nature biotechnology

A reference standard for genome biology

The Vertebrate Genome Project provides a new benchmark for those seeking to build reference genomes.

The first set of reference genomes recently released by the Vertebrate Genome Project (VGP) represents a watershed for genome sequencing. The VGP intends to generate reference genomes of species from all 260 vertebrate orders; ultimately, it plans to identify the nucleotide sequences of all 66,000 vertebrates. These reference genomes are notable not only for their breadth but also for their completeness, accuracy and haplotype-phased, chromosome-by-chromosome assemblies. The project employs state-of-the-art sequencing, mapping and computational technologies and draws on expertise worldwide to create a resource that promises unprecedented insights into vertebrate diversity and evolution. It is also establishing a benchmark of quality standards and best practices for the genome-sequencing field.

Reference genomes are the cornerstone of modern genomics. These high-quality genomes are differentiated from draft genomes by their completeness (low number of gaps), low number of errors, and high percentage of sequence assembled into chromosomes. Although the genomes of viruses and some prokaryotes have complete end-to-end sequence information, nearly all eukaryotic genomes do not. Indeed, even the latest (19th) version of the high-quality human reference has hundreds of gaps, mostly in or near centromeres, telomeres, segmental duplications and ribosomal DNA arrays.

The VGP's first release of reference genomes includes four mammals, three birds, one reptile, one amphibian and five fish, many of which are endangered species. Remarkably, according to the VGP, the September release almost doubles in one swoop the number of high-quality vertebrate reference genomes available to the research community.

The effort, part of the Genome 10K Consortium, uses several technologies for genome sequencing and assembly: PacBio long reads, 10x Genomics linked Illumina reads, Hi-C chromatin mapping data and Bionano Genomics optical maps. The consortium aims to standardize the assembly and validation process to avoid systematic biases introduced by any one strategy. The results will provide a unique opportunity to discover what combination of technologies yields the best outcomes.

Sequencing today is largely dominated by Illumina's high-throughput, short-read (~100–200 base pairs) technology, which has made the process of decoding genomes much faster and cheaper. But this speed comes with a cost. The millions of short, overlapping reads generated by these instruments represent a complex puzzle that must be pieced together in the correct order and orientation. Computational algorithms are needed to assemble reads into continuous segments of DNA sequence (known as contigs) and to subsequently order and orient these contigs into chains (known as scaffolds), which often contain gaps. To improve scaffolding, additional long-range information is needed from technologies like Hi-C (which chemically cross-links neighboring chromatin domains), optical mapping (which visualizes fluorescent probes bound to single immobilized DNA molecules) and linked reads (sets of barcoded short reads from the same DNA molecule). Even these approaches still fall short when attempting to decipher intractable genomic features such as repetitive DNA, G+C-rich sequence or structural rearrangements that span distances much longer than a short read. Polyploidy, which is also often encountered in crop genomes, also presents challenges.

In contrast, single-molecule sequencing technologies, such as those from PacBio and Oxford Nanopore Technologies, produce reads long enough to span many genomic regions that are difficult to reconstruct with short reads. The greater length (ten kilobases to over a megabase) and higher error rate of long reads have stimulated the development of tailored assembly algorithms for correcting errors, improving contig formation and polishing assemblies. Hybrid approaches that use short-read data to correct errors in long reads before assembly offer another way of increasing contig length and closing gaps.

Despite this progress, important challenges remain. For complex eukaryotic genomes, the sequencing and assembling of reads is only the first step. This is often followed by months of manual curation and data checking to improve the quality of the genome sequence. The rigor with which this laborious 'finishing' step is undertaken tends to be project-specific and dependent on the funding available.

Beyond these practical concerns, there is no definitive method to verify the correctness of the finished product. For some species, even simple information like the number of expected chromosomes is unknown. In most cases, researchers perform several checks to evaluate the quality of a final assembly. The assembly size can be compared with existing genome size estimates for that organism or can be estimated using statistical approaches. Algorithms can be applied to identify all the sequences of set length (called *k*-mers) in an Illumina library that are likely to be real, and then to work out what fraction of those *k*-mers are recovered in the new assembly. The final assembly can also be inspected for 'core' genes—a set of genes common in species related to the sequenced species.

In the absence of a perfect measure of genome correctness, the definition of a high-quality genome continues to evolve, together with advances in sequencing and assembly. The VGP is setting a new standard of quality through its efforts to benchmark genome assemblies. The consortium requires that all released genomes have an N50 size of at least 1 Mb for contigs and 10 Mb for scaffolds (N50 is the minimum length such that 50% of the genome can be covered by contigs or scaffolds of that size or greater), that sequence error frequency be no higher than 1 in 10,000 bases, that structural variants be confirmed by multiple technologies, and that at least 90% of the sequence be assigned to chromosomes and haplotype phased.

Of course, not all genome-sequencing efforts are obliged to meet these stringent standards. But 17 years after the publication of the human genome, it is time for biology to move beyond draft genomes, with their inaccuracies and poor annotations that too often lead researchers down blind alleys. The VGP is the clearest indicator yet that the era of high-quality reference genomes is upon us.