genomeweb

# Researchers Combine Single-Molecule Sequencing, Mapping Technologies to Analyze Human Genome

Jun 29, 2015 | Ciara Curtin

NEW YORK (GenomeWeb) – Combining single-molecule sequencing and single-molecule physical genome mapping is giving researchers a better view of the structure of the human genome.

Ali Bashir, an assistant professor of genetics and genomics at the Icahn School of Medicine at Mount Sinai, and his colleagues analyzed a diploid human genome using both Single-Molecule Real-Time (SMRT) sequencing data generated on the Pacific Biosciences platform and single-molecule genome maps from BioNano Genomics' Irys system, and used the resulting data to build a hybrid assembly.

Separately, the two approaches improve genome assemblies from short-read sequencing data as they produce long fragment sizes. BioNano's approach uses linearized DNA molecules up to megabases in length, combined with nicking enzymes, to generate physical maps with sequence motifs, called genome maps. PacBio's SMRT sequencing can generate reads that are tens of kilobases in length.

By contrast, short-read assemblies are more fragmented, and though there are approaches to minimize that, the sequence contigs still can't resolve structures in repetitive regions and there would be missing physical mapping sites, Bashir noted.

Together, long reads and physical maps as the researchers reported in Nature Methods today, yield an even better assembly, improving the contiguity of the initial sequence assembly by nearly 30-fold and that of the initial genome map by 8-fold.

"I don't think there's really any other set of technologies that could've been integrated in this way," Bashir told GenomeWeb.

In addition, the researchers reported being better able to resolve complex forms of structural variation, something they noted may be critical for understanding disease.

For their hybrid assembly approach, Bashir and his colleagues generated two de novo assemblies of the NA12878 diploid genome, a sample that's often used in benchmarking studies.

They sequenced NA12878 using 851 Pre P5-C3 and 162 P5-C3 SMRT cells. This gave the investigators 24x and 22x coverage of the genome, and aligned mean read lengths of 2,425 and 4,891 base pairs, respectively.

They assembled error-corrected PacBio reads using the Celera assembler as well as with the PacBio-developed Falcon assembler. These assemblies gave them an N50 of roughly a megabase.

At the same time, the researchers analyzed de novo BioNano nanochannel array-generated genome maps with 80x coverage and mean spans of 277.9 kilobases.

Rather than then mapping the genome maps back to the human reference genome, the investigators used the sequence contigs they'd generated as a sort of reference for the genome maps, and vice versa.

In addition to anchoring the sequence contigs to the scaffold, they performed the reciprocal anchoring, Bashir said, noting that every once in a while their sequence contigs could bridge multiple genome maps. Further, because the two approaches tend to break at different spots — repeats versus fragile sites where nick sites are proximally located to one another — the sequence contigs could be scaffolded to the physical maps, and the physical maps could be scaffolded to the contigs to give more contiguous maps.

"You are mapping these genome maps and these contigs symmetrically to each other," Bashir said.

With this reciprocal scaffold in place, the researchers could lay the sequence contigs out across that space to generate their final scaffold.

Hybrid scaffolding combining the Celera-assembled reads and BioNano maps yielded 377 hybrid scaffolds with a scaffold N50 of 13.6 Mb, the researchers reported. A second round of scaffolding based on the Falcon assembly resulted in hybrid scaffolds with a scaffold N50 of 31.3 Mb.

As compared to the reference genome, Bashir and his colleagues generated an assembly with a higher contig N50, scaffold N50, and scaffold accuracy measurement. However, their sequence identity was lower, though they said that could be due the detection of true variants or alternative alleles.

The combination of direct observation and long read length provides a better resolution to detect structural variants, Bashir noted, and he and his colleagues reported observing large structural variations in the hybrid scaffold.

With long contigs, rather than having to infer breakpoints from paired-end data, structural variants can be directly observed, as molecules span the entire structural variant. At the same time, because the physical maps were so long, the researchers could see structural variations that couldn't otherwise be seen with another technology.

For instance, the researchers would be able to see a 20-kilobase to 50-kilobase stretch of virtually identical repeats at the breakpoint of a structural variant. Even long PacBio reads, Bashir noted, wouldn't be able to span that, while the physical maps do.

In addition, the researchers uncovered tandem repeat expansions that can be tens to hundreds of kilobases in size. Such repeats, they said, have been underrepresented in the reference genome.

They were able to detect complex rearrangements, such as inversions located with insertions, deletions, and duplication, as well as inversions with overlapping boundaries.

This approach, Bashir noted, could be applied to other complex genomes, like plant and fungal genomes.

Bashir said that much of the interest thus far has been in humans, especially to apply their hybrid approach to trio studies to get a complete, accurate full-length diploid analysis. "I think that's where's there's a ton of interest," he said.

Bashir estimated that the combined method undertaken by the researchers would cost between $20,000 and $30,000, but he suspects that it will fall to about $10,000 in a year or so. As that occurs, he said that people would begin to assemble genomes without using the reference.

"That is the future for these sorts of complex variation studies," he said.