nature genetics

Framework for a fully powered risk engine

The tools for genome-wide association studies are now available. Here we present the journal's current criteria for manuscripts in this area of research.

The scope of the journal encompasses two main areas of research, both anticipated to occupy researchers for many years. One is the 'wiring diagram': the architecture of the genome sufficient to predict its cellular and organismal functioning (*Nat. Genet.* **37**, 1; 2005). The other, the 'risk engine', can be thought of as a comprehensive actuarial table in which one could look up the contributions of natural variants and mutations to any human disease phenotype in a given environmental and genetic 'background'. This catalog will be the sum of all the association studies in the field of genetic epidemiology.

What are *Nature Genetics* referees and editors looking for? Overall, the ideal study is one in which the weight of the evidence strongly supports the association reported. It should be written in a candid and transparent style, with a view to its utility to other researchers. It should contain a thorough and skeptical attempt to examine alternative explanations for the association (*Am. J. Hum. Genet.* **73**, 711–719; 2003). Although individual studies may not meet all these criteria, successful ones will meet many of them.

First, the genetic and statistical evidence for association should be sound. Molecular biological evidence for a functional variant is desirable in addition to, but will not substitute for, sound genetic evidence. Although conclusive evidence for functional variants is rare and not essential in this field, the conclusions drawn from the association should not be biologically impossible. The alleles described should affect the gene product in a physiologically meaningful way, and expression or activity of the gene or its downstream targets should be shown to be altered in affected individuals.

Because meta-analysis (*Nat. Genet.* **33**, 177–182; 2003) has shown that many published associations could not be replicated, we now stipulate that the association should be observed in two independent cohorts. At very least, the report should include the overall hypothesis, the numbers of individuals with each genotype and phenotype, the genotypic odds ratio with confidence interval, and the significance level before and after correction for multiple testing. Supplementary information can contain details of each statistical procedure done. This will allow referees and other researchers to check the assumptions and rerun the analysis with different parameters.

Threading a line between type I and type II errors is the classical problem (the Scylla and Charybdis, as it were) of statistical analysis. To discourage wrongly rejecting and wrongly accepting null hypotheses, we published recommendations for suitably large sample sizes (*Nat. Genet.* **29**, 306–309; 2001) and suitably small *P* values (*Nat. Genet.* **36**, 1045–1051; 2004). Explicit discussion of the assumptions of the study and frank evaluation of its potential pitfalls are appreciated by referees and readers alike. For example, the variant in question might be tested for genome-wide significance, or it may be considered a better candidate than a randomly picked marker because of other information. In the latter case, a Bayesian approach may be used, providing an estimate of the range of prior probabilities used in calculating the significance. Because this is not routinely done in a formal way, discussion of the nature and effect of the supporting evidence is crucial.

In controlling type I errors, it is necessary to state explicitly the number of genotypes, phenotypes and hypotheses tested and the procedures carried out to correct for the multiplicity of statistical tests. Although more power equals less chance of type II error, there is no strict consensus about the power that a study requires. Therefore, this should be discussed along with the range of effects that could be detected for a given population size and significance level. On page 1217 of this issue, Paul de Bakker and colleagues provide an analysis of simulated association studies to guide the use of HapMap markers in genome-wide association studies so as to retain full power and maximize efficiency.

Thorough investigation of alternative causes for a false positive association is considered essential. Referees frequently expect authors to present an analysis of linkage disequilibrium in the region of the most significant variant, particularly when no unambiguous functional variant is identified. Such analysis has been greatly aided by the publication of the common haplotype structure of several human populations by Perlegen Sciences, Inc. (*Science* **307**, 1072–1079; 2005) and by the International HapMap Consortium (*Nature*, published online 26 September 2005; doi:10.1038/nature04226).

The HapMap should become the collection of markers used by most researchers as a reference, and its population data provide a framework for the investigation of the genetic structure of sample populations. Methods for controlling stratification have been discussed (*Nat. Genet.* **36,** 1129–1131; 2004), but thinking in this area is evolving rapidly. For example, neutral markers or even ancestry-informative markers both throughout the genome and in the vicinity of the significantly associated gene can now be investigated in cases and controls to provide evidence for the absence of potentially confounding population substructure. Family-based investigations are more resistant to the effects of population structure than are larger population-based associations; hence, including both experimental designs in a study is likely to increase the chance of successful review. ■