Editorial

The impact of low-cost, genome-wide resequencing on association studies

With the completion of phase 1 of the HapMap project (www.hapmap.org), we are now close to the point where genome-wide association studies form a routine tool for trying to identify genes involved in human disease and drug response. In any of three large human populations, the HapMap provides more than 500,000 single nucleotide polymorphisms (SNPs), chosen as far as possible to be evenly spaced and highly polymorphic. Genotyping these SNPs in large samples of cases and controls should permit any common variant (for example, minor allele frequency [MAF] $>$ 5 per cent) implicated in disease causation to be identified, even if the size of its effect is rather small. Phase 2 of HapMap, expected to be delivered later in 2005, will generate an even more dense SNP resource.

Even before the HapMap results have been applied to benefit disease gene studies, the project is already yielding a wealth of information on human population genetics. The enormous fine-scale variability in recombination rates is being documented for the first time; much is being learned about the effects of selection on human genetic variation, and more generally about the nature of genetic polymorphisms and their distribution in the genome and across the globe.

As researchers investigating particular genetic diseases assess the implications of HapMap for their work, however, it is already possible to look beyond to an era in which much of the HapMap data becomes largely redundant. The major motivation for HapMap was that, by genotyping a small proportion of the genome (500,000 SNPs represents less than 0.02 per cent of it) and exploiting the power of linkage disequilibrium (LD) to illuminate the unobserved variants in the vicinity of a typed polymorphism, much of human genetic variation can be captured cost-effectively. If whole genome resequencing becomes fast and cheap, however, why bother with only part of the information?

Several companies are promising to make that prospect a reality within a short time-frame. In March 2005, 454 Life Sciences of Connecticut, USA, announced that it had sold and installed its first genome sequencing system, claimed to perform sequencing 100 times faster than conventional machines, at the Broad Institute of MIT and Harvard. The system, based on light-emitting sequencing chemistries and microfluidics, implements massively parallel genomic sequencing. According to the company, a single instrument can sequence over 20 megabases per four-hour run. In December 2004, the company announced that its technology had been used to sequence four entire bacterial genomes, one from *Mycobacterium tuberculosis* and three from related strains, facilitating the discovery of a potential antimicrobial treatment for tuberculosis targeting a newly identified pathway. Sequencing larger genomes — eventually that of *Homo sapiens* itself — seems not to be far away.

Solexa, based in California, USA and Cambridgeshire, UK, is in the same race towards fast and cheap resequencing of human genomes. In March 2005, it announced the resequencing of the virus Phi-X 174, whose small genome has often been used as a 'proof of principle' of new technologies. Solexa's current technology involves 'clusters' — dense collections of DNA molecules on a surface — and novel chemistry that allows a single base extension per cycle, thanks to reversible termination and fluorescence. A resequencing system based on these technologies — and capable of sequencing much larger genomes — is due for commercial release late in 2005. For the future, Solexa is investigating single-DNA-molecule technology, allowing massively parallel processing with around 10^8 sites per cm^2. Working at the single-molecule level avoids the need for amplification, allowing 'one-tube' sample preparation for a whole-genome analysis.

How much extra should researchers be willing to pay to get whole-genome sequence data, rather than genotypes at a dense SNP map? Clearly, the former includes the latter, and so should be preferred if cost is irrelevant. But, of course, cost is crucial. For genetic association studies, at least 500 cases and 500 control genomes are typically required, meaning a cost of less than $5,000 per genome for a $5m overall cost. This is still a long way off, even with the new technologies, but their small reagent costs makes the $5,000 genome a feasible proposition for the forthcoming years.

Data quality is as important as cost. SNP genotyping platforms have recently been improved, such that miscall and noncall rates can reach very low levels. The resequencers are also claiming very low error rates, but achieving this, and high coverage rates, will have cost implications. The issue of error rates is complicated by translocations, inversions, duplications, tandem repeats and indels of varying sizes, some of which will be virtually impossible to capture fully with any one technology. Since many of these are candidates for being involved in disease causation, however, capturing most of them would provide a major advantage for sequence over SNP data: SNPs can sometimes 'tag' non-SNP variants such as indels or inversions but, broadly speaking, these need to be common and known *a priori*.

More generally, the main potential advantage of sequence over genome-wide common SNPs is the capturing of rare variants, whether they be SNPs or more complex variants, and whether or not their existence is recognised *a priori*. Even cases attributable to multiple spontaneous mutations — resistant to study by classical genetics — can potentially be successfully unravelled using resequence data. The distribution of allele frequencies in natural populations is U-shaped, with most alleles having a frequency close to zero or one: common variants are rare and rare variants common. The belief that many of the variants causing common diseases are common — known as the 'common disease common variant' (CDCV) hypothesis — is backed by some theory and data, but sceptics may nevertheless suspect that wishful thinking has also played an important role in motivating adherence to CDCV. An alternative is that common diseases are caused by a number of different polymorphisms, many or even all of them rare: there may be a number of pathways potentially contributing to disease progression, each modified by a specific set of polymorphisms, or a common pathway may be interrupted by mutations at a number of widely-dispersed genomic locations.

Genome-wide resequencing would greatly simplify genomic variation analyses: the problem of highly-variable LD between a disease-predisposing polymorphism and even a close marker locus vanishes because the causal polymorphism itself will be typed in the study. This potentially leads to more power for a given sample size, because effect sizes are not attenuated by variable LD. Conversely, sequencing errors and coverage gaps can mar the analyses, whereas SNP genotyping is now highly robust. It may be optimal to accept a moderate level of resequencing errors and gaps, in order to increase the number of individuals sequenced within the available budget. Another possible route to cost reduction is multiple uses of a control sample for comparison with different case samples. This cannot be done indiscriminately, but some pooling of controls across studies may be feasible and help to make resequencing cost-effective.

Naturally, the large genome centres are taking an interest in these developments, which for the moment remain out of reach of most researchers. The time-scale for widespread implementation of resequencing technologies is hard to predict, but it seems prudent to bear in mind the possibility that fast, cheap genome resequencing may soon be within reach.

David Balding
Editorial Board
Human Genomics