One pangenome to bind them all

The pangenome provides a first glimpse of the scope of human genetic diversity. But its routine adoption into research and clinical practice faces several challenges.

This summer, the Human Pangenome Reference Consortium (HPRC) released its first draft human pangenome. The release contains 47 genomes from individuals in the 1000 Genomes Project that cover>99% of the existing reference genome sequence (GRCh38) with>99% accuracy at the structural and base-pair level. Thus far, HPRC has generated 119 million base pairs of additional genome sequence and identified 1,529 gene duplications missing from GRCh38. A key aspect of the HPRC effort is its ability to identify structural variants, over two-thirds of which are estimated missing from GRCh38. Given the importance of structural variants to disease, the pangenome represents a leap forward for clinical genetics. The challenge for its adoption, however, will be whether laboratories will spend the necessary time, effort and money to update their sequencing pipelines to incorporate HPRC's graph genomes.

Twenty-one years since the Human Genome Project's first 'draft', the current reference sequence (GRCh38) still covers only ~92% of the human genome. The latest patch released in May, GRCh38.p14, represents the 14th update on a reference that is now nearly nine years old.

Age is not the only problem for GRCh38. First, it is a biased consensus haploid genome built from a mishmash of different individuals: 70% of it derives from one person (RP11), with>20 individuals contributing the rest. GRCh38 reflects neither the most common alleles nor the longest nor the most ancestral. And it over-represents sequences of European and African origin.

The second problem is incompleteness. Of the reference's 3,099,441,038 bases, 8% remain unknown due to 349 sequence gaps, segmental duplications or other unknown errors. These gaps have remained for decades at centromeres, at telomeres, and on acrocentric chromosomes containing ribosomal RNA genes — where highly repetitive or homopolymer stretches have confounded most sequencing technology. This results in sequence reads failing to map properly.

Finally, the reference genome does a poor job of representing human genetic variation. Since the Genome Reference Consortium began stewardship of the reference assembly in 2007, the last two reference genomes have added alternative sequences at positions with large heterogeneity (GRCh38.p14 has 434 alternative loci). But homozygous alternative and rare alleles, segmental duplications and most structural variants (many of which are medically important) remain under-represented. To make matters worse, not all alignment tools make use of the provided alternative sequences — and those that do often preferentially align to the standard reference. This creates a 'streetlamp' effect where more research is conducted on sequences contained within the reference.

In the past three years, two consortia, the Telomere-to-Telomere (T2T) consortium and HPRC, have attempted to address these problems. Both capitalize on recent advances in long-read sequencing technology and a renaissance in 'graph-first' assembly methods.

The mission of T2T is to obtain accurate sequences for all 24 human chromosomes, starting with the publication in 2020 of the X chromosome. Earlier this year, the consortium published the first 'gapless' human genome sequence. Using a diploid female cell line (CHM13hTERT) that has near-complete homozygosity to minimize assembly uncertainties, T2T employed PacBio high-fidelity 'HiFi' and continuous long-read sequencing supplemented with Oxford Nanopore Technologies ultralong sequencing to tile across repetitive sequences and build a near-complete genome through graph assembly. The published genome has one error every 10.5 Mb (quality value, Q70.22; (CHM13v0.9); it has since been polished further, although ribosomal RNA genes remain unfinished.

HPRC aims to assemble 700 reference-quality haplotypes from 350 individuals at a quality similar to that of the recent T2T assembly, improving representation of genomic and geographic diversity. The effort uses a combination of T2T technologies and graph-first assembly approaches that rely on several common steps: homopolymer compression (collapsing error-prone stretches of bases to a single base), read cleaning and correction using statistical techniques, and string graphs that contain perfect (rather than 'fuzzy') overlaps because HiFi sequence is so good. 'Easy' and 'hard' tangles in the string graph assembly are resolved by Hamiltonian walks and spanning Oxford Nanopore long reads, respectively. HiFi sequence reads are then used to create the consensus sequence and the homopolymer tracts decompressed.

The result — two decades after the first draft — is a near-complete human genome reference. This is particularly important for clinical genetics, where the pangenome can produce fewer ambiguous mappings, provide more accurate analyses of copy number variation, and resolve multiallelic regions of high clinical importance (for example, the human leukocyte antigen (HLA) locus) that fail to be captured in the existing linear reference.

But routine clinical implementation of a human pangenome reference will require new bioinformatics methods capable of querying and operating on it. Unlike the gapless T2T genome, which has a single haplotype, the technology for querying pangenomes with gapless diploid assemblies remains a work in progress.

And most clinical genetics laboratories are unprepared: going from raw sequencing reads to a short-list of variants of clinical importance involves multiple computational tools maintained by different research groups or organizations, as well as queries to external databases like gnomAD, TCGA or GWAS catalogs.

Unlike basic research, a clinical genetics facility must also meet regulatory quality standards (such as the US Clinical Laboratory Implementation Amendments, or CLIA). And every time a sequencing pipeline is changed, it needs recertification by regulatory authority; indeed, many clinical laboratories still use the GRCh37 build from February 2009 for their alignments. More training will clearly be needed to explain how additional sequences included in the pangenome reference relate to GRCh37 or GRCh38.

Thus, a great deal of software development, standards, recertification and education will be needed to switch to a pangenome reference. This will not happen overnight. But already in basic research, its adoption is improving the mapping of structural variants. And the payoff in terms of reduced diagnostic odysseys for patients with rare diseases makes it a change for clinical laboratories that will be well worth the investment. ❐

Published online: 30 August 2022
https://doi.org/10.1038/s41587-022-01484-y