GOING LONG AND DIRECT

Sarah Webb explores how RNA sequencing innovations help scientists tackle the complexity of the transcriptome.

Biologist Chris Vollmers of the University of California, Santa Cruz relies on RNA sequencing to study the rich and unique biology of B cells, where each cell lineage recombines its genome to produce specific antibodies. As these cells modulate their response to immune challenges, they can further mutate and produce diverse transcripts, even within the same individuals. To study these cells, short-read RNA sequencing, the current gold standard, fell short.

After reverse transcription, Vollmers' team produced full-length cDNA that they then chopped into short sequences of roughly 200 bases in length. "You have this beautiful full-length isoform information that you then lose completely because of the technology in the library preparation," Vollmers said. The resulting data was cumbersome to assemble, limiting his ability to study transcriptional isoforms and the coding regions for antibodies.

So in a recent study (1), his team tried a long-read strategy. They took advantage of the full-length cDNA and used Oxford Nanopore's MinION sequencer to examine whole sequences in a single read. They soon realized that there was far more going on in B cells than they could see with short-read data. "You think you have a homogenous population, but really, cells make totally different isoforms, which may do totally different things," he said.

Vollmer's work is just one example of researchers fueling sequencing innovations for mining the rich biology of RNA. Scientists are also using sequencing to find and study chemical modifications to transcripts. Routine direct RNA sequencing is

on the horizon, and that technique could boost the emerging field of epitranscriptomics, allowing researchers to explore another layer of chemical and regulatory information in these molecules.

Delving into RNA

RNA occupies a sweeping middle ground in molecular and cellular biology. Its cousin, DNA, is far easier to study and sequence; it is more chemically stable and folds into more consistent secondary and tertiary structures. But much of the information about a cell's activity is embedded in its RNA sequences—in the isoforms transcribed, in splicing patterns, and in the function of noncoding sequences. Researchers are just scratching the surface of the biology in the transcriptome and its implications for both normal cellular function and disease pathology.

One critical challenge with today's RNA sequencing is trying to balance experimental goals with technology limitations. Illumina's systems remain the gold standard for nucleic acid sequencing. According to Vollmers, this sequencing strategy is fantastic for many applications, including quantifying gene expression and finding individual splice junctions, and it has unparalleled accuracy because of the volume of reads. But read length is a limiting factor.

Short-read sequencing can be a bit like trying to reconstruct paragraphs of text word by word—a daunting process. By contrast, long-read sequencing is like recording an entire

paragraph instead, said Hagen Tilgner of Weill Cornell Medical College. With that context, it's far easier to make overall sense of a transcript.

For almost a decade Tilgner has been interested in the frequency and importance of transcriptional variations such as start sites, exon inclusion patterns, and polyadenylation. As a postdoctoral fellow in Michael Snyder's laboratory at Stanford University, he has used long-read sequencing strategies to study these distinct sites and how they're combined. If such combinations were random, short-read sequencing would pick up the probability of each variable site separately, and researchers could figure out the combination frequency based on products of these probabilities. But when combination patterns are non-random—like words nested grammatically within a sentence—“You really need to record all variable sites within the same molecule,” Tilgner said.

To look across RNA with the depth and statistical power that they needed, Tilgner and Snyder couldn't use newer long-read sequencing approaches (see sidebar), so they developed synthetic long-read sequencing (SLR-RNA-seq), an approach that adds barcodes to assist with long-read assembly. By combining that approach with microdroplets, they successfully sequenced up to 10–100 million RNA molecules from 1 ng without amplification. Their studies (2,3) showed that non-random combinations are widespread, much like the patterns in human speech. “It's really the fundamental justification for long-read sequencing when it comes to RNA.” Tilgner's lab is now studying how these various non-random combinations are being used in tissues such as the brain and in various cell types.

RNA-seq is also enabling researchers to uncover some of the secrets of long, noncoding RNAs. Once thought to be junk, some of those sequences now appear to take active roles in cells, including as regulators of pluripotency in human embryonic stem cells and related cells (4).

Recently, researchers at the University of California, Davis, Medical Investigation of Neurodevelopmental Disorders (MIND) Institute and Pacific Biosciences used long reads to expand the number of known FMR1 gene isoforms from 16 to 49. Mutations in this gene can cause Fragile X Syndrome and predispose carriers to premature menopause. Women with

Long read pros and cons

For now, choosing an RNA sequencing strategy requires weighing a variety of pluses and minuses, ranging from scientific needs to budget to what's available at your institution. “Short read technology is somewhat of a stopgap. It tells you a lot about your sample, but not everything,” Vollmers said. The perfect RNA-seq experiment should provide quantitative information about isoforms without guesswork, and perfect technology would have hundreds of millions of reads of those full-length transcripts, he added.

The list price for a Pacific Biosciences Sequel system is \$380,000 in the USA. Oxford Nanopore's MinION sequencer, which plugs into a USB port, starts at \$1000, including a starter pack of reagents for sequencing DNA. (Direct RNA sequencing requires separate reagents.)

Read volume is far lower than for shorter reads ranging between half a million to a million, compared with hundreds of millions for Illumina short read sequencing. Error rates are also higher—up to 15% compared with less than 5% for Illumina.

an increased risk of giving birth to a son with Fragile X show different expression levels of some of these isoforms (5).

Mining modifications

In addition to isoforms and non-coding RNAs, researchers are also beginning to realize how much biological information may be encoded in chemical modifications to RNA strands, known as the epitranscriptome (6).

Pacific Biosciences' Single Molecule, Real-Time (SMRT®) technology uses an immobilized DNA sequence, a DNA polymerase, and fluorescently linked nucleotides within tiny microscope-like chambers. As the DNA polymerase incorporates a labeled base, such as A or T above, the identifying fluorophore is released and measured. Credit: Pacific Biosciences.

Tao Pan of the University of Chicago has been interested in these types of RNA modifications for more than two decades, with much of his work focusing on tRNA. Nearly 10 years ago, he teamed up with University of Chicago chemical biologist Chuan He to better understand the epitranscriptome and the most commonly observed modification of mRNA, 6-methyladenosine (m⁶A). In 2011, He's laboratory published their discovery that fat mass and obesity-associated protein (FTO) acts as a type of eraser, lopping off these added methyl groups. That discovery, coupled with work from two other groups that developed antibody-based sequencing strategies to detect m⁶A, spurred the emerging interest in epitranscriptomics.

Finding and distinguishing RNA modifications with sequencing typically involves variations of RNA immunoprecipitation sequencing (RIP-seq), one of which also incorporates UV-crosslinking, known as crosslinking and immunoprecipitation (CLIP-seq). Using antibodies that glom onto m⁶A researchers can identify, isolate, reverse transcribe, and sequence the short sequences surrounding the modification using an Illumina sequencer, said Francois Fuks of the Free University of Brussels in Belgium.

While studying Tet enzymes, which oxidize methylcytosine in DNA, Fuks' team became interested in studying the role of that same modification, hydroxymethylation of cytosine (hMC) in RNA molecules. Using a RIP-seq type strategy, they found that Tet enzymes mediate hMC in 1500 transcripts in *Drosophila* (7). This modification is critical for *Drosophila* development, and mRNA tagged in this way is more likely to be translated. Further work in mammalian cells has linked this modification with an inflammatory response to infection.

In mRNA, epigenetic modifications are variable and relatively rare, estimated at no more than one every few hundred bases. That contrasts with human tRNA, where an average of 13 modifications occur on each molecule, which is less than 100 bases in length. But the field needs to get a better picture of the modification fraction and needs ways to get accurate sequencing data from smaller samples, Pan said.

This work isn't trivial, even for the most commonly observed m⁶A modification where researchers typically need to work with a million cells or more. That's fine for cell culture work, but

not practical for studying cancer tumor biopsies or diseased neurons. "So, another challenge of all of these epitranscriptome sequencing methods for mRNA is, how are we going to reduce the sample size [needed to get meaningful data]?" Pan said.

Less common modifications are even more difficult. Pseudouridine is probably the second most abundant mRNA modification, but it doesn't lend itself to an antibody-based sequencing strategy. According to Pan, a chemical strategy is available, but it's difficult to perform and lacks precision.

Direct to the source

Most of today's RNA sequencing involves reverse transcribing RNA to cDNA and sometimes amplifying sequencing with PCR. Right now, those steps in library preparation can make the difference in whether it's possible to sequence RNA or not. But those procedures can skew the resulting data, noted

Daniel Turner, vice-president of applications at Oxford Nanopore Technologies. "Research transcription is not always as processive you might want," he said. "PCR can make lots of copies but it doesn't amplify all things equally." So as researchers want to answer questions that require quantitative data, they'll increasingly want to move away from reverse transcription and PCR.

First accomplished by now defunct Helicos BioSciences in 2009, directly reading RNA is possible with both Pacific Biosciences and Oxford Nanopore sequencers. Pacific Biosciences published a proof-of-principle study on direct RNA sequencing in 2013 (8) and continues to partner with academic

Pacific Biosciences' Single Molecule, Real-Time (SMRT®) technology uses an immobilized DNA sequence, a DNA polymerase, and fluorescently linked nucleotides within tiny microscope-like chambers. As the DNA polymerase incorporates a labeled base, such as A or T above, the identifying fluorophore is released and measured. Credit: Pacific Biosciences.

researchers to improve this technology.

Oxford Nanopore's MinION sequencer can also sequence RNA (9). Researchers need a different set of reagents and the speed is more than 80% slower—80 bases per second rather than 450 per second with DNA. But other than naturally occurring degradation, there is no fragmentation or bias based on RNA strand length, said Sissel Juul, Oxford Nanopore's Director of Genomic Applications.

Although not yet optimized for identifying chemical modifications to RNA strands, both systems show variations in modified RNA bases as they move through the sequencing process. For example, Oxford Nanopore's direct RNA sequencing picks up variations from modified bases, but the software is not optimized to accurately identify them.

Better data, higher throughput

As long-read strategies improve and direct RNA sequencing strategies become more practical, researchers will be able to dig more deeply into RNA, ensuring that the demand for more tools to delve into RNA will continue to grow.

On a practical level, the availability of long-read sequencing is already reducing the bioinformatics burden of RNA sequencing. With very short-reads, researchers have needed to rely on a variety of isoform prediction tools that come with a host of uncertainty, Montgomery noted. As long-read data is becoming increasingly available, bioinformatics specialists can redirect their efforts and resources from assembling transcriptomes from short-reads to addressing interpretive questions such as the regulation and downstream effects of the various gene isoforms, said Jonas Korlach, Chief Scientific Officer at Pacific Biosciences.

Pacific Biosciences has also been optimizing workflows to improve researchers' abilities to better understand gene expression on polyploid genomes, found in a variety of plants including crops such as wheat and strawberries.

As the quality and availability of RNA-seq improves, Montgomery expects that integrated transcriptome data could become an important component of personal genomics data. Stephen Montgomery of Stanford University is interested in aberrant gene expression patterns in rare genetic disorders, and his team has started to incorporate long-read data to augment their understanding of splicing effects in these genes. With information about isoforms, interactions between loss of function alleles, and gene expression levels, clinicians could have more tools for predicting and even managing patient risks from a known mutation such as BRCA. "One could imagine that if you're making lots of a really bad protein, that's going to be worse for you than if you're only making a little bit of it and the other copy—the normal copy—is compensating".

Although Vollmers and his team are still working on improving the molecular and computational biology in their long-read RNA sequencing work, he already sees the payoff. "Working with this data every day—it's great. It contains so much information," he said.

References

- Byrne, A. et al. 2017. Nature Communications 8:16027. DOI:10.1038/ncomms16027
- Tilgner, H. et al. 2015. Nature Biotechnology 33:736-742
- Tilgner, H. et al. 2017. Genome Research 28:1-12.
- Durruthy-Durruthy, J. et al. 2016. Nature Genetics 48:44. doi: 10.1038/ng.3449
- Tseng, E. et al. 2017. BBA- Gene Regulatory Mechanisms. 1860:1117-1126.
- Liu, N and T. Pan. 2015. Nature Structural & Molecular Biology. 23:98-102.
- Delatte, B. et al. 2016. Science. 351:282-285.
- Vilfan, I.D. et al. 2013. Journal of Nanobiotechnology. 11:8/
- Garalde, D. R. et al. 2018. Nature Methods. doi:10.1038/nmeth.4577

Written by Sarah Webb, PhD.

BioTechniques 64:39-43 (February 2018) doi 10.4155/btn-2017-0105

New Era in Single Cell Genomics

Single Cell RNA Analysis Using Drop-Seq Technology

- Highly Parallel Cell & Mol Bar Code DNA Coated Beads (fig.1); Lead to Single Cell Transcriptomes (STAMPS)
- Capable of High Level RNA Capture, rt PCR & cDNA Library (fig.2); Macosko et al., Cell 2015, 161 (5), 1202-1214, Gierahn et. al., Nature Methods 13 Feb 2017.

Fig. 1

Fig. 2

- A multiple particle type of Bar Code Beads are now available.

Your source of Barcode Beads confidently from ChemGenes

UNPARALLED Expertise in

NUCLEIC ACID Chemistry since 1981

ISO 9001:2008 Certified Company

ChemGenes
CORPORATION

Toll Free: (800) 762-9323

33 Industrial Way, Wilmington, MA 01887 • T: (978) 694-4500 • F: (978) 694-4502

For more information, visit us at www.ChemGenes.com