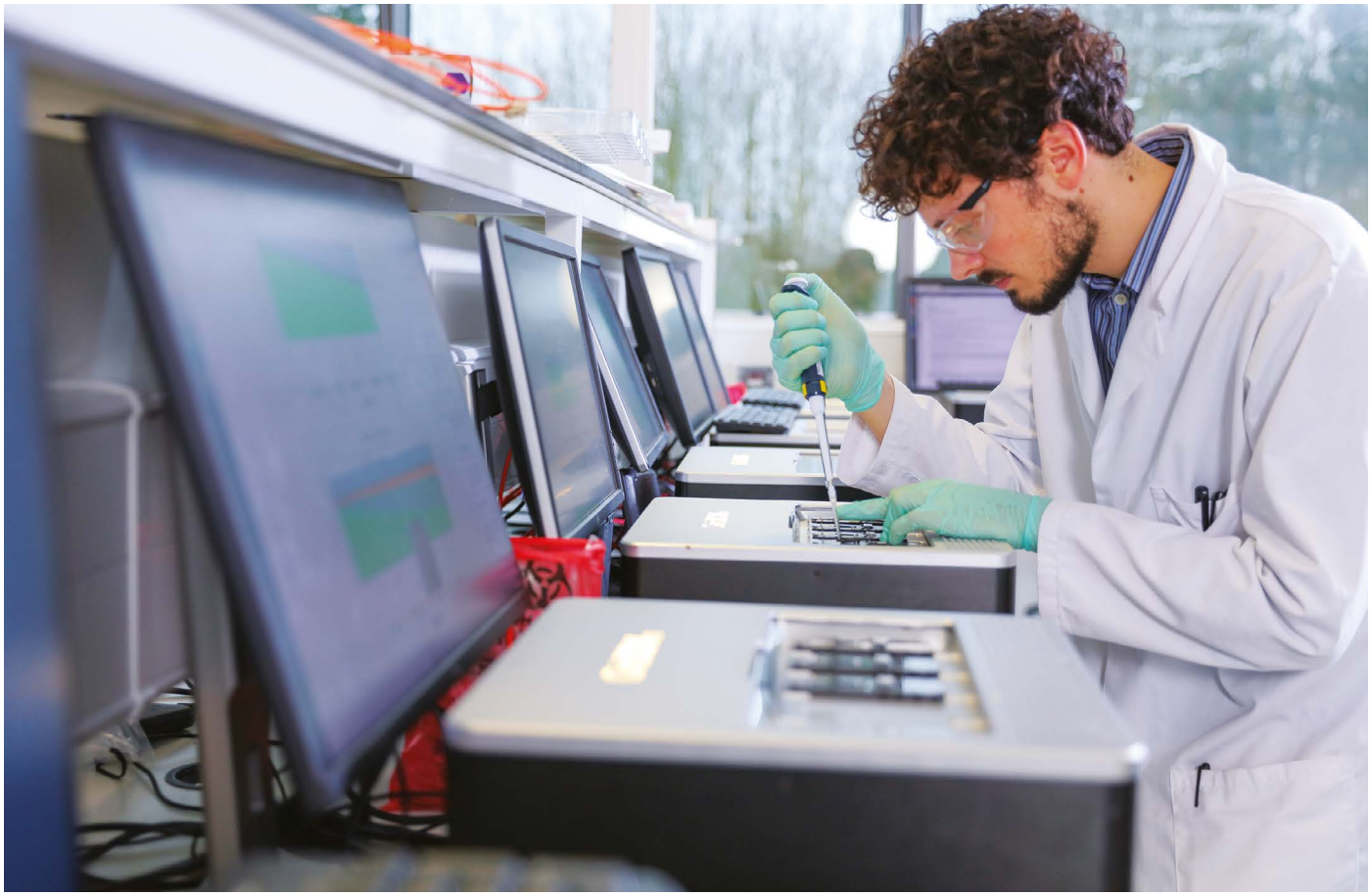

OXFORD NANOPORE TECHNOLOGIES

'Nanopore' sequencers, such as these GridION units, can decode tens of thousands of bases of DNA without interruption.

CLOSING IN ON A COMPLETE HUMAN GENOME

Advances in sequencing technology mean that scientists are on the verge of finally finishing an end-to-end human genome map. **By Michael Eisenstein**

With a complex and diverse topography of genes and regulatory sequences, the human genome is often likened to a landscape. But in many places, this terrain is less dramatic vista and more desert highway: vast and repetitive.

Consider a chromosome's centromere, which links its two gene-laden arms. Centromeres comprise thousands of near-identical α-satellite sequences — 171-base-pair units that need to be correctly organized to ensure chromosomal stability and cell division. Yet two decades after the publication of the draft human genome, these and other challenging DNA features remain as stubborn gaps in our chromosomal atlas. And, until a few years ago, some researchers despaired of ever filling them.

Beth Sullivan, a centromere researcher at Duke University in Durham, North Carolina, recalls a conversation in 2014 with Karen Miga, a genomics researcher at the University of California, Santa Cruz. "She told me, 'If something doesn't happen drastically with the technology, we're going to be stuck here for a long time'," Sullivan says.

But something did happen: the development of sequencing technologies that can read long stretches of DNA uninterrupted. Now Miga and her colleagues in the Telomere to Telomere (T2T) consortium are poised to complete the 20-year odyssey that began with the release of that first draft sequence. Their goal is to produce, for each chromosome, an end-to-end genome map that stretches from one telomere (the repetitive sequence elements that cap chromosomal ends) to the other. "This wasn't just doing it for the sake of doing it," says Miga. "It was because I think there's some really cool biology there." But to find it, the genomics world will need to sequence many such genomes, chipping away at the variation of these still poorly understood genomic regions.

Stuck in the middle

Published 20 years ago this month[1], the first draft of the human genome was a landmark achievement. But it was also full of holes. Scientists from the Human Genome Project generated vast numbers of short sequences from chromosomal DNA. Where they overlapped with their neighbours, these were assembled into larger, contiguous stretches known as contigs. Ideally, each chromosome would be represented by a single contig, but the

first draft consisted of 1,246 such fragments.

Since then, scientists working as part of the Genome Reference Consortium (GRC) have been fleshing out the assembly, manually checking it and using sequencing analysis to identify segments with errors and information gaps. The most recent version of the human genome, called GRCh38, was released in 2013. Since then it has been repeatedly 'patched'. Yet it's still missing 5–10% of the genome, including all the centromeres and other challenging regions, such as the large collection of genes encoding the RNA sequences that form protein-producing organelles called ribosomes. These are present in long stretches of numerous, repeated gene copies. "That's a large portion of the yet-to-be-closed gaps," says Adam Phillippy, a bioinformatician at the US National Human Genome Research Institute in Bethesda, Maryland, and T2T co-chair. The genome is also peppered with hard-to-map stretches of near-identical DNA called segmental duplications – the product of ancient chromosomal rearrangements.

These challenging sections have continued to stymie genome-assembly efforts. That's because most sequencing so far has been done with short-read technologies, such as the widely used platform commercialized by biotechnology company Illumina in San Diego, California. Illumina sequencers generate extremely accurate data, but typically over just a few hundred bases – too short to span the long repeats and position the sequences unambiguously. "Genes are usually easy to assemble," says Kerstin Howe, a computational biologist at the Wellcome Sanger Institute in Hinxton, UK, who is part of the GRC. "But everything else in that intergenic space or with lots of repeats was basically not addressable."

Reaching across the gaps

Two long-read technologies are now closing those gaps. Biotechnology company Pacific Biosciences in Menlo Park, California, uses an imaging system to directly read hundreds of thousands or even millions of DNA strands in parallel, each spanning thousands of bases. Another approach, commercialized by UK firm Oxford Nanopore Technologies, threads DNA strands through tiny protein pores, or nanopores, reading tens to hundreds of thousands of bases by measuring the subtle changes in electrical current that occur as nucleotides traverse the channel.

When they were first rolled out (Pacific Biosciences' technology in 2010 and Oxford Nanopore's in 2014), these technologies were more error-prone than that of Illumina, which delivers greater than 99% accuracy for individual reads. "We're talking about 15–20% error rates in the early PacBio reads," says Phillippy. First-generation nanopore sequencers could produce errors in more than 30% of the bases.

But performance steadily improved, and with it, read length. "Within the past three or four years, we could now get read lengths of over 100 kilobases," says Phillippy. "That's when Karen and I launched this T2T consortium."

Set up in early 2019, the consortium aims to produce high-quality, end-to-end assemblies for every human chromosome. More than 100 sequencing and genomics specialists from around the world have signed up, many of whom were already actively demonstrating the power of long-read-based analysis.

Two papers published in 2018 highlight their work. In one[2], computational biologist Matthew Loose at the University of Nottingham, UK, and his colleagues described the first human genome assembled entirely from Oxford Nanopore data. Previous long-read assemblies used Illumina data to correct the error-prone nanopore output. But Loose and his colleagues covered around 90% of GRCh38 with 99.8% accuracy using nanopore data alone, while also closing a dozen major gaps in the reference genome.

"We were able to have a backbone representation of those chromosomes from telomere to telomere."

In the second study[3], Miga and her team reassembled the centromere of the human Y chromosome, the genome's smallest. They produced numerous long reads across the region to generate high-quality consensus sequences in which random errors could be readily identified and eliminated. "We could actually traverse all the way across the centromere," says Miga. "But it was still very manual at that point – just looking at patterns and stitching them together."

First to finish

Such successes made it clear that the T2T's goal was within reach. To simplify its work, the consortium focused on CHM13, a tumour-derived cell line with a genome that comprises two identical sets of chromosomes. This eliminates the complexity of diploid genomes, with distinct chromosome copies from each parent.

In late 2020, T2T scientists published the first two complete assemblies, for chromosomes X[4] and 8 (as a preprint)[5]. The investigators used Oxford Nanopore technology to sequence pieces of the two chromosomes that routinely exceeded 70,000 bases in length, with one read surpassing one million bases. "With these, we were able to essentially have a backbone representation of those chromosomes from telomere to telomere, but at lower accuracy," says Phillippy. They then complemented those data with Illumina and Pacific Biosciences reads to polish their assemblies.

Glennis Logsdon, a postdoc in the lab of genome scientist Evan Eichler at the University of Washington in Seattle and first author on the chromosome 8 work, says that the different sequencing technologies have distinctive quirks. For example, T2T scientists have found that the Pacific Biosciences chemistry can struggle with genomic regions that are highly enriched in G and A bases, whereas nanopore technology sometimes stumbles over long repeats of the same nucleotide. "If one data set has a defect that the other one doesn't, they end up complementing each other well because of that," says Logsdon.

Completing and fact-checking the assemblies required specialized software tools developed by researchers, including Phillippy and computational biologist Pavel Pevzner at the University of California, San Diego. The team took a cautious approach. "We were only going to glue two sequences together if they're basically 100% identical over 7,000 bases of their length," says Phillippy. "Once you introduce an error into the assembly, it's very difficult to fix it." But by taking such care, he says, it became possible to produce assemblies with 99.99% accuracy at the nucleotide level.

The initial work[4] with chromosome X also benefited from previous knowledge of that chromosome's centromere, which has been well studied at the structural level. "We used a variety of molecular techniques to make sure that the size of the assembly of the α-satellite array from the sequencing information was correct," says Sullivan. "Overall, I was really impressed with the amount of validation that went into that first study."

The researchers also exploited mapping techniques, such as one developed by Bionano Genomics, a biotechnology company in San Diego, California, that make it possible to measure the distances separating DNA sequences on a chromosome.

Closing in on completion

Although successful, the T2T approach to chromosomes 8 and X was laborious and painstaking. But an important advance during this time gave the team's efforts a shot in the arm. Pacific Biosciences instruments support a process known as circular consensus sequencing (CCS), in which individual DNA strands are converted into closed loops that can be read over and over. By comparing these repeated reads, researchers can eliminate random errors to produce a highly accurate result.

Early versions of CCS topped out at a few thousand bases, limiting their use in genome assembly. But in 2019, the company revamped this process[6], and the resulting high-fidelity approach now produces consensus reads surpassing 20,000 bases with greater than 99% accuracy. "Some centromeres we now can assemble completely from high-fidelity

POWER AND SYRED/SPL

Human chromosomes imaged by a scanning electron microscope.

reads — no extra help is needed," says Pevzner, although he adds that well-calibrated algorithms that can work with such data are also required.

Pevzner compares centromere reconstruction to assembling a jigsaw puzzle of seemingly clear blue sky, in which all the pieces initially appear indistinguishable. "There are little, almost invisible, clouds there that can distinguish different pieces of the puzzle," he says. Finding those clouds reveals the puzzle's organization — and the revamped approach does the same with centromeres, sensitively detecting subtle sequence differences that can provide landmarks for assembly algorithms.

The combination of this approach with ever-longer nanopore reads markedly accelerated T2T's progress — Logsdon reports that hundred-thousand-base stretches are now routine. "It took us a year or more to do each of the chromosome X and 8 projects," says Phillippy, "but we were then able to essentially finish all the remaining chromosomes in a two-month span." Now the end is in sight. "We've green-lit all of the centromeric arrays except for the one on chromosome 9," says Miga. This centromere, she says, is massive — spanning 27 million bases — and has posed a special challenge in terms of validation. The team is also still finalizing the highly duplicated ribosomal RNA genes. But the consortium is already sharing its data on GitHub, and Miga anticipates that the complete genome release for the CHM13 cell line will arrive this year.

The data are already yielding insights. Logsdon and others have been using nanopore sequencing to find patterns of DNA chemical modification that can influence chromosomal function. "Most of the centromere is methylated, but there's this dip in methylation that seems to be found in all centromeres," she says. The dip seems to mark the location of the kinetochore, an essential centromeric structure that manages the equal partitioning of DNA during cell division. Logsdon hopes to use these findings to engineer minimal centromeres for synthetic chromosomes.

T2T's approach also made relatively short work of the vast and complex gene arrays that encode the variable regions of antibodies and receptors on the surface of the immune system's T cells. "They're highly repetitive and notoriously difficult to assemble," says Pevzner. "As of today, we have only two references for this region." The ability to access and characterize these challenging genomic segments could guide efforts to understand the immune response to infections and vaccines.

End of the beginning

Challenging as it has been to build, a single end-to-end genome offers researchers limited value without other genomes from diverse individuals against which to compare it. To boost its utility, in late 2020, the T2T began working more closely with a parallel effort, the Human Pangenome Reference Consortium (HPRC). The HPRC was launched in 2019 with the goal of replacing GRCh38 with a reference genome that better captures the scope of human diversity, based on whole-genome data from at least 350 individuals. "The more genomic medicine becomes routine, the more you will want to remove any bias that depends on the ancestry of a person," says Tobias Marschall, a computational biologist at the Max Planck Institute for Informatics in Saarbrücken, Germany, who is part of the effort.

Yuta Suzuki, a research associate in the lab of computational biologist Shinichi Morishita at the University of Tokyo, has used Pacific Biosciences sequencing to study the centromeres of 36 individuals from Japan and other parts of the world[7]. "Just within the Japanese population, we see different centromeres for virtually every sample we have investigated," says Suzuki. "It's not enough to have just one reference, or even just one reference for each population."

Morishita plans to analyse hundreds of additional human centromeres, and he notes that several dozen disease-associated genetic variations have been mapped to these regions. "This suggests there's something going wrong in the centromeric repeats, and our impression is that their stability might be destroyed due to structural variants," he says. For his part, Phillippy sees the opportunity to better understand diseases associated with cellular protein-production machinery once ribosomal RNA genes can be routinely resolved.

But first, researchers must work out how to apply the T2T process to a diploid genome. Determining which sequences reside on which chromosome copy requires scientists to identify enough unique genetic landmarks to confidently assemble distinct contigs for each DNA strand, a tough feat in ultra-repetitive regions such as the centromere. In their chromosome 8 preprint, Logsdon, Eichler and their colleagues describe the feasibility of reconstructing diploid centromeric regions from chimpanzees and humans, but only when the two chromosomes are highly genetically distinct. "We'll need much more accurate or longer reads to span the full centromere region for a diploid genome," says Morishita.

At present, most clinical-genomics efforts focus on known genes — a fast and cost-effective approach to genome analysis. But the pioneers exploring this new terrain expect that comprehensive analyses will ultimately become a standard, although probably more expensive, fixture in medical and research genomics — particularly as researchers begin routinely exploring the clinical impact of variations in these once-unmappable regions. "If my child was sick and I knew that I could get 100% of the genome with long-read, I would want to pay that difference," says Miga.

Michael Eisenstein is a freelance writer based in Philadelphia, Pennsylvania.

1. International Human Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).
2. Jain, M. *et al. Nature Biotechnol.* **36**, 338–345 (2018).
3. Jain, M. *et al. Nature Biotechnol.* **36**, 321–323 (2018).
4. Miga, K. H. *et al. Nature* **585**, 79–84 (2020).
5. Logsdon, G. A. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.09.08.285395 (2020).
6. Wenger, A. M. *et al. Nature Biotechnol.* **37**, 1155–1162 (2019).
7. Suzuki, Y., Myers, E. W. & Morishita, S. *Sci. Adv.* **6**, eabd9230 (2020).