The 1000 Genomes Project: Welcome to a New World

The 1000 Genomes Project is an international research consortium that was set up in 2007 with the aim of sequencing the genomes of at least 1,000 volunteers from multiple populations worldwide in order to improve our understanding of the genetic contribution to human health and disease. Global support was contributed by major institutions, including the Wellcome Trust Sanger Institute (UK), the Beijing Genomics Institute (China), and the US National Human Genome Research Institute. The project goal was to produce a catalogue of human variation down to variants that occur at 1% frequency or less over the genome, in order to facilitate genetic studies on common human disease (1).

A major paper, published in the October 1, 2015, issue of *Nature*, marks the completion of the final phase of the colossal project: a comprehensive, open-access database of genetic variation from 2,504 individuals from 26 populations across the globe (2). The genotypes were obtained using a combination of whole-genome sequencing, deep exome sequencing and high-density single nucleotide polymorphims (SNPs) microarrays. The characterization of the variants was based on a set of 24 sequence analysis tools. Overall, the project discovered and characterized more than 88 million variants, including 84.7 million SNPs, 2.6 million short insertions/deletions (indels), and 60,000 structural variants, that were integrated into a high-quality haplotype scaffold.

A few salient findings: As compared to the reference human genome, a typical genome differs at ~4 to 5 million sites, 99.9% of these variants being SNPs and short indels. The number of variant sites is greatest in individuals from African ancestry, as expected from the out-of-Africa model of human expansion. Analyses of the variants most likely to affect gene function revealed that a typical genome contained ~150 sites with protein truncating variants, ~10,000 sites with peptide-sequence altering variants and ~ 500,000 variant sites overlapping regulatory regions such as promoters, enhancers, or transcription factor binding sites. Importantly, ~2,000 variants per genome were associated with complex traits through genome-wide association studies (GWAS) and 24 – 30 variants per genome implicated in rare diseases through ClinVar (a database of the relationships among human variations and phenotypes). Other analyses provided information about population history, demography of ancestor populations, and resolution of genetic association studies (2).

The results of the 1000 Genomes Project, which attest to the benefits of "consortium-based science," complete a set of genomic information that has already been in use for several years. Such information is particularly useful for the design of genotyping arrays, population genetics (e.g. genotype imputation in GWAS, defining variants in regions of interest, filtering of likely neutral variants), and investigations on natural selection, population structure, and admixture. The major advantages of the 1000 Genomes Project data set include the broad representation of human genetic variation (with a much improved coverage of South Asian and African populations); the use of multiple analysis strategies, increasing the quality of filtering and mapping and allowing the capture of more diverse types of genetic variants; and the wide availability of samples and data resulting from the project. Altogether, these elements will help to provide further insights into the genetic basis of disease. They will be used, for instance, in the ongoing efforts to decipher the genetic basis of peritoneal transport and the outcome of peritoneal dialysis.

"Now this is not the end… But it is, perhaps, the end of the beginning" as Winston Churchill said. Large-scale sequencing projects will continue for more regional or ethnic groups, in order to extend the global coverage. Much effort will focus on a better understanding of the relationship between genetic variation and common disorders. The translation of this massive genetic information to human health will benefit from the development of complex databases gathering genetic, clinical, and biological data, such as multi-omics profiles, while maintaining protection of potentially sensitive personal information (3). Efforts are also underway to increase genetic awareness in the public and to educate health professionals (http://www.1000genomes.org/about).

Olivier Devuyst
University of Zurich, Zurich, Switzerland

olivier.devuyst@uzh.ch

REFERENCES

1. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, *et al*. A map of human genome variation from population-scale sequencing. *Nature* 2010; 467:1061–73.
2. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM,

Garrison EP, Kang HM, *et al.* A global reference for human genetic variation. *Nature* 2015; 526:68–74.
3. Devuyst O, Knoers NV, Remuzzi G, Schaefer F. Rare inherited kidney diseases: challenges, opportunities, and perspectives. *Lancet* 2014; 383:1844–59.
http://dx.doi.org/10.3747/pdi.2015.00261