genomeweb

# New Methods Enable Phased Human Genome Assemblies Without Parental Data

Dec 08, 2020 | Andrew P. Han

**Premium**

NEW YORK – Two independent teams of researchers have developed methods that allow them to assemble haplotype-resolved human genomes without the use of parental data.

Both groups published their respective papers on Monday in *Nature Biotechnology*. The methods are similar in concept but use different technologies — single-cell strand sequencing (Strand-seq) and the Hi-C assay, respectively, said Tobias Marschall, a professor at Heinrich Heine University Düsseldorf in Germany and senior author of one of the papers.

One team, led by a tandem of postdocs in Marschall's group and Evan Eichler's lab at the University of Washington, achieved their phased *de novo* genome assembly for a person of Puerto Rican heritage by combining Strand-seq with long-read sequencing data from two platforms.

"The assemblies are accurate (quality value > 40) and highly contiguous (contig N50 > 23 Mbp) with low switch error rates (0.17 percent), providing fully phased single-nucleotide variants, indels, and structural variants," the authors wrote.

The method works with long-read data generated on instruments from either Pacific Biosciences or Oxford Nanopore Technologies and is compatible with several haploid assembly algorithms, including Flye, Peregrine, Shasta, wtdbg2, and Canu, as well as different variant callers, such as Google's DeepVariant, FreeBayes, WhatsHap, and Longshot.

The other team, led by researchers at Harvard University in collaboration with several companies, including DNAnexus, used a method called DipAsm to assemble four human genomes with the help of PacBio HiFi sequencing data and data from the Hi-C chromosome conformation assay from Arima Genomics, both at approximately 30-fold coverage.

"This is just great work," said Mikhail Kolmogorov, a bioinformatician at the University of California, San Diego, who has developed genome assembly algorithms for long reads and was not involved with either study. "They have solved the haplotype assembly problem." While both studies resulted in assemblies that showed good contiguity, "the phasing is what makes these papers unique," he said. "They basically claim [the phasing] is complete," he said, meaning the different genetic variants have been assigned to one of the two haplotypes, adding that they appear to have the data to support those claims.

The Strand-seq-based paper claimed a switch error rate of about .17 percent; the Hi-C-based paper compared one of its assemblies to the Genome in a Bottle Consortium's SNP truth set for that genome and disagreed at only .49 percent of heterozygous SNPs.

"This has been percolating for quite some time," Eichler said of the push to phase genomes without parental information. "The reason we want to do that is to increase our sensitivity for sequence resolution of structural variants and put those into the context of all the other variation." Using a reference genome to discover disease-associated and -causing variants subjects the process to biases from the reference.

"Each one of these individuals now is its own human genome project," he said. All that was necessary was to find a way to provide ancillary data to the long reads, he said, to help physically phase the genomes.

Strand-seq, published in 2016 by researchers at the British Columbia Cancer Agency's Terry Fox Laboratory and the European Research Institute for the Biology of Ageing, uses a directional, single-cell sequencing method to develop haplotypes with short-read sequence data.

In addition to being able to help sort contigs to chromosomes and put them in the right order, Strand-seq provides the ability to phase contigs by helping to cluster them into single chromosomes.

DipAsm begins with an unphased Peregrine assembly scaffolded by 3D-DNA or HiRise, calls small variants with DeepVariant, phases them with WhatsHap and HapCUT2, partitions the reads, and assembles each partition independently again with Peregrine. "Grouping contigs into chromosome-long scaffolds is necessary for phasing of entire chromosomes by WhatsHap and HapCUT2," the authors wrote.

Shilpa Garg, first author of the DipAsm study and a research fellow at Harvard and Dana-Farber Cancer Institute, said the Hi-C assay costs about $500 plus the cost of sequencing, which is about $1,000 to $1,200 per sample, in addition to approximately $10,000 per sample for PacBio sequencing. The cost of computing is negligible compared to the cost of sequencing, she said, and assembly is fast. "We have now started to see it can be done in a day," said Garg.

Strand-seq costs approximately $1,200 to $2,500 per sample, "depending on how many single cells per sample one needs," Marschall said, with 40 to 50 cells being sufficient for the workflow described in the paper.

There's a potential tradeoff researchers deciding between the two methods will have to consider. "Conceptually, I think Strand-seq is more robust," Marschall said, "but as of now, the Hi-C technology is more broadly available." Arima and Dovetail Genomics both offer commercial kits, and while very high-quality Hi-C data are needed, they are obtainable, as shown in the DipAsm paper, he noted.

"This method should open the door for producing high-quality phased human genomes needed for personalized [structural variant] discovery in healthy and diseased individuals," the Eichler and Marschall-led team wrote. "Fully phased, reference-free genomes are also the first step in constructing comprehensive human pangenome references that aim to reflect the full range of human genome variation."

Sequencing costs will have to come down before these assemblies could be done routinely in clinical applications, but Marschall suggested they could be immediately valuable to researchers looking at SVs. "If you know what the structural variant sequences look like, you can genotype them from short reads, even if you could not have found them using short reads alone," he said. His lab is already fielding interest in the method from researchers studying rare diseases and cancer.

Eichler said his lab is beginning to study SVs using these kinds of assemblies. "Things are much more structurally variable than we anticipated," he said. "[SVs are] radically different in every human we've looked at." But he suggested that large, repetitive regions of the genome remain impenetrable and assembling chromosomes from telomere to telomere will require further advances in sequencing technology and algorithms.

Kolmogorov said he agreed with these assessments of the applications of the pipelines and their limitations. "Improvements are still needed," he said. "They're still getting fragmented chromosomes. We'll need better algorithms to get the tricky parts [of the genome.]"

He noted that phasing and assembly are related problems, and that the fact the Strand-seq-based method could use any of several assemblers was promising. "It gives users flexibility to choose their own software. If there are any improvements in one of the assemblers, you can easily plug it into the pipeline and get even better reconstruction," he said.

**Filed Under** Informatics Sequencing Genetic Research Next-Generation Discovery Workflows

genome assembly long fragment read technology massively parallel sequencing Next-Generation Sequencing

software strand sequencing structural variant analysis Europe Arima Genomics North America

DNAnexus peer-reviewed publication Google Dana-Farber Harvard University of Washington

haplotype data integration de novo assembly