Blend Long and Short Reads for Better mRNA Isoform Analysis

Friday, October 10, 2014

Print

Jeffrey M. Perkel

To mangle the old saw, there's more than one way to sequence a transcriptome. For some researchers, the goal is counting transcripts to assess expression levels—a sequencing-based alternative to DNA microarrays. Others are interested in transcript architecture. Eukaryotic genes are often alternatively spliced, and the choice to include or exclude particular exons can have profound biological consequences.

The former application is simpler and likely more widespread—it dovetails nicely with the characteristics of Illumina's popular sequencing platform, which offers short snippets of RNA sequence, but billions upon billions of them at a time. For researchers in the latter camp, bioinformatics tools and long-read technologies increasingly are the name of the game.

The long and short of it

The average mammalian transcript is between 1,000 and 3,000 bases and exists in multiple forms, says Jonas Korlach, chief scientific officer at Pacific Biosciences. A gene with five exons, for instance, could conceivably occur in such configurations as 12345, 1245, 1345, 245 and so on. Working out the structure and abundance of those different forms shouldn't be difficult—just sequence each RNA molecule from end to end and tally the numbers. The problem is, no current sequencing technology can do that by itself.

Illumina's HiSeq v4 reagents produce some 4 billion high accuracy reads per run—more than enough to deeply sequence a transcriptome and find rare variants. Yet each paired-end read measures just 2 x 125 bases long, making it difficult to determine which pieces go with which—or if, say, two exons always or never co-occur. Should those reads include repeated elements, they can be difficult to unambiguously map to the genome, to figure out whence they came. (Illumina declined to be interviewed for this article.)

"The way we figure out transcriptomes now is kind of crazy, if you think about it," said Michael Snyder, professor of genetics at Stanford University and director of Stanford Center for Genomics and Personalized Medicine, in a recent Mendelspod interview. "We take RNA. We blow it up into little fragments, and then we try to assemble them back together to understand what the transcription looked like in the first place. That's a horrible way to do this."

Pacific Biosciences' single-molecule sequencing PacBio RS II produces reads averaging 8,500 bases apiece, easily long enough to cover most transcripts, and a new longer-read chemistry is anticipated later this month, Korlach says. But the RS II yields only 50,000 to 80,000 reads per SMRT Cell, too few to comprehensively read every transcript

multiple times (though a user can queue up as many as 16 cells at a time with walkaway automation, Korlach notes). And the platform has an error rate of about 10% per base. (Alternative long read technologies include Illumina's Moleculo technology and Oxford Nanopore Technologies' nanopore sequencing.)

The hybrid approach

For many researchers, the solution is to combine the two approaches. In a recent study in *PNAS* [1], for instance, Snyder's team applied that hybrid strategy to sequence the lymphoblastoid transcriptomes of a child and its parents, using the long PacBio reads (approximately 711,000 per sample) as scaffolds on which to hang the shorter Illumina data (100 million reads per sample). At the same time, the Illumina reads provided an error check on the PacBio nucleotide calls. "It lets you do local cleanup of the actual bases, so you get all the calls correct," Snyder says.

Jason Underwood, director of technology development at the Northwest Genomics Center at the University of Washington (and formerly an R&D scientist at Pacific Biosciences) also applied that strategy in a recent analysis of the H1 human embryonic stem cell line transcriptome [2]. His team's "hybrid sequencing" approach identified "hundreds of novel genes/long noncoding RNAs (lncRNAs) and thousands of novel isoforms of known genes expressed" in H1 cells, the authors wrote—this despite the fact that H1 cells had already been extensively scrutinized as part of the ENCODE project.

That said, Underwood doesn't always apply short reads for error correction; when he analyzed the chicken-transcriptome architecture, he used only long-read technology [3]. "In that case, you just error-correct with the reference if you are looking to predict ORFs," he says.

According to Korlach, PacBio technology enables researchers to capture the entirety of transcript diversity. In its method, called Iso-Seq, users synthesize cDNAs, size-fractionate to create libraries of different lengths and then circularize and sequence the results. The company's SMRT Analysis software then polishes the reads by clustering transcripts with identical structure, thereby minimizing sequencing errors. (A complementary strategy is "circular consensus sequencing" (CCS), in which cDNAs are circularized and sequenced repeatedly to produce a more accurate average read.)

Given the relatively low read count of PacBio, some researchers couple the technology with methods to select one or a handful of genes for analysis. In one recent study, for instance, a team led by Peter Scheiffele at the University of Basel used PacBio's method to specifically sequence 370,000 neurexin transcripts from adult mouse brain, identifying and quantifying nearly 1,400 unique isoforms of that family [4].

Analytical tools

To make sense of those data, Scheiffele's team used an alignment program called GMAP, which Underwood also uses. Other bioinformatics tools for transcript structure analysis include Cufflinks, SpliceMap and SigFuge (see here for more options). Developed in the lab of D. Neil Hayes, associate professor of medicine at the Lineberger Comprehensive Cancer Center at the University of North Carolina at Chapel Hill, SigFuge is an "unsupervised" tool for identifying interesting structural variants—in Hayes' case, cancer biomarkers—across thousands of patient samples. "If [a variant] is important, it should be recurrent," he explains. And with SigFuge, "we can detect recurrent structural variants in RNA structure."

But how much sequence do you need to find them? There's no simple answer, says Hayes. "The more you have, in general, the better. But the more you sequence, the more expensive the study is." He shoots for about 60 million Illumina read-pairs per tumor transcriptome, he says.

As a general rule, Underwood recommends users interested in whole-transcriptome analysis by long-read sequencing shoot for at least a million reads per sample. “You have maybe five or six orders of magnitude between the lowest and highest expressed RNAs,” he says. So a million reads should be enough to get a couple passes at even the rarest transcripts. That would require about 20 SMRT cells (or, at 8 cells per run, 2.5 runs) on a PacBio instrument. “So it’s a chunk of change, but not prohibitive.”

References

[1] Tilgner, H, et al., “Defining a personal, allele-specific, and single-molecule long-read transcriptome,” *Proc Natl Acad Sci USA*, 111:9869-74, 2014. [PubMed ID: 24961374]

[2] Au, KF, et al., “Characterization of the human ESC transcriptome by hybrid sequencing,” *Proc Natl Acad Sci USA*, 110:E4821–30, published online November 26, 2013, doi: 10.1073/pnas.1320101110. [PubMed ID: 24282307]

[3] Thomas, S, et al., “Long-read sequencing of chicken transcripts and identification of new transcript isoforms,” *PLoS ONE*, 9:e94650, 2014. [PubMed ID: 24736250]

[4] Schreiner, D, et al., “Targeted combinatorial alternative splicing generates brain region-specific repertoires of neurexins,” *Neuron*, in press, 2014. [DOI: 10.1016/j.neuron.2014.09.011]

Image: Pacific Biosciences

Update: This article was updated 11 Oct. to add new information on Hayes' research.

Next Generation Sequencing / Whole Genome Sequencing »

Expression Analysis »