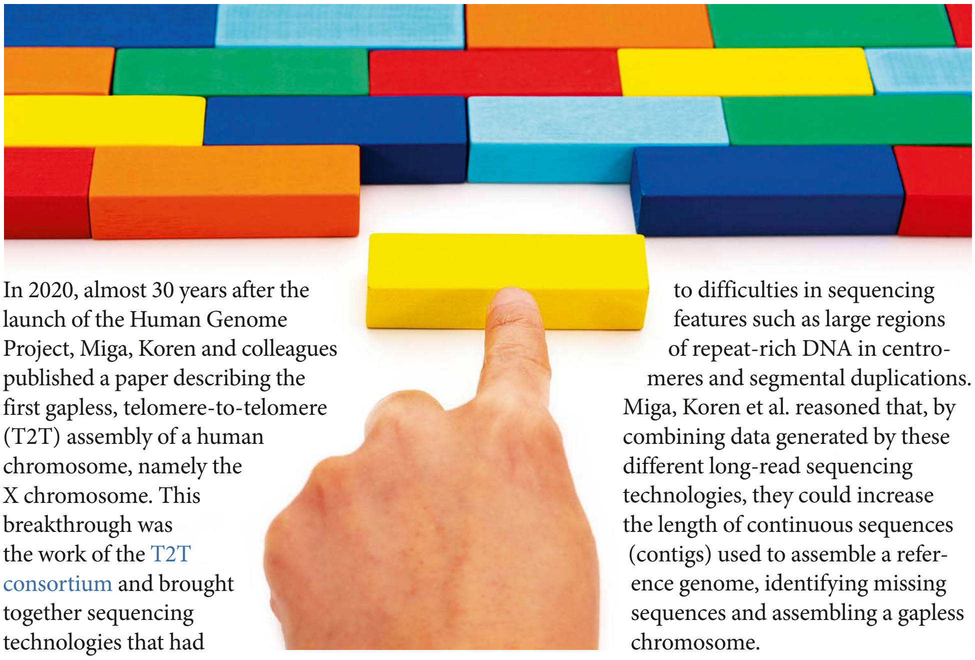
MILESTONE 17

Filling in the gaps telomere to telomere



Credit: PORNCHAI SODA / Alamy Stock Photo

In 2020, almost 30 years after the launch of the Human Genome Project, Miga, Koren and colleagues published a paper describing the first gapless, telomere-to-telomere (T2T) assembly of a human chromosome, namely the X chromosome. This breakthrough was the work of the T2T consortium and brought together sequencing technologies that had been developed in the preceding 6 years.

In 2015, Chaisson et al. showed that long-read sequencing technology from Pacific Biosciences (PacBio) could be used to sequence a human genome, specifically that of the complete hydatidiform mole (CHM) cell line CHM1. As CHM cells have a duplicated paternal (but no maternal) genome, bypassing the need to assemble both haplotypes of a diploid genome, they became a key reference genome. Later that year, Berlin, Koren et al. reported the first de novo assembly of a human genome based on PacBio sequencing long reads alone. Then, in 2018, Jain et al. revealed that ultra-long-read nanopore sequencing (from Oxford Nanopore Technologies) could also be used to assemble a human genome de novo (MILESTONE 8). Finally, in 2019, Wenger, Peluso et al. introduced PacBio high-fidelity (HiFi) sequencing, which was 99.8% accurate in sequencing the human genome reference standard HG002 over average read lengths of 13.5 kb.

Although these technological advancements were reported to have closed gaps in the GRCh37 or GRCh38 version of the human reference genome, no chromosome had been sequenced in full owing to difficulties in sequencing features such as large regions of repeat-rich DNA in centromeres and segmental duplications. Miga, Koren et al. reasoned that, by combining data generated by these different long-read sequencing technologies, they could increase the length of continuous sequences (contigs) used to assemble a reference genome, identifying missing sequences and assembling a gapless chromosome.

> “The sequencing of the first two complete chromosomes … suggested that it was technically possible to complete the human genome sequence”

Consequently, they sequenced 155 Gb of DNA from CHM13 cells with nanopore sequencing, using the genome assembly tool Canu to combine these ultra-long reads with data previously generated by PacBio sequencing. Nanopore sequencing, PacBio sequencing and linked-read Illumina sequencing were used to polish their assembly of the CHM13 genome, a 2.94-Gb assembly with a median consensus accuracy of ~99.99% and in which 50% of the genome was within contigs of ≥70 Mb. The presence of 41 of 46 telomeres at contig ends suggested that CHM13 was a more complete reference genome than GRCh38.

Indeed, Miga, Koren et al. noted that the X chromosome in their CHM13 assembly was broken in just three places. To fill in these gaps, they first mapped ultra-long reads against the assembly, manually identifying reads that joined breaks between contigs; this approach resolved two breaks resulting from segmental duplications. These findings were validated by mapping independent long-read PacBio HiFi data from CHM13 to the X chromosome. To resolve the third break, which was at the centromere, the researchers uniquely tiled ultra-long reads across the repeat-rich centromeric α-satellite array on the X chromosome, confirming the results with long-read PacBio HiFi data and benchmarking and improving the centromere assembly using an automated satellite assembly method (CentroFlye) and evaluation tools (TandemTools). After polishing, the gapless X chromosome assembly was ≥99.9% accurate and had resolved 29 reference gaps. By precisely mapping long-read data to the finished chromosome, the researchers also produced the first comprehensive, T2T profile of DNA methylation, enhancing our picture of epigenetic regulation over repeat-rich regions.

Sequencing of the X chromosome led the way to the T2T assembly of the first autosome, chromosome 8 from CHM13 cells, as announced by Logsdon et al. later in 2020. Combining nanopore, PacBio and PacBio HiFi sequencing, this work closed up five gaps in chromosome 8 and produced an assembly with an accuracy of >99.99%.

The sequencing of the first two complete chromosomes, 20 years after the release of the first draft human genome (MILESTONE 1), suggested that it was technically possible to complete the human genome sequence. Indeed, in September 2020, the T2T consortium announced that they had filled in all of the gaps, obtaining complete sequences for all the chromosomes in CHM13 cells (apart from the five ribosomal DNA arrays) and thus, outstandingly, a v1.0 assembly of a complete human genome.

Katharine H. Wrighton,
Nature Reviews Cross-Journal Team

ORIGINAL ARTICLE Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020)
FURTHER READING Berlin, K. et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015) | Chaisson, M. J. P. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015) | Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018) | Logsdon, G. A. et al. The structure, function, and evolution of a complete human chromosome 8. Preprint at *bioRxiv* https://doi.org/10.1101/2020.09.08.285395 (2020) | Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019) | Jain, M. et al. Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323 (2018) | Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017) | Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017)