GENOMICS

Finding copy-number variants

Several studies evaluate high-density single-nucleotide polymorphism (SNP) arrays for the detection of copy-number variations in human genomes.

Many debilitating diseases have complex genetic roots, and the challenge is to unravel the interplay between multiple mutations and their phenotype. No two genomes are alike; instead, each displays structural variability in the form of single-nucleotide polymorphisms (SNPs), deletions or insertions of various sizes, which are collectively called copy-number variants (CNVs), and inversions, which are copy number–neutral structural variants.

Since a map with the position of SNPs in the human genome was established a few years ago, SNPs have proven invaluable as markers in genome-wide association studies. The goal is to identify those SNP alleles that are more frequent in individuals with the disease of interest so that the region around the SNP can be analyzed for polymorphisms that contribute to the disease.

An effective tool to genotype SNPs in large populations are microarrays, in which each probe is designed to discriminate between a single nucleotide difference. These arrays are expected to yield genotype data that cluster into three discrete categories for each SNP: two clusters for homozygous individuals and one for the heterozygous phenotype. To achieve these results the companies that produce the arrays used to exclude genomic areas that appeared to violate Mendelian segregation and did not divide into three distinct clusters.

Consequently the arrays yielded clean SNP genotypes, albeit at the expense of more complex structural variants, which often fall into the regions that had been filtered out. Researchers were increasingly dissatisfied with the notion that more complex structural variations such as CNVs would have to be analyzed in separate studies and instead wanted to obtain CNV data directly from SNP arrays.

Common structural variations in the genome.

The two most important questions for researchers seeking to perform genome-wide analysis of SNPs and CNVs in the same study were: 'how effectively do SNP arrays capture CNVs?' and 'how can scientists analyze the data from such arrays to collect accurate information about CNVs in each patient?'.

A collaborative effort between Affymetrix and a research team at the Broad Institute, led by Steven McCarroll, Joshua Korn and David Altshuler, resulted in the redesign of a commercial SNP array (McCarroll *et al.*, 2008). This 'hybrid array', called Affy SNP 6.0, combines traditional SNP probes with what McCarroll calls "copy-number probes," targeted to regions known to contain CNVs. The scientists developed computational approaches to draw a high-resolution CNV map from the data.

Illumina, another company manufacturing SNP arrays, has also developed arrays that target potential CNV regions.

Gregory Cooper in the laboratory of Debbie Nickerson and Evan Eichler at the University of Washington took the lead in systematically testing common SNP platforms for CNV detection (Cooper *et al.*, 2008). The team developed new algorithms to extract the most information from older (Illumina HumanHap 300) and new (Illumina Human 1M) genome-wide arrays and found that although they could not detect the majority of CNVs on the older arrays, the new array performed better.

To accurately measure CNVs on these new array platforms in large cohorts of patients, researchers need new software tools.

With an algorithm developed for analyzing deletions, Cooper and colleagues genotyped deletions with an average size of 30 kilobases.

The group at the Broad Institute released Birdsuite, software that treats the analysis of common copy-number polymorphisms as a clustering problem, analogous to SNP genotyping, rather than a mutation-discovery problem. Birdsuite clusters the population of individuals into discrete classes corresponding to each person's integer number of copies at a given locus (Korn *et al.*, 2008). This allowed them to capture over a thousand common CNVs.

Matthew Hurles and his team from the Sanger Institute provide a comprehensive statistical framework for CNV association tests in case-control studies (Barnes *et al.*, 2008).

A prerequisite for genotyping CNVs is of course a detailed knowledge of the regions they occur in. As the authors of Cooper *et al.* point out, at least 20% of CNV-containing areas are still not covered even on the new arrays.

"The important thing is to get higher probe density in the messier regions of the genome," Cooper concludes, "that is, regions that are already duplicated several times. We need probes that can reliably distinguish four, five or six copies."

Once all regions prone to copy-number polymorphisms are known—large sequencing projects such as the '1,000 genome project' will supply the necessary data—scientists will be in a position to design even more informative arrays so that CNVs of all sizes can be queried during genome-wide association studies.

Nicole Rusk

RESEARCH PAPERS

Barnes, C. *et al.* A robust statistical method for case-control association testing with copy number variation. *Nat. Genet.* **40**, 1245–1252 (2008).

Cooper, G.M. *et al.* Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat. Genet.* **40**, 1199–1203 (2008).

Korn, J.M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).

McCarroll, S.A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).