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.

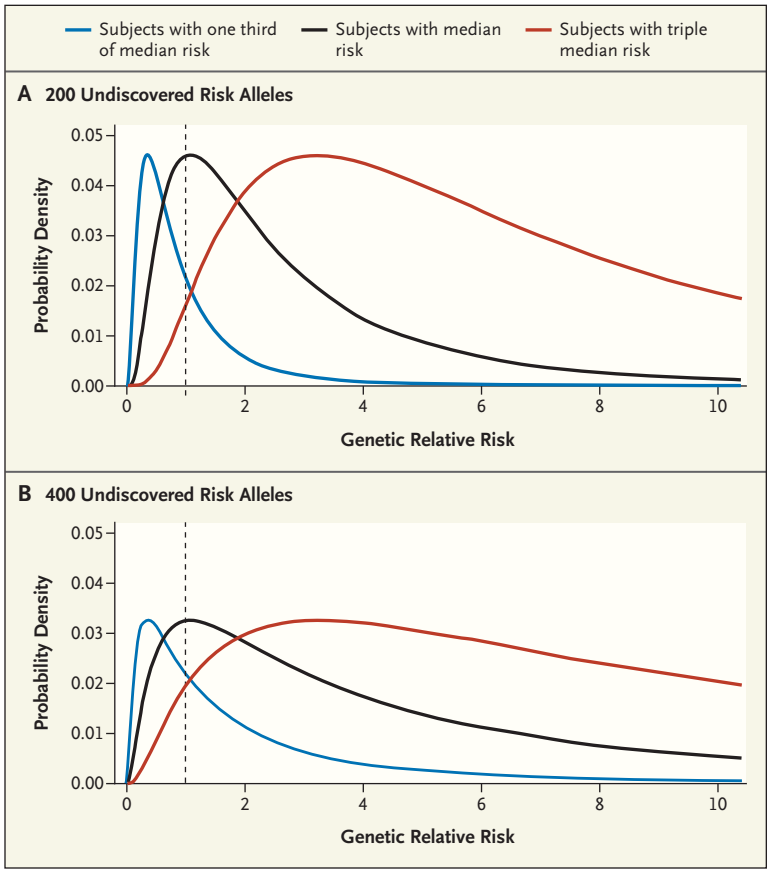
A major goal of the Human Genome Project was to facilitate the identification of inherited genetic variants that increase or decrease the risk of complex diseases. The completion of the International HapMap Project and the development of new methods for genotyping individual DNA samples at 500,000 or 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-




**Distribution of True Genetic Relative Risk of Disease for Three Subjects, According to Their Current Estimated Risk and the Number of Undiscovered Risk Alleles for the Disease.**

Panel A shows the genetic relative risk of disease for three subjects with differing estimated risks with the assumption that there are 200 undiscovered independent risk alleles for the disease, each with a frequency of 25% and a multiplicative relative risk of 1.1. Panel B shows the same scenario with the assumption that there are 400 such undiscovered risk alleles. Risks are relative to the prevalence in the general population. The dashed line represents a relative risk of 1.0, the population median risk.

ciated with a family history of the disease. Third, in the case of diseases that have been the focus of several genomewide association studies, some alleles have been detected more than once, but each study has identified multiple alleles that were not identified in other studies, suggesting that many more alleles remain to be discovered.

These factors suggest that many, rather than few, variant risk alleles are responsible for the majority of the inherited risk of each common disease. The good news is that pooling the results of multiple genomewide association studies has led to increased statistical power and the discovery of many new loci linked to small increases in the risks of major diseases and phenotypes. For example, pooling efforts have led to the identification of more than 16 new loci associated with diabetes and more than 30 loci linked to Crohn's disease. From a scientific perspective, we would like to know roughly how many risk loci remain to be discovered. From a clinical and policy perspective, we would like to know the extent to which the available associations are useful for measuring risk, both in absolute terms and relative to the expanded set of associations that are likely to be discovered in the next few years.

The table shows the approximate number of risk loci (with allele frequencies and relative risks similar to those of most markers that have been discovered through genomewide association studies) that would be needed to reach a level of risk equivalent to the sibling recurrence risks of complex diseases such as diabetes, heart disease, and many cancers. Even a relatively large genomewide association study (one with 5000 case subjects and 5000 control subjects) has a rather low power to detect any specific marker: 0.9% at a P value of $10^{-7}$, the least stringent accepted measure of genomewide significance, for an allele with a frequency of 25% and a relative risk per allele of 1.1. However, because there are so many risk loci, the probability of detecting at least 1 is good (83% if there are 200 risk loci). Although we know of many more risk loci than we did 2 years ago, there are probably many more associations that are yet to be discovered for these complex diseases. Less common variants in the prevalence range of 0.5 to 5.0% also remain to be discovered.

Estimates of risk based on established locus associations are therefore likely to change substantially in the next few years. The graphs show the distribution of true genetic relative risks under the assumption that there are either 200 undiscovered locus associations (Panel A) or 400 undiscovered locus associations (Panel

**Number of Risk Alleles Needed to Produce a Sibling Relative Risk of 1.5, 2.0, or 3.0.***

| Relative Risk Per Allele | Sibling Relative Risk | | |
|---|---|---|---|
| | 1.5 | 2.0 | 3.0 |
| | | *no. of risk alleles* | |
| 1.10 | 203–507 | 347–867 | 550–1374 |
| 1.20 | 51–135 | 87–231 | 138–367 |

* The number of risk alleles was calculated over a range of allele frequencies (10 to 90%); the minimum and maximum numbers are presented. All alleles were assumed to have the same frequency and relative risk and to be independent.

B), each with a risk-allele frequency of 25% and a per-allele relative risk of 1.1, for three hypothetical subjects: one who is estimated (on the basis of her current genetic profile) to have one third of the median level of risk, one who is estimated to have a median level of risk, and one who is estimated to have three times the median level of risk. There is a high degree of variability around the current estimates. If there are 200 undiscovered locus associations, more than 7% of the subjects who are estimated to have triple the median risk of disease actually have less than the median risk; if there are 400 undiscovered associations, 15% of such subjects would be in that category.

As the number of known risk loci increases, the correlation between the predicted risk and the actual risk will also increase.[4] But this correlation, though important, is only one of the factors determining whether knowledge of genetic risk is beneficial. The clinical value of a genetic test also depends on its sensitivity, specificity, and positive and negative predictive values; the costs and benefits of interventions; and the availability of data linking specific variants to improved clinical outcomes.[5] In particular, although we will be better able to distinguish subtle differences in risk as we discover more risk loci, most people will still be at or near the median level of risk. As a result, for less-common diseases (with a prevalence of 1% or less), the positive predictive value of a genetic test will almost always be low.

We are still too early in the cycle of discovery for most tests that are based on newly discovered associations to provide stable estimates of genetic risk for many diseases. Although the major findings are highly unlikely to be false positives, the identified variants do not contribute more than a small fraction of the inherited predisposition. Estimates that are based on combinations of the current risk alleles (even estimates indicating substantial relative risks for a very small number of persons who carry many risk alleles) will undergo constant revision as new loci are found. Such estimates are poor predictors of risk, both in absolute terms and in relation to risk estimators that will be available when more of the remaining locus associations are discovered.

The rapid progress being made through meta-analyses suggests that many more common variants conferring a risk of disease will be identified in the next several years, leading to increasing stability of individual risk estimates. Once risk estimates are more stable, the usefulness of genetic screening will need to be considered for each disease, and recommendations about potential interventions will need to be made for persons whose predicted risk exceeds some threshold. Although testing for inherited susceptibility on the basis of common risk alleles is premature for most diseases, the situation may be very different in just 2 or 3 years. Appropriate guidelines are urgently needed to help physicians advise patients who are considering this form of genetic testing as to how to interpret, and when to act on, the results as they become more stable.

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

This article (10.1056/NEJMp0810107) was published at NEJM.org on April 15, 2009.

Dr. Kraft is an associate professor of epidemiology and biostatistics at the Harvard School of Public Health, Boston. Dr. Hunter is a professor in the Departments of Epidemiology and Nutrition at the Harvard School of Public Health and serves as a statistical consultant to the *Journal*.

**1.** Couzin J. Genetics: DNA test for breast cancer risk draws criticism. Science 2008; 322:357.
**2.** Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 2008; 118:1590-605.
**3.** Janssens AC, van Duijn CM. Genome-based prediction of common diseases: advances and prospects. Hum Mol Genet 2008;17:R166-R173.
**4.** Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 2007;17:1520-8.
**5.** Katsanis SH, Javitt G, Hudson K. Public health: a case study of personalized medicine. Science 2008;320:53-4. [Erratum, Science 2008;320:316.]