nature genetics

New handles on genomic structural variation

Procedures for genotyping structural variants with SNP detection arrays now permit many of the larger and more common polymorphisms to be incorporated into association studies.

The study of copy number variation (CNV) has matured from the discovery phase to the point where it is feasible to undertake reliable genotyping for association testing. Using SNP genotyping for common variants and 270 of the same DNA samples used in the HapMap program, Steven A. McCarroll and colleagues (p 1166) report that they were able to re-identify 73% of the structural variants longer than 5 kb that had been previously reported from eight of these samples by fosmid sequencing or high-resolution oligonucleotide array hybridization (Kidd *et al. Nature* **453**, 56–64; 2008). Further, the group found that even the largest structural variants are much smaller than previously estimated using large insert arrays and are consequently likely to affect the expression of fewer coding genes than was predicted. On the basis of the variants detectable with their platform, the group claims that relatively common polymorphisms account for much of the structural variation between individuals of a population.

In contrast, Gregory Cooper and colleagues (p 1199) compare currently used genotyping arrays with paired-end clone sequences from nine genomes and conclude that although about half of common deletions are effectively tagged by SNPs on the arrays, duplicated regions are much less well covered. In theory then, it should be possible to improve the performance of tagging arrays incrementally, but there is of course the inherent problem that SNPs in segmentally duplicated regions have been selected against in the design of genotyping array platforms, and many CNVs arise in regions flanked by segmentally duplicated sequences. In some regions, there is a lack of SNPs, but more commonly, the SNPs that are present in these regions are confusing to call.

Recognizing the need for dedicated analytical procedures to make sense of the data, Joshua Korn and colleagues (p 1253) present a suite of software tools to process array-probe intensities. They do this sequentially, clustering diploid SNP genotypes, then making copy number calls on the resulting clusters. With this approach, it was possible to assign the genotypes of two copy number alleles in Hardy-Weinberg equilibrium to many loci that had previously looked rather more complicated. In a complementary Technical Report, Chris Barnes and colleagues (p 1245) develop a procedure that incorporates CNV scoring and association testing into a single statistical model. In this way, they are able to assess and deal with the inherently noisy data that is generated from the genotyping of structural variants. The method is also quite robust to differences in the distribution of genotyping errors between cases and controls.

An emerging picture is that there seems to be no single answer to the question, “Where in the genome is the missing heritable variance in the human population?” With much of the search still to be undertaken, it is now already clear that large polymorphic variants are unlikely to account for a large proportion of common disease phenotypes. Recurrent large rearrangements are diverse and individually rare and there is an abundance of small deletions (and much more besides) still to be investigated.

Thanks to these new papers, many of the tagged common variants of all sizes and the haplotypes upon which small structural variants are found can now be incorporated into genome-wide association studies. The discovery of structural variants is very much an ongoing process, and because most of the variants are small, high-density oligonucleotide arrays and even sequencing will be needed for reliable detection of them. These papers will accelerate discovery of additional structural variants by refocusing research effort where it is needed. ■