MinIon Early Access Users Demonstrate De Novo and Hybrid Assemblies, SNP Calling, Outbreak Analysis

September 16, 2014

MinIon Early Access Users Demonstrate De Novo and Hybrid Assemblies, SNP Calling, Outbreak Analysis

By Julia Karow

NEW YORK (GenomeWeb) – Early access users of Oxford Nanopore Technologies' MinIon sequencer and their collaborators have used nanopore data they generated to assemble a yeast genome *de novo* from error-corrected reads, to generate a hybrid *de novo* assembly of a bacterial genome, to call SNPs in bacteria, and to quickly analyze bacterial outbreak samples.

The results, some of which were presented during an online conference last week, demonstrate that the pre-commercial MinIon produces useful data in the hands of several research groups. This is in contrast to a peer-reviewed opinion article by another user earlier this month who found the data that he generated to be of little use.

Michael Schatz, a researcher at Cold Spring Harbor Laboratory and a participant in the MinIon Access Program since June, has used MinIon data, in conjunction with Illumina reads, to assemble the 12.5-megabase *S. cerevisiae* genome *de novo* and is planning to sequence larger genomes in the future.

"The instruments are very exciting, especially their low cost and small size, although they have required major retooling of our analysis software to use them effectively," Schatz told *In Sequence*.

He and his colleagues developed a new error correction and assembly pipeline called NanoCorr that uses Illumina data to correct errors in the nanopore reads. The pipeline is similar in design to the algorithms his team developed for error-correcting Pacific Biosciences reads but "it has a completely new method to overcome the new challenges of Oxford Nanopore reads," he said.

Using NanoCorr, Schatz's group was able to assemble 100x coverage data from a yeast strain into contigs with an N50 size of 363 kilobases, about tenfold better than an assembly of Illumina data alone.

As the MinIon, the library protocols, and the analysis software improves, he and his colleagues "are optimistic to scale this up to larger genomes and further improved assemblies," he said. Schatz plans to discuss his results in more detail next week at the Genome Informatics meeting in Cambridge, UK.

Researchers from St. Petersburg Academic University of the Russian Academy of Sciences employed a different strategy for a *de novo* hybrid assembly of the *E. coli* genome, using nanopore reads in conjunction with Illumina data. Nick Loman, a researcher at the University of Birmingham, presented the results at an online conference called Balti and Bioinformatics last week, which was webcast.

The Russian group, led by Anton Korobeynikov, incorporated *E. coli* MinIon data generated by Loman's team into an assembly of Illumina data from the same strain in order to resolve repeats, close gaps, and create scaffolds. They employed the SPAdes genome assembler, in particular a read-to-graph aligner that was originally developed for PacBio data.

Using Illumina data alone, the assembly had 92 contigs with an N50 of 133 kilobases, and 98 percent of the genome was covered. Adding the nanopore data decreased the number of contigs to four, one of them covering the entire 4.6-megabase *E. coli* genome. The researchers found six misassemblies that appear to be associated with transposon repeat units, which Loman said are likely due to subtle differences between the *E. coli* batches his lab and Illumina sequenced.

It is unclear at this point how much nanopore data will be needed to generate good hybrid assemblies, he said, and the SPAdes reads-to-graph aligner can be further improved.

Loman's own group has performed approximately 20 successful runs on the MinIon so far – he said his lab has five of the instruments at the moment.

Last week, he posted data from four runs of the *E. coli* K-12 MG1655 sub-strain, including analysis scripts, to the GigaScience database, which is publicly available. Three runs used the R7 chemistry, which recently replaced the initial R6 chemistry, and one used the more recent R7.3 chemistry.

The first run, which he said was not the lab's best, generated about 44,000 or 272 megabases of forward reads; 23,000 or 125 megabases of reverse reads; 20,000 or 131 megabases of normal 2D reads, which incorporate both forward and reverse reads and are of better quality; and 1,600 or 10 megabases of "full 2D" reads, which represent the best 2D reads where the reverse strand was slowed down.

The two other R7 runs included an overnight incubation step and resulted in much lower overall yields but a greater percentage of full 2D reads. The mean 2D read length for all three runs was 6 to 7 kilobases, and some reads were as long as 45 kilobases.

The R7.3 run had a slightly smaller overall yield than the first R7 run but a higher proportion of full 2D reads.

Oxford Nanopore started holding a weekly competition for the greatest yield per run among

MAP participants this summer, Loman said, which his group has won three times so far. His team also currently holds the overall record among users, with 583 megabases from a single run, about half of Oxford Nanopore's internal record of about 1 gigabase.

Using only 2D reads, Loman's team was able to cover the *E. coli* genome with around 10x coverage, missing only 15 bases in the genome, which he said might actually not be present in the sample they sequenced.

They were also able to call SNPs from an alignment of MinIon reads, he said, showing an example of a *Salmonella* dataset. "The nanopore alignment looks a bit messier" than an alignment of Illumina data, he said, "but you convince yourself that there is a SNP there by setting the appropriate allele frequency."

While it may be difficult to call heterozygous SNPs, he said, "if you set an allele frequency around 50 to 60 percent, you can generally partition true and false SNPs."

"Importantly, the data are quite usable," Loman said. "The experiments we can do now with, say, 80 percent accuracy are very similar to what we can do with 85 percent accuracy."

One application Loman's team has explored is the quick analysis of samples from a disease outbreak.

As a proof of concept, they sequenced an outbreak strain and a non-outbreak strain from a recent *Salmonella* outbreak in the Birmingham area on the MinIon using the R6 and the R7 chemistry. They mapped chunks of data generated during certain time periods of each run to a metagenomics reference database that contains genes which define particular species or genera.

The researchers found that after 10 minutes of generating data, they were able to determine the species of the sample and detect chromosomally encoded phages. After 30 minutes, they could determine the serotype, and after 100 minutes, they were able to distinguish the outbreak strain from non-outbreak isolates.

Loman plans to reveal further details about the project in a journal publication.

Julia Karow tracks trends in next-generation sequencing for research and clinical applications for GenomeWeb's *In Sequence* and *Clinical Sequencing News*. E-mail Julia Karow or follow her GenomeWeb Twitter accounts at @InSequence and @ClinSeqNews.

Related Stories

- Oxford Nanopore Presents Details on New High-throughput Sequencer, Improvements to MinIon
  September 16, 2014 / In Sequence
- Oxford Nanopore Opens Registration for MinIon Early Access; 'Robust and Diverse' Response so Far
  December 3, 2013 / In Sequence
- Accelrys Works with Oxford Nanopore to Prep Analysis Tools for Upcoming GridIon,

MinIon
February 24, 2012 / BioInform

- MinIon Review by Early Access User Suggests Technology Not Ready for Routine Use Yet
  September 9, 2014 / In Sequence
- Oxford Nanopore Shows off MinIon at ASHG; 'Hundreds' to be Shipped for Early-Access Program
  October 29, 2013 / In Sequence