genomeweb

New Human Genome Reference Projects Combine Long Nanopore Reads, Other Data Types

Mar 26, 2019 | Julia Karow

NEW YORK (GenomeWeb) – Two ongoing efforts to produce new human genome assemblies are combining ultra-long nanopore reads with other types of sequencing and mapping data to generate gapless or near-gapless assemblies that can serve as reference genomes for future studies.

The Telomere-to-Telomere (T2T) consortium, led by researchers at the National Institutes of Health and the University of California, Santa Cruz, focuses on the first gapless assembly of a human genome, finishing each chromosome from one end to the other. Meanwhile, another effort, involving many of the same members as the consortium, is exploring high-throughput sequencing and assembly methods for a pan-genome reference project that aims to generate hundreds of new human genome assemblies.

At the Advances in Genome Biology and Technology conference in Marco Island, Florida, last month, Adam Phillippy, a bioinformatician in the intramural research program at the National Human Genome Research Institute, presented first results from the T2T consortium — a gapless assembly of the human X chromosome and a human genome assembly, from Oxford Nanopore and PacBio data, that is more continuous than the current GRCh38 human reference genome.

Phillippy and Karen Miga, an assistant research scientist in David Haussler's laboratory at UCSC and a satellite DNA expert, explained in an interview that the T2T consortium, which they jointly coordinate, resulted from the realization that the current human reference genome, which has hundreds of unresolved regions and gaps, leaves out important areas of the genome that play a role in disease and other processes. "All of these regions that we now know are not junk DNA but are functional regions were left off the map" Miga said.

Both researchers had previously participated in the first *de novo* assembly of a human genome using ultra-long reads from Oxford Nanopore's MinIon sequencer — the genome of the well-characterized HapMap sample GM12878. Obtaining these reads, which exceed 100 kb in length, involved the use of a gentle DNA extraction protocol that leaves long DNA molecules largely intact.

That work, led by Nick Loman at the University of Birmingham and Matt Loose at the University of Nottingham, resulted in a publication in *Nature Biotechnology* a year ago. "It was our first flexing of our muscles using long reads," said Miga, noting that the project only generated 10X coverage of the genome.

As part of that project, Phillippy's group had predicted, based on the known number and size of large repeats in the human genome, that at least 30X coverage with ultra-long reads would be required to obtain an assembly that would have the same continuity as the current human reference genome.

Another publication a few months later, led by Miga's group, described the assembly of a human Y

chromosome centromere — a region with highly repetitive DNA — using a combination of ultra-long nanopore reads and short reads. "So it was really natural, now that we have this long-read protocol and a demonstration that you can close one of these big, gnarly tandem repeat gaps, to say 'What can we do with the whole genome?'," she recalled.

From telomere to shining telomere

Following a phone call, Phillippy and Miga decided to embark on a project to generate the first complete, gapless assembly of a human genome, resulting in the T2T consortium. The group also includes Loman and Loose, as well as researchers at NIH, UCSC, the Stowers Institute for Medical Research, Washington University in St. Louis, the Wellcome Sanger Institute, the University of Pittsburgh, Duke University, the Magee-Womens Research Institute, and Arima Genomics. Most recently, Evan Eichler's group at the University of Washington joined the consortium.

The T2T consortium chose to sequence the CHM13hTERT cell line, a hydatidiform mole that is essentially haploid because both sets of chromosomes are paternal. This has advantages because the assembly does not require phasing of two separate haplotypes. "When you have an effectively haploid genome, that problem really goes away," Miga explained. "It simplifies the problem."

In addition, the CHM13 cell line had already been sequenced with Pacific Biosciences' sequencing technology a number of years ago, and Phillippy said that whenever his group published a new assembler, it would benchmark it on data from the CHM13 cell line.

In May of last year, NHGRI's intramural sequencing center acquired an Oxford Nanopore GridIon instrument, learned the ultra-long DNA prep, "and ran that instrument basically all summer long to collect as much data as we possibly could," a total of almost 100 flow cells, Phillippy said. The reason for accumulating so much data, he explained, was that only a fraction of it consisted of very long reads above 100 kb. "So we kind of sequenced it to death, in the hope that we would collect enough of these very longest reads to enable a good assembly," he said.

In the end, they obtained 50X coverage of the genome in nanopore data, with an N50 read length of 70 kb. This included 99 Gb of reads exceeding 50 kb, and 44 Gb of reads longer than 100 kb, with the longest read reaching 1.04 Mb.

In November, they combined the nanopore data with existing 70X coverage of PacBio data for an initial assembly, using the Canu assembler from Phillippy's lab, which he said is designed to be able to take both PacBio and Oxford Nanopore data as combined input. He said the two data types are complementary, correcting each other's errors.

The resulting draft assembly has 657 contigs and an NG50 contig size of 86 Mb, exceeding the continuity of the GRCh38 human reference genome, which has an NG50 contig size of 56 Mb.

Over the following months, the consortium worked on manually closing two remaining gaps in the X chromosome and validating the assembly. "For most of these regions on the X, we were really lucky because they were about 100 to 150 kb [in size] but our read lengths were reaching upwards of 300 kb," Miga said. "So if we could anchor uniquely on either side of these repeats with these long reads, we could derive consensus and manually repair."

The consortium has also generated 10X Genomics/Illumina sequencing data, as well as Bionano Genomics optical maps, Arima Genomics Hi-C data, and CRISPR/Cas9 duplex sequencing data. "We're

using those technologies as complementary validation sources," Phillippy said, as well as to improve the overall base accuracy of the assembly, currently estimated at QV36.

In addition, Eichler's group will contribute ultra-long nanopore reads it has generated for the same cell line to the next release of the dataset. His group also has data from BAC libraries that represent "a solid truth to work into some of these more difficult regions," Miga said. In addition, the consortium is exploring the use of data from flow-sorted chromosomes.

The consortium has made all its data so far available on GitHub and plans to publish its X chromosome assembly as a preprint within the next few months, following the completion of ongoing validation and polishing experiments. In particular, this will involve the addition of 10X Genomics/Illumina data to improve the consensus accuracy.

"Now, we're planning to go through the additional chromosomes kind of one by one and get the rest of them closed like we've done for the X," Phillippy said. Some of this work will be split up by chromosome, while other parts will be tackled by groups with specific expertise, such as in *de novo* assembly or repeat structures.

Not all of it is expected to be straightforward, though. "There are a few regions throughout the genome that we know will be very, very difficult, much more difficult than we encountered with the X chromosome," he said. The plan is to proceed first with chromosomes that are expected to be relatively easy to close, while looking for new methods to finish the hard ones.

Chromosomes 1 and 9, for example, as well as the acrocentric chromosomes, which are rich in repeats and have interchromosomal homology, are expected to be particularly tricky. "Within the consortium, we have a number of groups that have expertise and interest in these especially difficult regions and are going to work on those in parallel," he said. "But I think that will require some new methods on the computational side and also, probably, on the wet side."

Using current methods, 90 percent of the chromosomes should be completed within the next two years, he said, but the remainder will take somewhat longer. However, improvements in nanopore read length and accuracy could speed things up. "If a year from now, we're able to generate, say, 10-megabase reads, that completely changes the picture, and a lot of these very difficult regions would just fall out and be done," he said. "The field is progressing so quickly that it's hard to predict the rate of change, and I think if that rate of change continues as it has been in the past, it's really feasible to get these done within a couple of years."

A pan-genome reference

While the T2T consortium is chugging along, another project seeks to generate a pan-genome reference that captures all human genome variation. This will require sequencing and assembling hundreds of new human genomes from diverse backgrounds. Earlier this year, the NHGRI released a request for proposal for "research and development for genome reference representations", seeking to fund up to four awards with $1.25 million in fiscal year 2020.

"For historical reasons, we've always had these giant databases of SNPs and copy number variants that are really hard to integrate into our current high-quality [GRCh]38 reference map," Miga said. "So we're trying to kind of reboot and rethink how to change the data structure so we can begin to show all of these variants that exist in the human population."

Generating hundreds of new high-quality genome assemblies is expensive and takes a long time with current methods, though, so researchers have begun looking for new combinations of technologies and assembly methods to streamline the process. Miga's UCSC team, in particular, has evaluated the Oxford Nanopore PromethION. In a pilot project, results of which she presented at the AGBT meeting last month, the researchers sequenced the offspring of 10 parent-child trios from the 1,000 Genomes Project. The goal is to sequence all 30 individuals of those trios with various methods.

The advantage of sequencing family trios is that the parental genomes help with phasing. "A lot of our best phasing algorithms and assembly algorithms right now benefit from trio binning," a technique developed in Phillippy's group, Miga said. "This helps us to take these long reads and assign them to either the paternal or maternal chromosome to do phasing, and this is a huge advantage."

One goal of the pilot project is to demonstrate that nanopore data is ready for prime time, Phillippy said. "By doing this initial set of 10 trios, I think we're hoping to show that nanopore can compete with these other technologies and maybe produce better results, but that is yet to be seen," he said. "And out of this analysis of these initial 10 trios, we hope to bring the community some clarity of what the best, most cost-effective way is to generate these high-quality human haplotype references. It's still a bit of an unanswered question."

The first 10 trio genome assemblies will involve different data types, including PacBio data generated by Phillippy's group, 10X Genomics/Illumina data, Bionano Genomics optical maps, and Hi-C chromatin interaction data. "We're going to have the full complement of technologies on these initial 10, which I think will make a great reference set," he said.

Researchers can then compare other assembly strategies against these. "As people develop new methods for quickly assembling the PromethION data alone, or PromethION combined with 10X data, or any of the possible combinations, we can come back and benchmark against this initial set of 10 that we're doing as a pilot that can serve as a great, super-high-quality set of 20 haplotypes," he said.

For its pilot project, which generated 11 genomes in 9 days, using three flow cells per individual, Miga's team used UCSC's PromethION, which currently can run 15 flow cells in parallel and is expected to run up to 48 flow cells in the future. Output differed somewhat between flow cells but reached up to 100 Gb, with an N50 read length of about 50 kb, which compared favorably to the higher N50 read length of the T2T consortium, even though the pilot did not use the ultra-long DNA prep protocol.

Instead, it employed the Nanobind DNA extraction kit from Circulomics, which Miga said made DNA extraction less variable, as well as Circulomics' Short Read Eliminator kit, which removes DNA up to 25 kb in size through selective precipitation. Overall, using more efficient sample preparation "sacrifices a little bit on the read length but gives you a much greater throughput," Phillippy said.

Challenges with getting the PromethION up to full production will include scaling up the base calling, Miga said, and "just running the device," making sure it can sense signals and process data from 48 flow cells in parallel. While running the pilot project, "we were updating constantly" on the PromethION, she recalled, which indicated that the technology is "moving very quickly."

The goal of the pilot was to see whether the PromethION could be "highly reproducible, gives consistent results, and reaches the read lengths that we need. And I would say, across the board, we're starting to see that this is in fact promising," she said. "But we're still pushing our R&D team at UC Santa Cruz to get even longer reads, push the ultra-long read coverage up, [and] push our N50s up."

The same amount of data that the T2T consortium generated on the GridIon over six months, she said, could probably be produced in just a few weeks on the PromethIon now. However, what is still under debate is what proportion of reads needs to be ultra-ultra-long, exceeding 300 kb, in order to get the best assembly.

Based on its results, her team concluded that going forward, it could reasonably generate 10 human genome assemblies in 10 days. The cost for sequencing and computing, including assembly, scaffolding, and polishing, is currently on the order of $10,000 per genome, which does not include phasing yet. "But I still think there is room for improving our pipeline, so that number is constantly going to be changing," she said.

Meanwhile, what types of data the NHGRI pan-genome project will eventually use for the 350 or so human genome assemblies it plans to generate is not clear yet. "There could be a recipe that involves all the data types or there could be a recipe where you use one data type more than others," Miga said.

Filed Under Sequencing | Informatics | North America | NHGRI | NIH | UCSC | Human Genome Project | de novo assembly | genome assembly | nanopore | nanopore sequencing | Oxford Nanopore