genomeweb

All of Us Researchers Make Case for Long Reads in Population Sequencing

Feb 03, 2023 | Andrew P. Han

Premium

NEW YORK – Researchers associated with the National Institutes of Health's All of Us research project are calling for the increased use of long-read sequencing technologies in this and other population-scale sequencing efforts.

Even using the technologies available three years ago, the authors of a pilot study posted to *BioRxiv* last week recommended the use of long reads on their own, and not necessarily as a supplemental technique to short-read sequencing. The group included Fritz Sedlazeck of Baylor College of Medicine, Evan Eichler of the University of Washington, Michael Schatz of Johns Hopkins University, and Shawn Levy, formerly of the HudsonAlpha Institute for Biotechnology and now CSO at sequencing startup Element Biosciences.

"This study shows the strong value of long reads for simple and complex medically relevant genes and gives clear indications that long reads are on par with, if not better than, the short reads," they wrote. "All of Us and other population-scale projects should investigate the usage of long reads at scale and how to utilize and understand the clinical relevance of the so-obtained novel alleles in the setting of larger short-read cohorts."

The researchers used long-read technologies from Pacific Biosciences (HiFi protocol on Sequel IIe) and Oxford Nanopore Technologies (R9 pore on PromethIon). Using the Genome In a Bottle v4.2.1 small variant benchmark data, they demonstrated F-scores (a combination of precision and recall performance) of 99.87 percent for PacBio and 98.74 percent for ONT, compared to an F-score of 99.47 percent for Illumina. Long reads predictably outperformed Illumina on calling structural variants, as short reads struggle to identify large insertions due to read lengths, they said. The manuscript included work on improving variant calling pipelines for both small variants and structural variants for long reads.

Sedlazeck stressed that the study shows long reads can be broadly useful in clinical genomics. In addition to resolving 386 "highly challenging" genes, long reads showed good coverage of a "general set" of 4,641 medically relevant genes as well as the ACMG 73, the list of genes recommended to be analyzed for secondary findings by the American College of Medical Genetics and Genomics. For the ACMG 73, HiFi's average F-score was 85.24 percent, compared to 93.64 percent for Illumina. The preprint did not provide an F-score for ONT, but Sedlazeck said it was 73.98 percent.

With even better long-read technologies available now, potentially offering epigenetic data for no additional cost, scalability and cost still remain the biggest hurdles to adopting them for clinical genomics.

"These long reads are improving at a significant pace," Sedlazeck said. "They've been perceived to be highly erroneous, but that has changed, massively. They've been perceived to be super costly, but they've dropped to a $1,000, or less" for a human genome at 30X coverage.

Greg Elgar, director of sequencing R&D at Genomics England, agreed that the pace of improvement in long-read sequencing has been tremendous. "There has been such a step change every few months," he said. "By the time you publish anything, you're bound to be six to 12 months out of date." He called the study a "good promotional paper" for long reads, but suggested it was slightly biased in favor of PacBio's HiFi sequencing.

The preprint contains the results of a pilot study started three years ago, when HudsonAlpha won $7 million from the National Center for Advancing Translational Sciences to generate long-read data from more than 6,000 All of Us participants.

Initially, the project intended to find complex SVs or analyze genes otherwise difficult to sequence with short reads. But at some point, the researchers decided to put the three technologies head-to-head. "Illumina is still the work horse for All of Us so it's good to showcase the benefits and the disadvantages compared to long reads," Sedlazeck said.

The authors noted that long reads exhibit "a slight reduction in accuracy across small indels." However, they predicted that new long-read sequencing platforms and analytical methods would close the gap.

Indeed, since the pilot study began, PacBio has announced Revio, a new high-throughput instrument that can produce 15 times more HiFi data than its predecessor.

"Based on our internal benchmarks of Revio data, we see additional increase in indel accuracy [compared to] the Sequel IIe," said Medhat Mahmoud, a postdoctoral fellow at Baylor and first author of the preprint.

Oxford Nanopore's current state-of-the art technology includes an entirely new pore design, a new chemistry, and new basecallers. "The teams will see increased performance as they are switching to the latest R10.4.1 and kit 14 nanopore chemistry on their PromethIons," a company spokesperson said in an email.

Already, other labs participating in the All of Us project are diving into long-read sequencing. In October, Broad Institute Genomics Platform director Stacey Gabriel shared her lab's plans to use lower coverage HiFi sequencing — about 8X to 10X — on 10,000 samples from All of Us.

While many labs associated with All of Us are using long reads, not everyone is pursuing the low-coverage strategy. "We have chosen not to pursue lower coverage techniques because we want to capture rare alleles," Sedlazeck said.

Also, while Levy said back in 2019 that All of Us was hoping to zero in on a sequencing platform for the program by the end of the year, as a whole the group has kicked the can down the road.

Sedlazeck didn't think there was sufficient data in the preprint to declare a winner despite a few areas where PacBio seemed to edge Oxford Nanopore.

"Read length didn't matter as much as accuracy of the reads," he said. "I'd rather go for higher coverage, more accurate reads than longer, longer, and longer reads." PacBio also performed a bit better on indels, he added, "but the difference is getting smaller and smaller."

Sedlazeck's lab is processing thousands of samples with Oxford Nanopore sequencing with 30X to 40X coverage. For hundreds of these samples, they're adding PacBio sequencing at 25X coverage "to achieve high quality assemblies," he noted. The preprint noted that Sedlazeck has "received support from Illumina, PacBio, [and] Oxford Nanopore."

Sedlazeck also didn't think it was worth considering a wholesale takeover by long reads at the expense of short reads.

"Oxford Nanopore was probably further behind three years ago," Elgar said. But its accuracy has improved, and its platform for whole-genome sequencing is now "fully mature." Moreover, the PromethIon is the only instrument other than Illumina's NovaSeq that can offer the scale required by population sequencing programs, he said. PacBio is catching up, though — especially with Revio — "but not in the same ballpark," he said.

Elgar suggested that short read sequencing is "just not equipped" to handle some applications, such as complex rearrangements in cancer genomes.

"I think people will start making choices in the future," he said. "Long reads will not necessarily diminish the short-read market, just expand the range of applications for whole-genome sequencing. The great thing about Oxford Nanopore and PacBio is, you get that for free."

However, "cost is still an issue, and it still drives a lot of decision making," he added.

In the meantime, Sedlazeck said, large-scale, clinically focused long-read sequencing projects will continue to pump out data. The NIH's Center for Alzheimer's and Related Dementias (CARD), for instance, will be sequencing about 4,000 brain samples with Oxford Nanopore sequencers. In addition, the Genomics Research to Elucidate the Genetics of Rare Diseases (GREGoR) consortium is analyzing hundreds of unsolved Mendelian disease cases with long reads. And, of course, there's All of Us.

"There will be more and more long-read data coming out of All of Us this year," he said.

Filed Under Sequencing | Informatics | North America | All of Us | population genetics | Pacific Biosciences | Oxford Nanopore | Illumina | Advances in Clinical Genomics Profiling