

Next-generation genome

Sequencing technology is now advanced enough to decode individual human genomes. Will it prove to be better than existing methods for discovering the genetic basis of human phenotypic variation?

A year ago we designated as Method of the Year the now rather inaptly named next-generation sequencing—the set of faster and cheaper sequencing technologies that make it possible to obtain genomic sequence information on a previously unimaginable scale. ‘Next-gen’ sequencing of an entire human genome had not been achieved at the time. This has changed in the past year, with the sequencing of James Watson’s genome using 454 technology, and of the genomes of a Yoruban and a Han Chinese individual using Illumina technology.

The year has also seen the initiation of large collaborative sequencing projects aimed at individual genomes: the international 1,000 genomes project and a project to sequence the genomes of at least 100 Chinese individuals at the Beijing Genomic Institute at Shenzhen. The goal of these efforts is not, as it was for the human genome project, to sequence a few people to build a single reference human genome, but rather to sequence very many people to build a detailed resource of genomic variation, including single-nucleotide polymorphisms (SNPs) and structural variations.

Next-gen sequencing is therefore poised to change the way the genetic basis for human complex traits, including disease risk, is studied. It receives the baton, so to speak, from microarray technology. Using high-density SNP microarrays, and based on the previous genome-wide mapping of human SNPs in the HapMap project, large-scale population studies have allowed geneticists to uncover SNPs that are associated with common human traits or diseases.

However, the variants that have been uncovered by these genome-wide association (GWA) studies are usually responsible for only a small fraction of disease risk. This probably reflects both the underlying biology of complex traits, as well as the fact that GWA studies may so far not have found several of the contributing genetic variants. Will more widespread use of sequencing technology prove more adept at comprehensively identifying the genetic variants underlying human disease and other traits? This will depend on the nature of the contributing, as-yet unidentified variants and on where in the genome they lie.

One set of SNPs that are likely to have been missed by existing GWA studies are rare SNPs, present in the population at a frequency of 1% or lower. Resequencing of thousands of individual genomes at loci that have already been

associated with a trait is likely to bring these rare, potentially causal SNPs to light; this approach has already borne fruit in some cases, such as for identifying genes involved in lipid metabolism. But distinguishing the rare, functional single-nucleotide change from noise comes with its own set of challenges, and will require high technical and computational accuracy.

In addition, structural variants—duplications, small insertions and deletions, and larger structural rearrangements—are likely to be important contributors to phenotypic variation and to disease. Sequencing is in principle more suitable for detecting such structural changes, although array-based tools continue to be improved. However, the shorter read lengths of next-gen methods are still a limitation for characterizing such variants, and mapping them within repeat-rich regions of the genome also remains a challenge.

Some traits or diseases, in contrast, may result from a combination of many common variants of extremely low penetrance. These would remain undetected in GWA studies if the sample sizes are not large enough. Sequencing technology in itself will not add power to the analysis; to discover these types of variants, still larger association studies would be needed. At present cost, this is unlikely to be widely feasible using sequencing.

So next-gen sequencing will help identify elusive genomic variants with different degrees of success, depending on the variant properties. As costs decrease—with the maturing of the third generation of sequencing technology, for instance—it will become increasingly feasible to apply whole-genome sequencing directly to large population studies. In the meantime, current efforts will catalogue human genomic variation at finer resolution and help identify optimal methods for studies at population scale.

And there are additional applications to which the current technology is amply suited. Interactions between genes, regulation of gene expression at transcriptional and post-transcriptional levels, and epigenetic effects, among others, are all very likely to play a role in human phenotypes. Next-gen sequencing can already be applied to study some of these phenomena—methylation, transcription, protein factor binding to DNA, or microRNA biology, to name just a few—and should thereby contribute to functional annotation of the genome and to a deeper understanding of how genomic variants translate into phenotypic variation and disease.