Team Uses Single-molecule Sequencing to Close Gaps, Chart Complexity in Human Reference Genome

Nov 11, 2014 | Andrea Anderson

Premium

NEW YORK (GenomeWeb) – A team from the US and Italy that included representatives from Pacific Biosciences has used single-molecule, real-time sequencing on a haploid sample to close gaps in the human reference genome and get a more complete view of the structural and sequence variation present in it.

In an effort to sort out the sequence and structural complexity found in the human genome, the researchers did SMRT sequencing on DNA from a sample of haploid hydatidiform mole, a rare growth that can form during early stages of pregnancy when a nucleus-free egg becomes fertilized.

Because all of the variants present in a haploid genome are inherently homozygous, it is far more straightforward to call variants in such samples than in diploid human tissues, first author Mark Chaisson told *In Sequence*, noting that the coverage at each allele is also essentially doubled when dealing with a haploid sample.

"While we're going about sequencing just a few genomes at a time, it makes sense in the beginning to try to track down these haploid samples for assessing genetic diversity," said Chaisson, a former PacBio employee and current post-doctoral researcher in senior author Evan Eichler's University of Washington genome sciences lab.

As they reported in *Nature* yesterday, he and his colleagues were able to map nearly 94 percent of the SMRT reads from that sample onto version GRCh37 of the human reference genome, closing or winnowing down more than 55 percent of the gaps present in the reference when they began.

The majority of the sequences needed to fill such gaps — some 78 percent — were composed of short tandem repeat runs, though the team also saw examples of sequence that had been misaligned in the human reference genome.

Past studies done with microarrays have delineated large structural variant patterns in the human genome and high-throughput sequencing has offered a look at single nucleotide variants, Chaisson explained. But much of the structural variation in between has been more difficult to detect.

In the hopes of filling in some of those gaps, he and his team took advantage of a newly generated genome sequence for a haploid hydatidiform mole sample called CHM1, which was sequenced using SMRT reads long enough to span some of the tricky-to-decipher parts of the human genome.

Last fall, PacBio announced that it was releasing long-read data from the CHM1 sample to the research community.

For the current analysis, Chaisson, Eichler, and colleagues had access to SMRT reads representing 41-fold sequence coverage of that haploid sample, generated with the PacBio RSII instrument using P5C3 scaffolding sequencing chemistry.

In an effort to profile small structural variants in the human genome that are between 50 and tens of thousands of kilobases, the team started by stitching the new long-read human genome sequences into the reference with the help of software tools such as Celera and Quiver, which were used to create local assemblies, and a modified form of the PacBio software BLASR.

"I was able to make modifications of BLASR that made it possible to retain all the sequence information, even with sequences that we partially mapped to the genome," explained Chaisson, who was involved in data processing pipeline development during his time at PacBio.

That made it possible to use all of the information from these partially mapped reads when feeding the data into the latest version of the long-read assembler Celera without performing additional error correction steps or circular consensus sequencing.

"Previous methods [for assembling PacBio reads] do some sort of error correction, either with high coverage PacBio sequences or with Illumina sequences," Chaisson said.

"The newest release of the Celera assembler is able to go in and do sensitive overlap detection of the long read collection without external error correction and then come up with a fairly good assembly," he added.

To further improve local assemblies spanning 20,000 to 60,000 base regions of the genome assembly, the researchers then applied the PacBio consensus software Quiver, which uses the full range of information present in PacBio and creates a consensus by evaluating multiple reads sequenced from the same starting position.

The analytical approach works particularly well when dealing with haploid samples, Chaisson said, though additional tweaks are needed to routinely apply the software to reads generated from diploid samples.

"It's not quite ready yet for different samples," he said. "That's something were working on fixing right now."

Using these analytical approaches, together with an iterative mapping and assembly method, the team closed 50 of the 164 known interstitial gaps in euchromatic regions of the genome and reduced the size of another 40 of those gaps, adding more than a million bases of new sequence in the process.

When it began delving into the newly resolved sequences, the team saw a slew of short tandem repeat expansions, including some simple tandem repeat sequences that were repeated hundreds or even thousands of times.

Such expansions would have been "virtually impossible to resolve using other technologies," Chaisson said, because they require reads that start outside the STR sequence and carry on until after the expansion concludes.

"The largest of the STR expansions that we saw in the mole are likely to have just been misassembled in the genome because the original Sanger sequences were not long enough to spam the full STR sequence," he noted.

When they mapped sequences from the ENCODE project back to all of the newly closed gaps and larger STR sequences, Chaisson noted, the researchers saw evidence for a large amount of functional sequences such as promoters and repressors in the gap closures.

In its search for new structural information in the human genome, meanwhile, the team uncovered tens of thousands of insertions and deletions larger than 50 bases apiece, including almost 7,000 indels that impacted protein-coding genes.

The researchers were also able to identify more complex structural variations and inversions, leading to increased resolution at a region already suspected of being misaligned based on bacterial artificial chromosome sequencing studies, for instance.

That large stretch of chromosome 10 sequence had a different orientation in the CHM1 mole sample than in the reference genome — misalignments that could only be fully resolved using both long reads and additional BAC-based sequencing.

"The next step is to improve the algorithms to be able to call variation using diploid samples, which will allow us to re-sequence patients and other populations," Chaisson said.

The researchers also plan to expand this type of sequencing to additional haploid and diploid samples to get a clearer sense of how some of the newly described structural variations differ from one individual to the next.

In particular, Chaisson said, it will be important to profile the short tandem repeat runs in as many individuals as possible — something that may be helped along in the near term by the development of new long-range capture sequencing technologies.

"The sheer number of these and the cost to sequence the entire genome is still a bit high," he said.

Provided sufficient time and resources are available, he noted that the same sort of single-molecule sequencing used to fill in the human reference genome could also be applied to find structural variants and improve on genome assemblies produced with data from large-scale sequencing efforts such as the 1000 Genomes Project.

Given sufficient resources it should also be possible to generate single-molecule long reads as a means of improving the genomes of other organisms, including non-human primates.

"If you look at the assemblies of a lot of the primate genomes, they're both highly fragmented and sort of humanized, where a lot of the contigs are ordered and oriented according to what the human sequence is," Chaisson said. "So there's a huge bias in our knowledge of primate sequences in that they're viewed in light of the human genome."