Credit: Cosmo Condina / Alamy Stock Photo

MILESTONE 13

Cataloguing a public genome

Within 10 years of the first human genome being sequenced, the cost of next-generation sequencing was reducing rapidly. Combined with advances in sequencing technology, this meant that exploring the genetic variation in large numbers of people was becoming increasingly tangible.

The 1000 Genomes Project was launched in 2007 with the aim of forming a public reference database to catalogue human genetic variation across major human population groups. In 2012, Gil McVean and colleagues presented phase one, the genomes of 1,092 individuals across 14 populations (including Europe, East Asia, sub-Saharan Africa and the Americas). They combined targeted deep exome sequencing, low-coverage whole-genome sequencing and dense single-nucleotide polymorphism (SNP) genotyping to establish one of the first high-quality resources of its kind. A key feature of this resource included a haplotype map consisting of 38 million SNPs. Low-frequency variants in the non-coding genome were also characterized, showing the ways in which purifying selection acts at functionally relevant sites in the genome. It highlighted the need for rare variants to be interpreted in the context of local genetic background and produced the first 'null expectation', that is, the number of rare, low-frequency and common variants one would expect to find in individuals across different populations.

The methods, technologies and systems developed to form the dataset provided a framework for the generation of large-scale genetic catalogues.

By the time the project reached its final phase in 2015, the dataset represented more than 2,500 individuals from 26 human populations and had contributed to or validated 80% of the variants in the dbSNP catalogue. In the final iteration, Adam Auton and colleagues painted a picture of what a typical human genome looks like, that is, one that consists of around 4–5 million sites that differ from the human reference genome and with variation that differs substantially among population groups.

More recent catalogues of human genetic variation have included the Exome Aggregation Consortium (ExAC) and the Genome Aggregation Database (gnomAD). Released in 2016, the ExAC database was the aggregation of exome sequencing data from more than 60,000 individuals of diverse ancestries (European, African, South Asian, East Asian and Latino ancestries). ExAC has now been replaced by gnomAD, which contains 125,748 exomes and 76,156 genomes in its current release, representing the largest catalogue of human variation.

> “The aggregation of population-level genetic datasets has given us unprecedented access to the depths of the genome”

This catalogue includes high-quality loss-of-function (LOF) annotations, allowing researchers to zoom in on per-base variation and gene-level selection. LOF variation can be used as an in vivo model for human gene inactivation and thus an indicator of a gene's intolerance to inactivation. By comparing expected versus observed variation, Konrad Karczewski, Monkol Lek and colleagues created metrics to quantify the sensitivity of individual genes to variation. These gene-level metrics, as well as estimates of allele frequencies, have been invaluable for the human clinical genetics community when filtering or classifying candidate pathogenic variants in rare disease analyses.

The aggregation of population-level genetic datasets has given us unprecedented access to the depths of the genome. Moreover, the public access of these databases set a precedent for related fields to continue in this spirit of data sharing.

While efforts have been made to include diverse population groups, many of these datasets are nearly devoid of populations from regions such as the Middle East, Oceania and large proportions of Africa. There is still a long road ahead to improve the representation of understudied population groups, making initiatives such as GenomeAsia 100K and H3Africa crucial. We also expect to see more initiatives linking genotype and phenotype information, such as the *All of Us*, 100,000 Genomes and TOPMed programmes. Each of these initiatives are already showing potential to fine-tune our understanding of the impact of genetic variation on human health and disease.

Ingrid Knarston,
Nature Communication

ORIGINAL ARTICLES McVean, G. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012) | Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016) | Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020)
FURTHER READING Durbin, R. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010) | Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015)