV-1. Science 2007; 317:944-7.