genomeweb

Long-Read Nanopore Sequencing Boosts Structural Variant Analysis in 1000 Genomes Project Dataset

May 03, 2024 | Huanjia Zhang

Premium

This article has been updated to clarify the start date of the project led by researchers in Germany.

NEW YORK – Two teams of researchers have been harnessing long-read nanopore sequencing for reanalyzing samples from the 1000 Genomes Project in order to build a comprehensive structural variant catalog.

In a preprint published in *MedRxiv* in March, researchers from the 1000 Genomes Project ONT Sequencing Consortium presented data from their pilot analysis of 100 samples from the 1000 Genomes Project. Similarly, another preprint published last month in *BioRxiv* by a team in Germany highlighted results from sequencing more than 1,000 samples from the same project using long reads.

"We know that short-read sequencing only identifies about a third to a half of structural variants compared to long reads," said Danny Miller, a physician-scientist at the University of Washington and the corresponding author of the *MedRxiv* paper. "As a clinician thinking about rare unsolved cases, I was really excited that we could use the structural variant calls from this dataset to filter known disease-associated structural variants."

Led by Miller, who sits on a scientific advisory board for Oxford Nanopore Technologies, and UW professor Evan Eichler, the 1000 Genomes Project ONT Sequencing Consortium was kicked off in 2022 as a collaborative effort to systematically reanalyze all of the approximately 3,300 samples in the 1000 Genomes Project with nanopore long-read sequencing.

So far, the consortium, which also includes collaborators from the New York Genome Center, Cold Spring Harbor Laboratory, Stanford University, and elsewhere, has sequenced about 250 samples, Miller said. A detailed analysis of the first 100 samples, which represent all five superpopulations and 19 subpopulations from the original 1000 Genomes Project dataset, were presented in the current preprint.

According to Miller, a highlight of their study is the high sequencing quality and depth. The samples were sequenced with the Oxford Nanopore PromethION 24 platform in Miller's and Eichler's labs with an average coverage of 37X and N50 read length of 54 kb. The first 250 samples were sequenced with the older Oxford Nanopore R9 flow cells, Miller said, but the team has switched to the newer R10 flow cells for the remaining samples.

"These are really high-coverage long-read samples, and that was very intentional," Miller said, adding that the high sequencing coverage enabled SV detection with high sensitivity using both assembly- and alignment-based methods.

Using the nanopore long-read data, Miller and his team analyzed selected 1000 Genomes Project samples comprehensively for SNVs, indels, and SVs. In the first 100 samples, they identified an average of 24,543 high-confidence SVs per genome.

Additionally, the UW researchers described variations that "would be difficult or impossible to detect or fully resolve using short-read technology," according to the paper, including disease-associated repeat expansions in the genes RFC1 and ATXN10, skewed patterns of X-chromosome inactivation, and differentially methylated regions (DMRs).

Consistent with previous studies, the researchers also noted that the long-read nanopore data demonstrated higher recall and precision than Illumina short-read data for single-nucleotide variants (SNVs) in well-characterized genomic regions and performed "well" for indels, specifically outside of homopolymers.

Independently of the UW-led consortium, researchers from the European Molecular Biology Laboratory (EMBL) and Heinrich Heine University in Germany, along with their collaborators in Vienna and elsewhere, kicked off a similar project in 2021 to reanalyze 1000 Genomes Project samples using nanopore long reads.

"We are interested in new technologies and how they can advance our understanding of genetic variation, including particularly structure variation," said Jan Korbel, head of data science at EMBL and a corresponding author of the *BioRxiv* study.

In contrast to the UW researchers' high-coverage approach, Korbel and collaborators carried out shallower nanopore sequencing to accommodate more samples in their analysis. For their paper, they analyzed 1,019 nanopore long-read genomes, representing 26 human populations from the 1000 Genomes Project database.

On average, these samples, which were sequenced on the PromethION 48 platform using the R9 flow cell, achieved a median coverage of 17X and an N50 read length of 20.3 kb.

By integrating linear and graph-based approaches for SV analysis via pangenome graph augmentation, the European research team

characterized 167,291 sequence-resolved SVs in their samples.

Korbel said the analysis has provided new insights into distinct classes of sequence insertions, such as auto insertions, tandem duplications, and variable number tandem repeats (VNTRs), as well as other types of genomic variation that were hard to detect using short-read technologies.

"For disease studies, what we are providing here is an instrumental resource that can be used for variant prioritization," Korbel said. "We will give the community the chance to discriminate what's normal from what's potentially disease-causing."

"These two papers are showing very good reconstruction of many parts of the genome, either as assemblies or as structural variant calls," said Ewan Birney, deputy director general of EMBL who was not involved in either study. "I think just knowing what the [variant] catalog looks like is a good starting point … there are plenty of things you can do straight away with this data."

Birney, who is a consultant and shareholder of Oxford Nanopore, said an interesting aspect of the preprint by the German team is that by sequencing "quite a lot of samples" with intermediate coverage, they still achieved "pretty good results about structural variation analysis."

Korbel said while lower sequencing depth could lead to lower detection sensitivity, an important take-home message from his team's study is that a hybrid approach of higher and intermediate sequencing coverage "is probably the way to go to achieve completing the whole resource quicker."

"I do agree that [the EMBL preprint] will be a valuable dataset for filtering and prioritizing SVs in challenging unsolved cases," Miller said. "That said, I do not want the community to lose sight of the primary goal of producing high-coverage, long-read data from all 1000 Genomes Project samples on both the ONT and PacBio platforms."

"The combination of longer and higher-quality reads will provide deeper insight into clinically relevant variants in regions difficult to analyze using short-read sequencing and provide a full picture of the spectrum of structural variation across populations," he added.

While both studies demonstrated the utility of long-read sequencing for population-scale studies, researchers still believe the technology has a way to go.

For one, long-read informatics tools are "still jarring," Birney said. "It's a lot better than it was two or three years ago, but it's still not on the same level of maturity as the short-read downstream pipeline."

Echoing Birney's point, Miller said one takeaway from the project is that it is still difficult to do SV merging. "We do really need to find better methods for this," he said. "I hope this dataset will help people start working on that and move that forward."

Additionally, Miller said his team has spent "a lot of time" on extracting high molecular weight DNA and prepping high-quality libraries as the workflows are currently manual. "I think it still shows us the need for automation and improvements in the extraction and library prep stages to really scale this into clinical service," he added.

Similarly, nanopore sequencing "still has a certain error rate, and we're still not at the level yet of an error rate where Illumina sequencing is," Korbel said. "It has not reached the clinical greatness of short-read sequencing. The [clinical] adoption will come, but it's not going to be tomorrow."

Nonetheless, by making their datasets publicly available, both the German and UW researchers are hoping to pave the road for long-read sequencing's wider adoption, eventually benefiting more researchers and patients.

"I hope the combination of this paper [and others] can help push long-read sequencing into the clinical environment, helping to get this technology to patients faster," Miller said.

Filed Under Sequencing | Preprint | University of Washington | nanopore sequencing | Oxford Nanopore | Europe | North America | 1000 Genomes Project | structural variant analysis | methylation | Genetic Research | EMBL | Editor's Pick | Advances in Single-Cell Multiomics