genomeweb

Baylor Team Aims for 'Gold Standard' in Structural Variant Calling

Apr 17, 2015| Monica Heger

Premium

NEW YORK (GenomeWeb) – Calling structural variants from next-generation sequencing data has always been a bit tricky. Short-read sequencing technologies do not always span the entire structural variant and there is not yet a plethora of data from longer-read sequencing technologies in human genomes. Complicating matters further, is the fact that there isn't a reference set of structural variants by which to compare calls.

Researchers from Baylor College of Medicine have taken one step toward alleviating the problem. Reporting in*BMC Genomics* last week, they described how they employed multiple sequencing technologies, library preparations, assembly methods, and genome mapping tools to work toward creating a reference diploid genome.

"If we want high-throughput structural variant calling, we need a standard set, and this paper was the result of us trying to figure out how best to call structural variants in a high-throughput manner," co-lead author William Salerno, a bioinformaticist at Baylor's Human Genome Sequencing Center, told GenomeWeb. "It is by no means a gold standard, but it is our first step in that direction."

"This study definitely moves in the direction of a gold standard because it is able to filter out many calls as false positives without having to revert to tedious validation at the bench," said Mark Chaisson, a postdoctoral fellow in Evan Eichler's lab at the University of Washington, who was not involved in the research. Chaisson has previously published on work using Pacific Biosciences' technology to close gaps in the human reference genome.

Salerno said that the study came about because the Baylor team works with a lot of whole-genome sequence data, and "we found it tricky to compare [structural variants] without having a truth set."

In the study the researchers analyzed multiple data types from one human genome, HS1011, a genome that the Baylor team has studied extensively via whole-genome and exome sequencing, as well as array CGH.

The team designed an infrastructure called Parliament, which can incorporate multiple data types to call structural variants. They used previously collected data, including Illumina HiSeq whole-genome sequence from both a paired-end 100 bp library as well as a long-insert library; single-end fragment sequencing on the SOLiD instrument; and array data from Agilent and NimbleGen. The new data included a 4.2 million probe aCGH assay, 10x coverage with Pacific Biosciences' RS II, 2x coverage with an Illumina Nextera long-insert library, and 51x coverage with BioNano's Irys platform.

In total, the data represent around 300 billion sequenced bases, or 90x coverage, and 7.3 million aCGH probes covering the HS1011 genome.

The group also used a variety structural variant calling methods, including commonly used ones like BreakDancer, as well as a recently developed method that relies on local assembly by Spiral Genetics. To call structural variants from the PacBio data, they used PBHoney, a method the Baylor group developed that uses local error and tail mapping.

The methods were then all integrated into Parliament. The initial discovery step in Parliament identified 47,706 events ranging in size from 100 bp to 1 mb. In order to pick out true structural variants from mapping artifacts, the team performed a local hybrid assembly with the HiSeq and PacBio data and were able to narrow down the structural variants to 9,777, which included 5,044 deletions, 4,463 insertions, and 270 inversions.

Using a variety of technologies proved helpful to assessing structural variation. "All of the technologies have something to offer and bring to the table," Adam English, co-lead author and lead bioinformatics programmer at Baylor, told GenomeWeb.

However, the long-read data from PacBio was especially important. "Highly accurate short-read data with Illumina resolves base pairs, but it gets mixed up and confounded by repeats, and that's where you need the long reads," English said.

One of the "big takeaways" from the study was "how much more we're able to see, in terms of structural variation, with these long reads," Salerno added. Simply "increasing coverage with short reads will not be sufficient to get at these structural variants. We're going to need other types of long-range information."

The data from BioNano's Irys system was also helpful for some of the very large variants, those around 50 kb, said English. The Irys mapping system is not as high resolution as NGS, he added, but it is high accuracy, so a call from Irys is "likely to be large and real," he said.

Chaisson agreed that the study illustrated the importance of long reads for calling structural variants. "Longer reads of any type will be superior for detecting structural variation," he said. Reads that are "shorter than the actual structural variant have to be inferred through computation as opposed to directly assaying it."

To illustrate the impact of long-read data, the team developed an Illumina-only workflow to call structural variants and perform local assembly and compared it to a PacBio-Illumina hybrid method and an Illumina method for calling SVs combined with a hybrid method for performing local assembly, showing that "the addition of long-read data can more than triple the number of SVs detectable in a personal genome," the authors wrote. The Illumina-only method identified 3,082 SVs, while the PacBio-Illumina hybrid method identified 9,193 SVs.

The team was also able to estimate what percentage of individuals' genomes are likely to contain structural variation, calculating that between 1.8 percent and 4.5 percent of a genome is likely to contain a structural variant.

Salerno said that the next steps include incorporating the most recent technologies. For instance, since completing this study, PacBio has released an updated chemistry that has longer reads averaging over 10 kb and higher throughput.

In addition, he said, the group also tested Spiral Genetics' anchored assembly method, and since doing the study, it has also been refined and they plan to make more use of that in the future. English added that he has also used the results of this study to make improvements to PBHoney and Parliament. And, said English, in the next round they plan to work with Hi-C sequencing data, "a completely different type of long-range information" that was developed to detect chromatin interactions.

One aspect of PBHoney is that it is not specific to PacBio reads and can work with any type of long-read sequence data, Salerno said, and they would like to figure out how to incorporate all those different types of long-read data into Parliament to improve on it.

Salerno said that the team would like to keep incorporating new sequencing technology data to improve on the genome and work toward developing a gold standard of SV calls. In actuality, the field will ultimately need multiple reference sets of structural variation specific to different populations, Salerno acknowledged. But, he said, a main hurdle is to first figure out how to develop that initial gold standard and then the same practices can be applied to other references.

A secondary goal is to figure out the best practices for calling structural variants from whole-genome sequencing, he said. Realizing that the comprehensive, exhaustive methods that they used in the current study is not feasible for every sequenced genome, Salerno said he would like to figure out using different combinations of data and different coverages at what point saturation is reached, "so we can have a sense of what are the best practices for finding SVs."

Then, when researchers approach the sequencing center with a sample, a budget, and an idea of the types of information they want to gain, the team will be able to determine the appropriate sequencing to do to "get the most bang for your buck," Salerno said. "People have different questions, so we want to be able to customize the response."