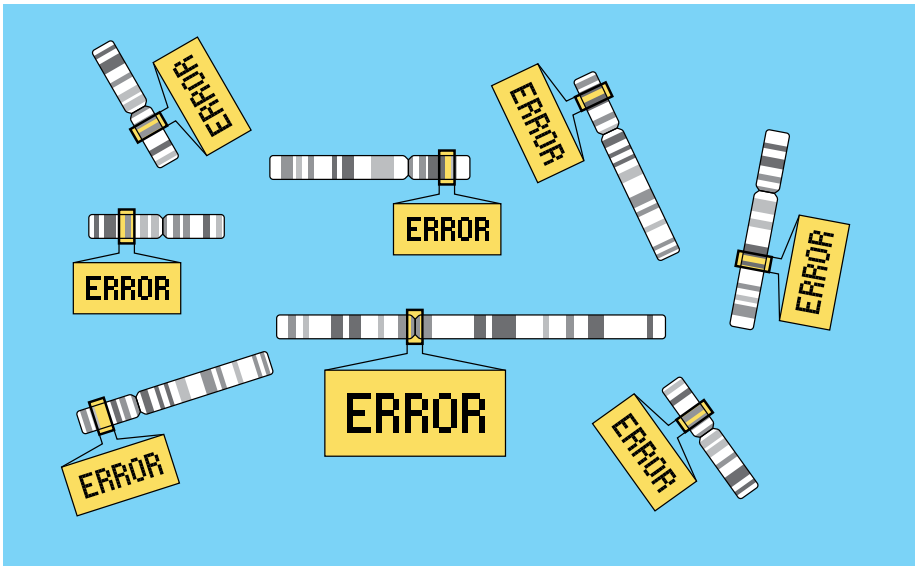
Long road to long-read assembly

Genome assembly projects get a boost from high-accuracy long-read sequencing.

Vivien Marx

When the assembling gets tough, the tough get assembling. Ever since the first sequencing technology came on the scene, it's been a tough computational jigsaw puzzle to assemble human, animal, microbial or plant genomes from DNA sequence reads[1]. Many reference genomes have imperfections such as misassemblies and gaps. The human reference genome GRCh38 has hundreds of gaps and is missing around 150 megabases of sequence.

Now that genomic data generation and analysis are faster, cheaper and more accurate, researchers can expect high-quality, haplotype-resolved genome sequences that run telomere to telomere, note Eric Green, who directs the National Institutes of Health's Human Genome Research Institute (NIH NHGRI) and colleagues in their 'strategic vision[2].' Reference sequences can also increasingly reflect human variation and diversity on a global scale. A group of NIH-funded researchers in the Telomere-to-Telomere (T2T) consortium have taken on the human genome's gnarly bits to build a completely contiguous reference. "I'm a bit of a perfectionist," says NHGRI senior investigator Adam Phillippy, who co-leads the consortium with Karen Miga, a researcher at University of California Santa Cruz. What they started has grown into a large endeavor. "Closing windows can sometimes be seen as a sport for nerds," says Pavel Pevzner, a University of California San Diego computational biologist, referring to gap-closing in genomics. Much biology can be learned by resolving sequence such as the highly repetitive centromeric regions, he says. "We've never had them before to look at," says Phillippy. The T2T Consortium is joining up with another NIH initiative, the Human Pangenome Reference Consortium (HPRC). The plan is to sequence 350 human genomes to represent alleles in people of various ancestries, says Miga. The HPRC effort leverages existing data, such as the 1,000 Genomes Project, which ran from 2006 and 2015 and led to a catalog of human variation. The two groups have been collaborating informally, so now it's a "natural coming together," says Miga. International partners are joining.



Imperfections such as misassemblies and gaps are found in reference genome sequences used around the world. But that is changing. E. Dewalt/T. Phillips, Springer Nature.

No gaps

Closing gaps and completing genomes can make it easier to parse differences between genomes. And the methods can be directed not only at human genomes but other organisms too, such as microbes, or vertebrates in the Vertebrate Genomes Project.

Using a cell line called CHM13hTERT, the T2T team has published the sequence of chromosomes X and 8. At a recent T2T meeting, says Miga, it was exciting to discuss the near-completed entire CHM13 genome. "At that point we had only five gaps remaining," she says. They hope to close these gaps of ribosomal DNA loci and expect to release the complete sequence in early 2021. CHM13 is derived from uterine growths that usually need to be surgically removed. The growths can form when a potential father's sperm enters an egg that lacks a nucleus. In the egg, the sperm's haploid genome is duplicated. Urvashi Surti at Magee-Womens Hospital in Pittsburgh developed the cell line. "The assembly problem is greatly simplified when we just have to assemble one genome versus two genomes," says Phillippy. The team is using CHM13 sequence to fill in gaps in the reference genome GRCh38. They plan to look at aspects such as epigenetics, but heed the caveat that CHM13 might differ from diploid human cells. Their next ambition is a telomere-to-telomere assembled diploid human genome. Going from haploid to diploid may not sound hard, but it's much more complex, he says. "Doing a T2T diploid genome is an unsolved problem to date."

Busy centromere

Both haploid and diploid genomes can be represented as graphs with nodes and edges. Assembly is about finding a path through the graph. With a haploid genome, "I just need to find one path through the graph rather than two paths," says Phillippy.

"Diploid is definitely the next big algorithmic problem," says Pevzner. "It will be interesting to see how quickly we can arrive at near-perfect diploid assemblies." The new genome assembler hifiasm[3] from the lab of Heng Li at the Dana-Farber Cancer Institute shows how one can tackle challenging regions such as centromeres, he says. Miga has long studied these

Karen Miga, UCSC, and Adam Phillippy, NIH NHGRI, co-lead the Telomere-to-Telomere consortium. Credit: P. Driscoll; E. Del Aguila III, NIH NHGRI

regions, often called DNA satellites[4]. In the X chromosome, the centromere has a fundamental repeat unit 171 base pairs long, she says, and its tandem repeats are a combination of 12 of these 171-bp repeats—a stretch around 2 kb long—that is repeated over and over again. In 2001, when the human genome sequence was published, the gaps were no secret, says Miga. Areas such as DNA satellites were left out because, for example, they couldn't be easily cloned. As she and Phillippy began collaborating, they realized that new sequencing technology—ultra-long reads from Oxford Nanopore Technology (ONT) and HiFi reads from Pacific Biosciences combined with Illumina short-read technology could help them reach their goal. Satellite DNA got its name from distinctive bands found upon cesium chloride density gradient centrifugation, says Miga; the bands "turned out to be this tandem repeated DNA." What is challenging about repeats, in both haploid and diploid genomes, is figuring out where they belong, she says. It's called "the blue sky puzzle." Even when this jigsaw puzzle is assembled, the research community does not yet have tools to check the assembly's accuracy in diploid genomes, she says. That's a task that people like Arang Rhie, who is wrapping up a postdoctoral fellowship at NHGRI, are working on.

The centromere makes assembly hard because of its repeats and so-called higher order repeats, which are "repeats on steroids," says Pevzner. He and others use his lab's algorithm centroFlye[5] to assemble the centromere sequence. Algorithmically, he says, centroFlye does the assembly by looking for microscopic clouds in the blue sky. In the millions of assembly steps, he says, correct decisions are needed early on. Missteps can preclude centromere assembly. Once the region is assembled, scientists can explore centromere biology and compare organisms. It will be exciting, he says, to "look at these newly discovered territories."

Being long

"Just over a million bases," is the longest read Phillippy has handled. It was from the CHM13 cell line, as was Rhie's longest read, which she says was around 1.3 million bases. Hardip Patel from the National Centre for Indigenous Genomics at Australian National University has handled a 1.8-megabase read from the X chromosome and a 700-kilobase one obtained during sequencing of the bearded dragon's genome. During graduate training, Miga used 'old school' approaches such as pulsed field gel electrophoresis and worked with short reads by today's standards, but she hears of colleagues around the world generating reads hundreds of kilobases long. Colloquially, she says, researchers call these reads "whales." When Pacific Biosciences and ONT launched long-read technology, with reads longer than 10 kilobases, the technology was quite error-prone. But this has changed dramatically, says Phillippy.

Pacific Biosciences developed Continuous Long Reads (CLR), a sequencing mode in which the instrument does one long pass over a DNA molecule. On a good day, says Phillippy, the error rate for a CLR read has been around 10%. In 2019, the company introduced circular consensus sequencing built from multiple passes over a DNA molecule, so-called HiFi reads. "The instrument spits out this consensus read that can be 99.9% accurate for that single molecule," he says. Five years ago, error rates of 30% with ONT were not unheard of, he says. "They have made tremendous improvements, primarily with their base-calling algorithms," by means such as applying neural net architectures that had been used mainly for speech processing. "They can get well above 90% accuracy now." HiFi reads, says Pevzner, have an error rate of "a couple errors per thousand nucleotides." ONT's ultra-long sequencing is less accurate but also less costly. This landscape is dynamically changing. Biologists still mainly use short-read technology, he says, but the future for complete assemblies is about long reads.

Mapping biases can trip up efforts to assess the assembly quality, says Arang Rhie.

Credit: E. Del Aguila III, NHGRI

Assemblers

A number of assemblers have been used for long, error-prone reads, such as Falcon, miniasm, Flye, Hinge, Canu, wtdbg2, Shasta and Wengan. When HiFi reads emerged, the list of applicable assembly tools shrank to mainly HiCanu[6] and hifiasm[3], says Pevzner. He and his group have just developed one called jumboDB[7]. In Heng Li's view, PacBio's IPA is another HiFi read-optimized assembler. Although noisy read assemblers can be applied to HiFi reads, he says they don't take advantage of the high base-accuracy and won't match HiCanu and hifiasm.

The assemblers HiCanu and hifiasm use string-overlap graphs to represent genomes, encode information for algorithmic analysis and show a reference and alternative paths along a DNA sequence, says Pevzner. The alternative paths represent variation at different loci. Graph genomes make it easier to resolve haplotypes. With string graphs, the nodes are reads and overlapping reads make up the edges, he says.

In his lab, he uses the de Bruijn assembly approach, which splits reads up into *k*-mers — strings of sequence with length *k*. "De Bruijn is little a bit counterintuitive," he says. Reads are converted to *k*-mer strings. Each *k*-mer is a node in the graph, and "edges represent consecutive strings of length *k* that are present in reads," he says. De Bruijn graphs are "the algorithmic engine" in assemblers such as SPAdes, Flye and wtdbg2, but they are not designed for making graphs with large *k*-mers. Memory and computing time become prohibitive. Overall, it remains to be seen which graph approach, the overlap/string approach or the de Bruijn approach, will be most efficient for long-read assembly, he says.

Accurate long-read technology is helping people to work toward haplotype-resolved assembly, says Li who is also part of the T2T consortium. "Not many realize that hifiasm/HiCanu assemblies we produce today are of much higher quality than the assemblies we could get a year ago," he says. "The difference is already night and day." The assembly field is targeting telomere-to-telomere assembly of diploid samples and is set on making techniques more accessible to the larger community, he says. "Then there are polyploid genomes and metagenomes, which are even harder to assemble," he says. "These will keep us busy in the next five years at least."

His assembler wtdbg2 is comparable to other assemblers in terms of accuracy, but it's faster, "mainly due to better engineering," says Li. In his view, Shasta is probably the fastest assembler for nanopore reads, "though it consumes a lot more memory."

A major problem with wtdbg2, he says, is that it tends to collapse similar segmental duplications or repeats into one copy. "This results in an apparently smaller genome," he says; the assembly is smaller than the 'true' genome. It's a common issue with noisy read assemblers, says Li, but it's more severe in wtdbg2 and Shasta. Hifiasm and HiCanu "don't have this problem, which makes them better assemblers."

Hifiasm is faster than wtdbg2 because accurate reads can simplify most algorithms, Li says. Hifiasm can be used for haplotype-resolved assembly, but it's applicable only to PacBio's HiFi reads. He reckons the tool can be adapted for ultra-long ONT reads, "but this requires a lot of engineering work."

A computational method she co-developed, called trio-binning, can be used for haplotyping, says Rhie. Using parental *k*-mers as markers, the tool can pull out reads with more markers from one parent. "But there's always this small fraction of reads that gets misassigned," she says. In her view, hifiasm is likely to prove good for partitioning haplotypes.

Even with more accurate reads, assemblies still need to be assessed for quality. For this task, she has co-developed Merqury[8]. The software can show what different assemblers are getting right and getting wrong, says Phillippy. Rhie says that parental *k*-mers can be used to validate haplotype phasing and that Merqury generates assembly assessment metrics with *k*-mers and does not use a reference. "Mapping biases" can trip up efforts to assess the quality of assemblies. For example, when evaluating an Asian assembled genome against the human reference, "any sort of Asian-specific variation would be called as an error," she says. Especially in the genome's repetitive regions, such bias can be more pronounced. "The *k*-mer-based approach sort of lifts away those mapping-based biases," she says.

A pangenome

Highly accurate telomere-to-telomere assemblies give a better understanding of human diversity and adaptations to local environments, says Patel, who is an external collaborator on the HPRC. When the data reveal such signatures, it "provides an avenue to understand how flexible the genome is," he says. Many software tools are emerging to find genomic nuances in assemblies.

Much of the world's population lives in the Global South and should be included more in genomics projects, says Hardip Patel.

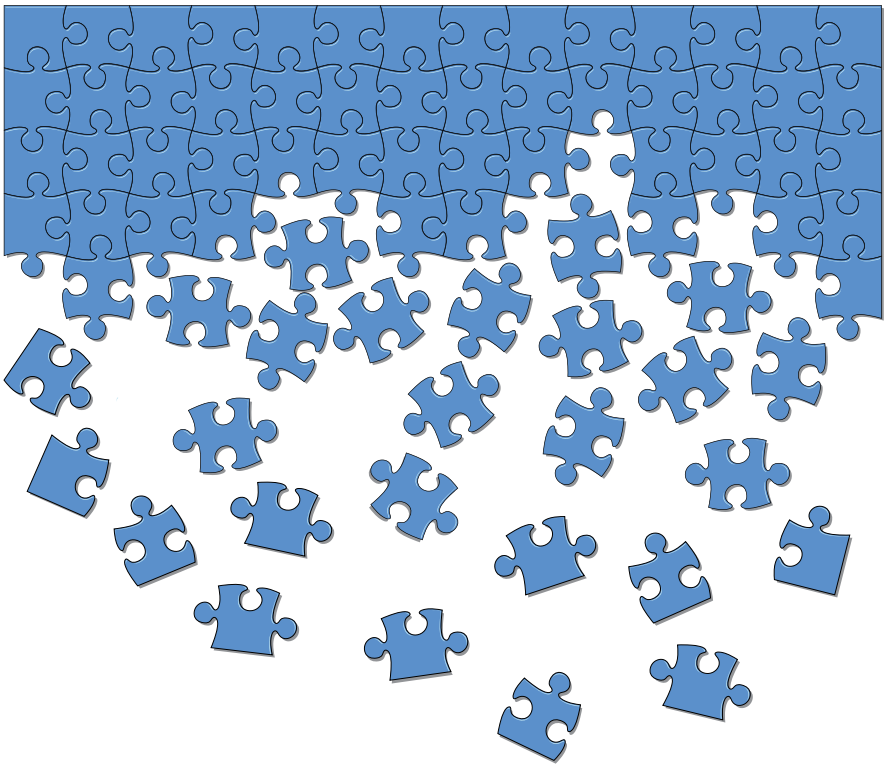
In the Human Pangenome Reference Consortium, scientists are sequencing DNA from hundreds of people so that alleles from people of many ancestries are found in reference sequences. Credit: A. DaSilva/Getty

The pangenome can enable analyses across ancestries[9]. The pangenome project draws on population-focused projects that have cataloged variants, such as the Genome Aggregation Database or Human Heredity and Health in Africa (H3Africa), a platform of genomic resources for exploring heath and disease across Africa. Much of the world's population lives in the Global South and should be included more in genomics projects, says Patel. Such data may be publicly available, but may require permission. Sometimes permission has not been granted, "which is ok, we need to respect the consent process," he says. Individuals might wish to share their data but seek to control how it's used. Communication and trust can help to overcome skepticism, he says. Project organizers must be open about intended data use and storage. "We as scientists and we as humans have to respect an individual's interests, a community's interests, a group of people's interests in how they share their information," he says. "Trust is the key element." Australia and New Zealand have guidelines on data sovereignty for such contexts. "Our indigenous leaders are showing us the right path," he says.

When pangenome data are appropriately collected and their complexity is assessed, he and others in the bioinformatics community will develop tools to analyze it with graph-based approaches. Multiple high-quality reference genomes can be represented as pangenome graphs. Researchers can define distinct variants by finding the most appropriate path through the graphs, he says. Haplotype-resolved genomes are important, he says, because they are accurate representations of how the genome exists in the cell. A representation that is not haplotype-resolved "will have its own set of errors." This pangenome effort is the next big challenge, and, like climate change, "these are the challenges of the twenty-first century," he says. "We just need to get on with the job."

The T2T and Human Pangenome consortia are joining up, "with the hope we'll make it routine in the coming years to do hundreds of human genomes T2T," says Phillippy. "We still have a lot of work to do on the methods side to make that happen."

The pangenome field could move in two directions, says Li. A pangenome will offer a way to encode complex variations, including those in clinically important genes, a task that would fail with most existing methods, he says. "With complex variations annotated, we can more systematically study their evolution and functional impacts," he says. An approach from his lab called minigraph is a way to start working on the next generation of graph-based tools, "but there are quite a few unresolved problems."



Assembling the centromere with its repeats and higher order repeats is a bit like assembling a puzzle of only a blue sky. Credit: DigitalVision Vectors/Getty

A minigraph, says Li, "collapses orthologous sequences if they don't have differences longer than 100 bp." A pangenome also helps with accurate genotyping of structural variations using short reads, he says. His collaborators at other institutions have developed the tools Giraffe and PanGenie, which he calls "promising," and "the combination of the two will have an even bigger impact," says Li. "The biggest question on pangenome is the acceptance of the community. I don't have a clear answer for now."

Completing metagenomes

As an aquatic geomicrobiologist at Friedrich Schiller University, Kirsten Küsel, along with her lab, including postdoctoral fellow Will Overholt, studies microbial diversity and interaction in environments such as groundwater. Among their projects is one on the teeming microbes in sampling wells in Germany's Hainich Critical Zone Exploratory, one of several such sites around the world[10]. "We found that including Oxford Nanopore long reads greatly improved the quality of microbial genomes we were able to recover," says Overholt. They were able to recover more genomes, which "reflected a greater diversity of groundwater microorganisms."

The team performs metagenomic analysis on samples to, for example, find an aquifer's dominant metabolic pathways. But it's been hard to link detected pathways to specific microorganisms. Combining Illumina-based short-reads with ONT's long reads helped them. For metagenomic assembly, they chose metaSPAdes and metaFlye and used Illumina short reads to 'polish' the ONT reads. This approach more than doubled the discovered number of bacterial and archaeal metagenome-assembled genomes, and the data had greater phylogenetic diversity. The quality metrics were more favorable than with Illumina-based reads alone. Using the long reads by themselves produced fewer genomes, says Overholt. They obtained some reads in large chunks, each hundreds of thousands of base pairs long. Having such contiguous information, he says, minimizes and in some cases avoids problems such as those from contaminating genes or sequences misidentified as belonging to another organism. Long reads are more likely to contain important phylogenetic marker genes that help the scientists place the organism within established phylogenetic relationships. It's also easier "to link your new genome to already produced datasets that used those marker genes."

The team's hybrid approach, with long- and short-read-based metagenomic assembly, stands to improve the group's ability to reconstruct genomes from environmental samples and yield improved data for microbial and viral comparative genomics projects, more single markers for phylogenomic studies and a better way to do more complete metabolic reconstructions. With groundwater, says Overholt, it remains difficult to get enough DNA from a sample to sequence on an ONT flow cell. Techniques for maximizing the amount of recovered DNA tend to "shred" long pieces of DNA, which is not good for long-read sequencing. But, he says, the field advances quickly, also developing new ways of working with less input DNA for sequencing. Such progress, along with further optimization of extraction methods, will help with recovering DNA from environmental samples.

Generating long reads from environmental samples is very difficult, says Pevzner, whose group developed the metagenome assemblers metaSPAdes and metaFlye, among others. With such samples, "even 50 kilobases is already a success," he says. Sample prep is challenging given that cells lyse differently under different conditions as a result of differing membranes. "No matter how good your assembler is, you will never get complete genomes with short reads," says Pevzner. To some degree, the assembly will always be fragmented. MetaFlye is the algorithm he and his team developed for assembling complex metagenomic datasets generated with long-read sequencing technology. What makes metagenome assembly hard is that a sample can contain many microbial species with similar sequences. Just as with haplotyping, metaFlye uses a graph-based approach to deconvolve such similarities.

HiFi reads in metagenomics are not yet common. "The first datasets are being generated right now," says Pevzner. In recent work, he and his group performed metagenome assembly with HiFi reads, an approach the team informally calls "complete metagenomics." The scientists found they could increase the coverage of a sheep gut microbiome dataset generated with HiFi reads and assemble complete metagenomes. It had not been possible to, with the push of a button, generate 100 or 200 complete genomes from a metagenomic sample, but HiFi read assembly is making this possible. Highly accurate metagenome assemblies will help people find aspects otherwise missed. For example, naturally occurring antibiotics result from the repetitive biosynthetic gene clusters

encoding enzymes that yield non-ribosomal peptides. Fragmented assemblies have previously prevented their discovery, he says but new long-read tools will emerge to help with such projects.

Welcome, newcomers

The T2T and pangenome work attracts many interested in sequencing technology, long-read-related assembly and other computational biology challenges, says Miga. As a DNA satellite biologist, she is hopeful some new entrants will want to explore repetitive DNA. "What I'm finding is that we have a whole new cohort of people who are showing up, who are junior colleagues," she says. The marriage between technology and a new kind of discovery in genomics brought on by long-read assemblies intrigues them, as do the challenges these assemblies present. She sees curiosity and interest building in people who may be starting their own labs in the next few years. The consortium, she says, welcomes new entrants from all areas of expertise to join existing efforts. "We try to put all of our stuff out as soon as possible so folks can work independently, too, and develop their own questions and research projects." ❐

Vivien Marx ✉
Nature Methods.
✉*e-mail: v.marx@us.nature.com*

Published online: 1 February 2021
https://doi.org/10.1038/s41592-021-01057-y

References
1. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. *Nat. Rev. Genet.* **21**, 597–614 (2020).
2. Green, E. D. et al. *Nature* **586**, 683–692 (2020).
3. Cheng, H. et al. *Nat. Methods* https://doi.org/10.1038/s41592-020-01056-5 (2021).
4. Hayden, K. E. *Chromosome Res.* **20**, 621–633 (2012).
5. Andrey, V., Bzikadze, A. V. & Pevzner, P. A. *Nat. Biotechnol.* **38**, 1309–1316 (2020).
6. Nurk, S. et al. *Genome Res.* **30**, 1291–1305 (2020).
7. Bankevich, A., Bzikadze, A., Kolmogorov, M. & Pevzner, P. A. Preprint at *bioRxiv* https://doi.org/10.1101/2020.12.10.420448 (2020).
8. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. *Genome Biol.* https://doi.org/10.1186/s13059-020-02134-9 (2020).
9. Hudson, M. et al. *Nat. Rev. Genet.* **21**, 377–384 (2020).
10. Overholt, W. A. et al. *Environ. Microbiol.* **22**, 4000–4013 (2020).