Genome-wide association studies getting more complicated but help is on the way

For several years now, human geneticists have been expecting that, with the availability of a dense set of genetic markers across the genome and high-throughput genotyping technologies, genome-wide association studies will be feasible and will lead to the discovery of major genetic factors contributing to susceptibility to common human diseases. By coordinated international efforts, a reference human genome sequence was determined and millions of single nucleotide polymorphisms (SNPs) were identified. Several million SNPs have been genotyped by the International HapMap Consortium and the resultant HapMaps have been used to develop sets of markers that can be typed on high-throughput genotyping platforms at reasonable cost. Just as numerous genome-wide association studies to investigate a number of complex traits are getting started, however, new developments in the field are forcing researchers to re-think their strategy. Despite these complications, however, human geneticists are already coming up with strategies to address these problems, and the papers found in this issue of *Human Genomics* are representative of this movement.

As Stranger and Dermitzakis point out in their review, susceptibility to many complex traits may involve multiple loci and their interactions with each other and the environment, often associated with altered gene regulation. They argue that, in addition to the genetic profile, one may have to examine the gene expression patterns of the cases and controls in order to establish the links between genotype and phenotype. Because gene expression is often tissue specific and affected by environmental changes, patient recruitment and sample acquisition will be complicated. With newer expression arrays that study all of the exons in the genome, however, and a careful choice of tissues in these studies, gene expression data will probably be combined with SNP genotyping data in association studies in the future.

In order to increase the chance of finding the 'major' genetic factors involved in a common disease, the phenotype of the cases (and controls) needs to be as homogeneous as possible. This is hard enough for most common diseases but it is especially difficult in the study of cancer, a multi-stage process. In their paper in this issue, Savage and Chanock describe the pitfalls in over-interpreting association study results in cancer studies and propose a way to move towards the goal of using genetic data in clinical oncology. As better clinical definitions of diseases are used in patient collections, and as similar populations are used in replication studies, the chances of establishing 'real' associations will improve with time.

The recent finding that there are a significant number of structural variations in the human genome is making human geneticists re-think the way that association studies are conducted. Because these variations and polymorphisms (in the form of copy number variants, inversions, insertions, deletions and complex rearrangements of multiple kilobase regions) are difficult to detect and cause aberrant behaviour in SNP genotyping data, they pose a difficult problem for the experimentalist and statisticians alike. In this issue of the journal, Carson and colleagues describe the strategies one can use to detect structural variants and show that realisation of their presence in a candidate region can help the researchers to interpret seemingly confusing data to their benefit.

A vexing problem for those engaged in genome-wide studies is that false-positive results are the rule because of multiple testing with millions of markers. Hidden population structure may result in genetic differences between the cases and controls that have nothing to do with disease status. This increases the chance of obtaining false-positive results in genome-wide studies. Moreover, it is a difficult problem to solve in studies involving members of one ethnic group, such as those of European descent. In their paper, Liu and Zhao describe a non-parametric strategy for addressing the issue of population structure and show that it performs well against other methods designed to solve the same problem.

Additional help is found in the area of direct association studies based on a large set of functional SNPs across the genome. As Carlton and colleagues report in this issue, functional SNP-based studies are efficient because they focus on the SNPs with the highest likelihood of being responsible for the disease. Coupled with comparative genomics showing that functional elements tend to be found in conserved regions, application of bioinformatics methods (see paper by van Driel and Brunner) will soon enable one to design a comprehensive set of functional SNPs for genome-wide direct association studies.

As the papers in this issue have shown, these issues are being recognised and addressed in ways that give one hope that genome-wide association studies will begin to identify genetic factors responsible for disease susceptibility. Examples of genome-wide studies using promising approaches are

found in a number of studies, including those on human ageing described herein by Kaeberlein. With deeper understanding of the occurrence of structural variations in the genome; the development of new tools to study them in association studies; new methods for the assessment of population structure and expression profiles; a comprehensive set of functional SNPs; better defined patient samples; and appropriate design of replication studies, one can expect genome–wide association studies to bear fruit in the near future.

Pui-Yan Kwok, MD, PhD
Professor of Dermatology
Henry Bachrach Distinguished Professor and Investigator
University of California, San Francisco
USA