E pluribus unum

If the human reference genome is to reflect more of the actual genomic diversity in humans, community participation is needed.

The human genome is ten years old. We acknowledge its reference assembly as an invaluable resource essential for many purposes such as the assembly of short reads from high-throughput sequencing platforms into chromosome context during resequencing projects. At the same time, we think necessary improvement of the reference genome depends on the willingness of the research community to provide data for the genome's less accessible regions.

First published in 2001, the human reference genome has, since 2007, been in the hands of the Genome Reference Consortium (GRC) a small group of fewer than 20 scientists from the European Bioinformatics Institute, the US National Center for Biotechnology Information, The Sanger Institute and The Genome Center at Washington University in St. Louis, who have committed to the improvement and completion of this reference, with very little financial support.

The reference genome is now in its 19th rendition, and probably the best measure of its improvement over the last ten years is the number of fragments it consists of. The very first version had ~150,000 gaps; the most recent build, GRCh37, has only around 250 gaps.

The only other publicly accessible *de novo* assembly of a human genome that contains chromosome sequences is HuRef. Obtained by traditional capillary sequencing, HuRef is the diploid genome of Craig Venter. It comes in 4,500 pieces and, like any individual genome, it contains many rare alleles.

GRCh37, in contrast, is a mosaic haploid genome derived from about 13 people. It still contains rare alleles, but the GRC recently decided to convert these to common haplotypes. Deciding which alleles are common and which are rare is proving challenging, and the GRC members are collaborating with members of the 1000 Genomes project to collect enough data to make these decisions.

A display of common haplotypes will help in the alignment of some sequences that currently cannot be placed. This will increase the appeal of the reference, which in turn should increase the willingness of researchers to help address another challenge the GRC faces, namely that of closing the gaps.

In very broad strokes, there are two types of gaps, nonstructural and structural. Sequences in the former do not contain structural variations—that is, extensive insertions, deletions or inversions—whereas sequences in the latter do. Next-generation sequencing technology, in particular the longer reads of the 454 platform, can help close the nonstructural gaps, as demonstrated by Chad Nusbaum's team at the Broad Institute who closed three gaps in chromosome 15. But gaps in areas with structural variation require classical sequencing technologies because longer reads are needed to span these regions and anchor them in known sequence. On page 365, Evan Eichler's team shows that Sanger sequencing of fosmid clones containing the genomes of nine individuals can add high-quality sequence to many genomic locations, many of which show clear differences between Europeans, Asians and Africans.

Incorporation of areas in which sequences diverge greatly between individuals or populations, where a common haplotype cannot be easily designated because many haplotypes of similar frequency exist, presents another difficulty for the GRC.

In the current GRCh37 build, there are only three loci that have alternative haplotypes. The MHC region, which has four million base pairs on chromosome 6, a 0.7-million-base-pair locus on chromosome 4 encoding an enzyme for drug metabolism, and the 1.5-million-base-pair *MAPT* gene on chromosome 17, a medically important region for which one haplotype predisposes carriers for a microdeletion syndrome associated with mental retardation. All these alternative assemblies were added to the reference in collaboration with researchers who work on the respective regions. There are likely hundreds of these regions and many of them may be medically interesting.

Of course, data collection is not everything. The GRC also needs to decide how to incorporate alternative loci while keeping the genome easy to navigate for the average user. Currently the public has access to the corrections that are made on a daily basis, but it takes considerable expertise to access and work with the most up-to-date version of the reference genome. The GRC is considering a system that keeps the coordinate system stable but allows locus-specific updates.

Eventually all these changes will entail a major overhaul of the sequence and its coordinates. High-quality data from the research community are required if the still unscheduled next build, GRCh38, and future updates are to adequately reflect the true diversity of human genomes. The GRC has neither the money nor the staff to single-handedly pay sufficient attention to all the regions that need it.