STATISTICS AND MEDICINE

Drinking from the Fire Hose — Statistical Issues in Genomewide Association Studies

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

Related article, page 443

The past 3 months have seen the publication of a series of studies examining the inherited genetic underpinnings of common diseases such as prostate cancer, breast cancer, diabetes, and in this issue of the *Journal*, coronary artery disease (reported by Samani et al., pages 443–453). These genomewide association studies have been able to examine interpatient differences in inherited genetic variability at an unprecedented level of resolution, thanks to the development of microarrays, or chips, capable of assessing more than 500,000 single-nucleotide polymorphisms (SNPs) in a single sample. This "SNP-chip" technology capitalizes on a catalogue of common human genetic variations that is provided by the HapMap Project, which was made possible by the completion of the consensus human-genome sequence.[1]

The amount of data in these studies is four to five orders of magnitude greater than that in the previous generation of case–control studies, which tested only a handful of variants, often in a specific candidate gene. This unprecedented volume poses unusual statistical challenges for the analysis, display, and interpretation of the data.

The chief strength of the new approach is that it permits an "agnostic" genomewide comparison of gene-variant prevalence between cases and controls, obviating the need for guessing which genes are likely to harbor variants affecting risk. Most of the robust associations seen in this type of study have not been with genes previously suspected of being related to the disease. Some of these associations have been found in regions not even known to harbor genes, such as the 8q24 region, in which multiple variants have been found to be associated with prostate cancer.[2] Such findings promise to open up new avenues of research, through both the discovery of new genes relevant to specific diseases and the elucidation of new genetic mechanisms (e.g., the mechanism explaining why a region without known gene-coding loci would be associated with a disease).

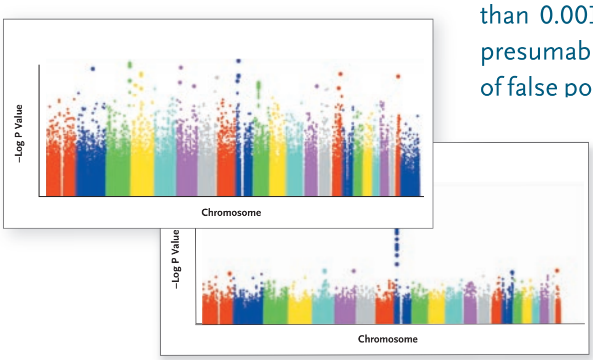
The chief strength of the new approach also contains its chief problem: with more than 500,000 comparisons per study, the potential for false positive results is unprecedented. One proposed solution is to adopt the approach conventionally used in much medical research — choosing a stringent P value at which statistical significance will be declared. To address the 500,000 or more comparisons, a Bonferroni approach can be used; for example, one can divide the commonly used P value of 0.05 by 500,000 to obtain a cutoff P value of 0.0000001 (10^{-7}), which is sometimes referred to as the threshold of "genomewide significance."

The main problem with this strategy is that, because of the high cost of SNP chips, most studies are somewhat constrained in terms of the number of samples and thus have limited power to generate P values as small as 10^{-7}. In addition, most variants identified recently have been associated with modest relative risks (e.g., 1.3 for heterozygotes and 1.6 for homozygotes), and many true associations are not likely to exceed P values as extreme as 10^{-7} in an initial study. On the other hand, a "statistically significant" finding in an underpowered study is more likely to be a false positive result due to chance than is such a finding in an adequately powered study,[3] and "statistically significant" associations could be attributable to systematic bias (e.g., from confounding due to ethnic ancestry, also known as population stratification). Thus, the sine qua non for belief in any specific result from a genomewide association study is not the strength of the P value in the initial study, but the consistency and strength of the association across one or more large-scale replication studies. Robust replication should permit the identification of true positive results and the weeding out of false positive results.[4]

Even with access to all the available primary-association data (see sidebar), it will probably still be desirable to select a subgroup of SNPs with the strongest associa-

Sharing the Rankings

The unprecedented number of comparisons being made in genomewide association studies using "SNP chips" has led to the recognition that no initially identified association can be relied on until it has been replicated in one or more studies of adequate size.[4] The process usually involves a multistage design, in which replication is attempted for a number of the SNPs that were found in the original study to have the most significant associations with the disease in question. Since genotyping a small number of SNPs is less expensive than using a SNP chip, such a design results in lower overall costs than using SNP chips for all studies. The main drawback is that if the P value for association for a given SNP in the initial study is not sufficiently small, the SNP will not be carried forward to the second stage of analysis — yet the association thus dismissed may actually be falsely negative.



If more than one genomewide association study has been conducted for a specific disease, an obvious alternative process of replication is to use one study to assess all the SNP associations found in the other study. In this issue of the *Journal,* Samani et al. compare a genomewide association study for coronary artery disease conducted in the United Kingdom with one conducted in Germany. Another use of two or more studies is to combine their data to provide increased statistical power for selecting the SNPs for smaller-scale replication in future studies. Again, Samani et al. provide an example of this approach, although by limiting their joint analysis to SNPs for which associations had P values of less than 0.001 in at least one scan, presumably to limit the number of false positive results, they have probably missed an opportunity to "resurrect" false negative results in either scan that did not meet their P value cutoff.

The benefits of these approaches suggest that if groups conducting genomewide association studies agree to share data — or better still, to make their data public in a format that permits other groups to obtain the results easily — progress in identifying causal loci will be accelerated. Although many groups conducting such studies have not declared their intentions regarding data availability, there are some encouraging examples. The National Cancer Institute's Cancer Genetic Markers of Susceptibility project has made the P values, relative risks, and confidence intervals from its genomewide association study of breast and prostate cancers available before publication (at http://cgems.cancer.gov), and investigators from the Diabetes Genetics Initiative have done the same (www.broad.mit.edu/diabetes). Samani et al. have committed to making the primary data from their two genomewide association studies available through a registration procedure (see the Data Access section of their article). The National Institutes of Health is finalizing a policy that may oblige grantees to make such data available through sites such as its Genotype and Phenotype database (dbGaP; www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gap), again through a registration procedure. This level of access to the full results from human studies is novel and should speed the identification of genetic variants associated with the diseases and other phenotypes that are the subject of genomewide association studies. We hope that many more examples of the benefits of data sharing will be forthcoming.

tions to use in studies of additional sets of samples for less than the cost of a full genomewide study. An obvious approach is to pick the highest-ranked SNPs, according to either a P-value threshold or the number of SNPs that can be genotyped by the platform being used in the second-stage study.

Samani et al. use a variation on this approach — identifying the highest-ranked SNPs using the false-positive-report probability,[3] which incorporates a priori assumptions about the number and strength of expected true associations between SNP markers and coronary heart disease — along with a SNP-specific estimate of the statistical power of the study to detect the association. The authors assumed that the probability of the association of a marker with disease did not depend on the genomic context (i.e., that intronic and nongenic markers were as likely as nonsynonymous coding SNPs to be associated with disease) and that all markers with a true association would have the same effect size, so the only factor that varies in the calculation of the false-positive-report probability is the allele frequency. The rankings of probabilities of association were similar to those based on crude P values. In principle, this approach could be used to increase or reduce the weights of markers, depending on the genomic context, and could be expanded to a fully Bayesian analysis incorporating the expected distribution of effect sizes.[5] The approach has the virtue of providing an estimate (based on the strong assumptions listed above) of the probability that the association is falsely positive.

Another statistical measure frequently used in the reporting of results from genomewide association studies is the population attributable fraction, often called the population attributable risk — an estimate of the percentage of cases of disease that would be avoided if the exposure were removed. This statistic combines information about the strength of the association, or relative risk, with information on the prevalence of the exposure (in this case, the genotype). Thus, mutations that convey very high relative risks of disease (such as mutations associated with familial hypercholesterolemia) but that are rare in the population are estimated to have low population attributable risks. Common polymorphisms imparting much smaller increases in risk may be estimated to have substantial population attributable risks.

For example, Samani et al. estimate that the variants they identified have population attributable risks of 10%, 11%, and 22%, with a combined estimate of 38%. Although this value suggests that they have discovered the causes of an impressively high percentage of cases of coronary heart disease, readers should be aware of some awkward properties of this measure. Individual population attributable risks cannot simply be summed to give the combined value; the sum of 10%, 11%, and 22% is 43%, as compared with the combined estimate of 38%. This fact complicates the combining of the estimates across studies, since their sums can exceed 100% — and clearly, we cannot prevent more than 100% of cases of a disease. Nor can we factor in the contribution of the additional, yet undiscovered, gene variants that researchers are confident they will find as they continue to comb through data from genomewide association studies. Thus, the population attributable risk provides a rough guide to the relative contribution of a gene variant to disease but should not be interpreted too literally, not least because its literal interpretation — which involves the hypothetical removal of the relevant exposure — does not apply as readily to gene variants as it does to modifiable environmental exposures.

The avalanche of recent data provided by genomewide association studies represents a quantum leap in information about the inherited component of certain diseases. However, a few caveats should be noted. Although SNP chips provide a vast quantity of information on common genetic variation, there is a substantial proportion of the known common variation that they do not capture. Manufacturers are producing newer chips, with probes for as many as 1 million SNPs, that will increase coverage, particularly for persons of African ancestry, suggesting that the rescanning of samples would uncover some loci missed by earlier generations of chips. Non-SNP gene variants, such as small deletions and insertions, are not formally represented on the SNP chips (although some of them may have SNP surrogates). Gains and losses of larger chromosomal segments, including variation in the number of copies of genes, have recently been found to be more common than previously appreciated. Identifying such variants will require special analysis of the chips, which has not been performed by most researchers to date.

This first wave of genomewide association studies is producing an impressive list of unexpected associations between genes or chromosomal regions and a broad range of diseases. There have been few, if any, similar bursts of discovery in the history of medical research. Relatively conventional statistical techniques are adequate for the analysis and interpretation of these initial studies. But as we delve further into the genome in the search for networks of interacting gene variants and interactions between these networks and environmental factors,[5] much more sophisticated methods of statistical analysis are likely to be required.

Dr. Hunter is a professor of epidemiology at the Harvard School of Public Health, Boston, a statistical consultant to the *Journal,* and codirector of the National Cancer Institute's Cancer Genetic Markers of Susceptibility project. Dr. Kraft is an assistant professor of epidemiology and biostatistics at the Harvard School of Public Health, Boston.

This article (10.1056/NEJMp078120) was published at www.nejm.org on July 18, 2007.

1. Christensen K, Murray JC. What genome-wide association studies can do for medicine. N Engl J Med 2007;356:1094-7.
2. Witte JS. Multiple prostate cancer risk variants on 8q24. Nat Genet 2007;39:579-80.
3. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 2004;96:434-42.
4. NCI-NHGRI Working Group on Replication in Association Studies, Chanock S, Manolo T, et al. Replicating genotype-phenotype associations. Nature 2007;447:655-60.
5. Thomas DC, Clayton DG. Betting odds and genetic associations. J Natl Cancer Inst 2004;96:421-3.

FOCUS ON RESEARCH

Rheumatic Heart Disease in Developing Countries

Jonathan R. Carapetis, Ph.D., F.R.A.C.P.

Related article, page 470

Only 30 or 40 years ago, rheumatic fever was a common topic in the *Journal.* A PubMed search for articles on rheumatic fever published between 1967 and 1976 returned 55 *New England Journal of Medicine* articles — fewer than for endocarditis (77) but more than for stroke and syphilis (24 entries each). A similar PubMed search for the decade 1997 through 2006 yielded just eight entries for rheumatic fever. This trend holds for all Medline-indexed journals: an average of 516 articles on rheumatic fever per year from 1967 through 1976, but only 172 per year from 1997 through 2006. Most observers would probably consider this decrease to be a reasonable reflection of the waning incidence of the disease. After all, in the mid-20th century, children with rheumatic fever occupied many of the beds in pediatric wards in industrialized countries — indeed, entire hospitals were dedicated to the treatment of, and rehabilitation from, rheumatic fever. But in the latter half of the 20th century, rheumatic fever receded as an important health problem in almost all wealthy countries. Today, most physicians in these countries are unlikely ever to see a case of acute rheumatic fever, and their experience with rheumatic heart disease will be limited to heart-valve lesions in older patients who had rheumatic fever in their youth.

The reality, however, is that the decrease in publications reflects only the waning burden of disease among the less than 20% of the world's population living in high-income countries. For everyone else, rheumatic fever and rheumatic heart disease are bigger problems than ever. It was estimated recently that worldwide 15.6 million people have rheumatic heart disease and that there are 470,000 new cases of rheumatic fever and 233,000 deaths attributable to rheumatic fever or rheumatic heart disease each year.[1] These are conservative estimates — the actual figures are likely to be substantially higher. Almost all these cases and deaths occur in developing countries.

How did rheumatic fever become rare in wealthy countries? Medical science can take some of the credit, thanks largely to the use of penicillin for primary prevention, but most of the reduction is attributable to improved living conditions, which have resulted in less overcrowding and better hygiene, with consequent reductions in transmission of group A streptococci. In other words, rheumatic fever is a disease of poverty. That it is in many ways the epitome of diseases of poverty and social injustice is exemplified by the situations in Australia and New Zealand. In these countries, which boast living standards that are among the best in the world, there are indigenous populations,