genomeweb

Google DeepVariant Boosts Accuracy of Indel Calls in PacBio Long Reads

Sep 22, 2020 | Neil Versel

CHICAGO – Improvements to DeepVariant, Google's deep-learning-based variant caller, have solved a major accuracy problem with indel errors in Pacific Biosciences HiFi long reads, according to data from a recent US Food and Drug Administration PrecisionFDA challenge.

With these changes, Google felt confident enough in DeepVariant to move the software into "production" mode, which it did last week with the release of version 1.0.

In the PrecisionFDA challenge, the second Truth Challenge, which closed in June, DeepVariant achieved 99.9 percent precision and recall in calling single-nucleotide variants and 99.4 percent precision and recall with HiFi long reads, for the best overall performance in the competition. That represents a 75 percent reduction in PacBio indel errors, bringing HiFi indel calls up to a level of accuracy previously only achievable with Illumina datasets.

DeepVariant also produced a hybrid model including both PacBio and Illumina reads for the PrecisionFDA challenge. The hybrid PacBio-Illumina model performed as well as the PacBio-only model for SNVs, while the hybrid model works slightly better for indels.

Lucas Hickey, senior director of strategic marketing at PacBio, said that there will be research presented at the virtual American Society of Human Genetics conference next month highlighting how HiFi reads analyzed with DeepVariant can increase the diagnostic yield in rare genetic disorders.

PacBio introduced HiFi reads with the release of Sequel II sequencers in April 2019. While it had been more accurate in finding SNVs and indels than previous long-read technologies, indel calling still fell far short of short-read technologies until now, mostly due to shortfalls on the bioinformatics side.

"The challenge is that although the PacBio HiFi reads are just as accurate as Illumina short reads, the types of errors that are made by the two platforms are different," said PacBio Principal Scientist Aaron Wenger.

Bioinformatics tools such as the Broad Institute's Genome Analysis Toolkit (GATK) had been developed for Illumina sequences. While legacy variant callers do work with PacBio and Oxford Nanopore sequences, they had not been optimized for non-Illumina instruments, according to Wenger.

DeepVariant also was initially developed for Illumina reads, but because it is part of a newer generation of variant caller that relies on machine learning rather than hand-coded statistical models, it can learn errors and adapt to the environment without the need for human bioinformaticians to create predictive rules. "These deep-learning methods learn from the input data those features that are important for prediction," said Andrew Carroll, project lead for genomics on the artificial intelligence team at Google.

Human understanding of sequencing analysis is still uneven. "We know tandem repeats tend to expand or contract and we know at the ends of Illumina reads that the quality goes down," Carroll said. "Those can be encoded into what humans write."

Other things such as the lack of diversity in reference genomes are harder for human programmers to account for. "Because DeepVariant is learning for itself, it has the ability to capture those features which humans haven't been able to encode as rules into traditional software," Carroll said.

This leads to higher accuracy of calls and better calibration of call confidence, he said.

Earlier iterations of DeepVariant have been available to the open-source community since late 2017, according to Carroll. The most used is version 0.7.17.

Since the initial appearance of version 0.4 to the public nearly three years ago, there have been eight incremental releases, with additions including increased speed, support for PCR-positive data, and the ability to call exomes and large-scale cohorts.

In the PrecisionFDA Truth Challenge competition, the most accurate single-technology submission was DeepVariant with PacBio HiFi data. When the FDA announced that challenge in the spring, the Google team prioritized several improvements to meet those requirements, according to Carroll.

Google actually entered the challenge with the University of California, Santa Cruz, Genomics Institute. PacBio itself did not directly take part, but regularly advised all participants because entrants were asked to run Illumina, PacBio, and Oxford Nanopore FASTQ datasets from three human reference genomes through their informatics pipelines to generate VCF files.

In the competition, PacBio and UC-Santa Cruz worked with Google to train DeepVariant's machine learning. PacBio also updates Google whenever the sequencing instrumentation manufacturer releases new chemistry so that information can be used to update the trained models, Wenger said.

Wenger and colleague William Rowell, a senior bioinformatics scientist, serve as PacBio's subject-matter experts as part of the partnership with Google and other bioinformatics software developers. They can point out why the variant caller might be making a specific type of mistake, for example.

Wenger said that Google made two major changes from earlier versions of DeepVariant that dramatically improved results with PacBio data.

HiFi reads have about 20,000 bases, and genomes generally have variants every 1,000 bases or so, meaning that typical PacBio reads turn up 20 variants. Wenger said that PacBio reads can distinguish whether or not a variant is heterozygotic.

"Because you have this long-range information you can use the single-nucleotide variants spread across the reads to divide them into the acute piles, the maternal and paternally inherited reads," Wenger said. "Then when you're calling variants, you can look only at the maternal pile and now you know that a variant will either be at 0 percent or 100 percent frequency in that pile because there's only one maternal chromosome."

Google thus decided to improve short-range calls with long-range phasing. The company also changed DeepVariant by realigning the layout of pile-ups in its genome browser, the Broad's Integrative Genomics Viewer, because the standard vertical pile-up works better for mismatch errors than for insertion variants, Wenger said.

"There was not a sufficiently rich view of the PacBio data to just mark something as having an insertion or having the exact thing you guessed," Wenger explained. Google thus expanded the reference genome for evaluating insertions and then compared HiFi reads to the modified version.

"That way you have the ability to put all the bases in the reads and get a more complete view of the data to DeepVariant," he said.

According to Wenger, those changed reduced PacBio indel errors by 75 percent in the PrecisionFDA competition. "With these improvements to DeepVariant, now that gap [with Illumina] is closed." Wenger. "We see it as completing the picture there of allowing people, with a single technology, to get best precision and recall."

Wenger said that Google has used the modified truth sets to improve Illumina calls as well. An updated Oxford Nanopore version is in the works as well.

DeepVariant measures accuracy by generating a visual report on metrics including the length of indels the ratio of transitions to transversions, and Mendelian rates. At the population level, the software looks for deviation from the Hardy-Weinberg equilibrium; there are fewer variant-calling artifacts in DeepVariant.

Google researchers including Carroll explained these metrics in a preprint article posted to BioRxiv in February and updated in May, based on a study run with an earlier version of DeepVariant.

"I think that DeepVariant will be an essential component in being able to accurately deliver results, both diagnostic and research, in the long-read technology," Carroll said. "DeepVariant has this unique potential to help these long-read technologies realize the full possibilities of what they're able to achieve."

After the PrecisionFDA competition closed, Google further tested its technology before last week's release. "DeepVariant 1.0 is the manifestation of the things that were submitted for PrecisionFDA," Carroll said.

According to Carroll, each sequencing hardware platform has a "ceiling" in terms of capabilities and accuracy. "The goal of our team is to get as close to the ceiling of what can be achieved with each sequencing technology as possible," he said.

"We do perceive that going forward, there's a fair amount of remaining headroom for the long-read technologies," he added.

But Carroll noted that capabilities will vary between Illumina, PacBio, and Oxford Nanopore, and there will be parts of the genome that remain difficult to map. "When we're developing new features, the first priority is to see if we come up with an idea that will improve variant calling on each of the three technologies," he said.

Carroll said that Google is committed to keeping DeepVariant open-source and plans on continuing to collaborate with entities like PacBio and UC-Santa Cruz. "This [PrecisionFDA challenge] is really a strong validation for the idea of open science and working with people," he said.