Full-length isoform sequencing yields a more comprehensive view of gene activity

RNA-seq has revolutionised how scientists can interrogate gene expression. But after years of performing RNA-seq studies with short-read sequencers, many have realised that there is more to be discovered. Comprehensive transcriptome analysis – an essential tool for characterising disease, studying cell lines and measuring drug response – requires a more thorough approach.

By Luke Hickey

New studies have demonstrated that long-read sequencing, which spans full-length isoforms without the need to reassemble fragmented data, can detect novel genes, transcripts and gene fusions even in well-characterised samples. Long-read sequencing has provided an in-depth view of alternative splicing, revealing far more of this mechanism than has previously been observed. It also enables detection of critical elements such as long, non-coding RNAs.

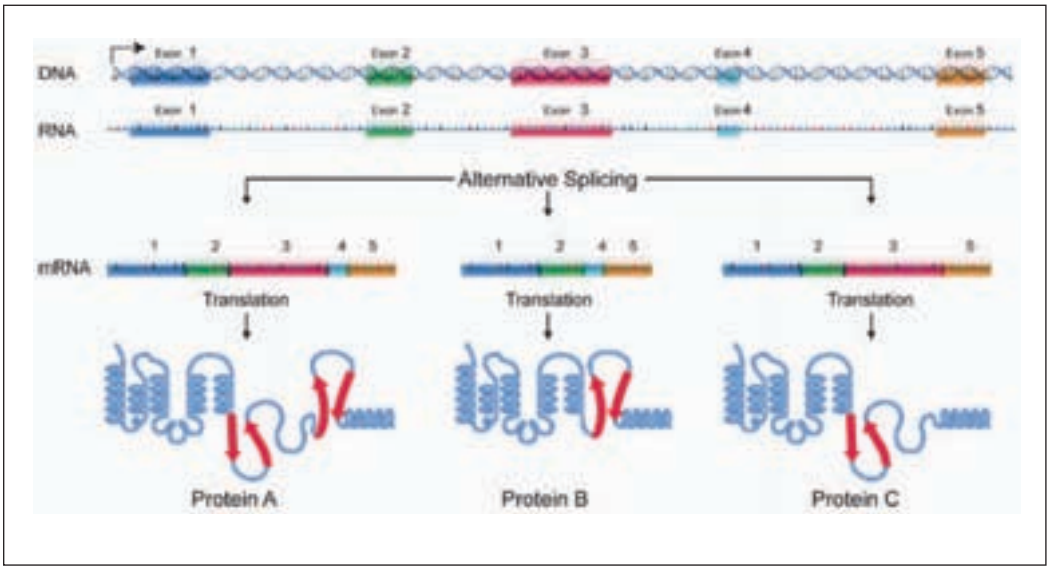
One approach, known as the Iso-Seq method, has been used to examine gene expression patterns for cancer and other diseases, organisms such as fungi that may cause infections, cell lines regularly used in drug discovery and even whole personal transcriptomes. These studies underscore the potential for using long-read sequencing to analyse full-length, protein-coding gene transcription across the drug discovery and development pipeline.

In this article, we look at the addition of long-read data to short-read analysis of RNA and review several studies conducted using the Iso-Seq method to generate new findings. We will also discuss how long-read sequencing can benefit drug discovery.

The rise of RNA-seq

When scientists began deploying next-gen sequencers rather than microarrays for RNA analysis, they were able to produce far more gene expression data. Perhaps even more importantly, they could analyse gene activity across the genome instead of being restricted to known regions of interest.

With these advantages, RNA-seq rapidly became the preferred method for measuring gene expression. First introduced eight years ago, RNA-seq is now a common technique used in genomic labs around the world; a quick PubMed search finds the term cited in thousands of papers to date.

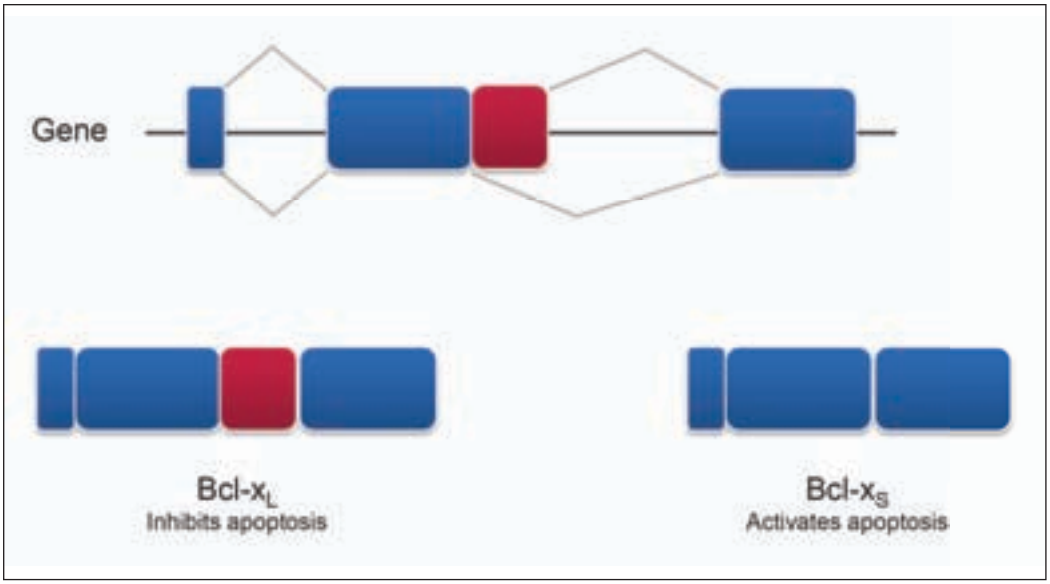


Alternative splicing refers to the process by which a gene is spliced into more than one type of mRNA molecule. These distinct mRNA gene isoforms are then translated into individual proteins which can perform unique and sometimes opposing functions in a biological pathway[1]

But at a certain point, users of the method began to realise that RNA-seq studies were not revealing the whole picture of gene activity. In a review of RNA-seq published in *Nature Methods*, Stanford University scientist Michael Snyder, who was one of the first to describe the method, acknowledged: "The way we do RNA-seq now is... you take the transcriptome, you blow it up into pieces and then you try to figure out how they all go back together again... If you think about it, it's kind of a crazy way to do things"[3].

The problem is really quite simple: reads produced by next-generation sequencing (NGS) instruments, usually no more than a couple of hundred bases long, are too short to span entire gene isoforms. When one tries to piece these snippets of information back into transcripts, accurate mapping is all but impossible. Too often, the assembly does not accurately reconstruct alternately spliced isoforms, misrepresenting the transcripts that exist in the sample.

When scientists began looking into this phenomenon, they found the issue was even more widespread than they thought. The RNA-seq Genome Annotation Assessment Project (RGASP) evaluated several RNA-seq protocols and found that short reads miss exons, limiting the ability to identify transcripts. "For a significant fraction of transcripts not all exons are identified, ranging from 30% in *C. elegans* to greater than 60% in *H.*



The Bcl-x gene is known to regulate apoptotic cell death in numerous cancer cell types. However, the gene encodes two alternatively-spliced mRNA isoforms Bcl-x(L) and Bcl-x(s) with opposite effects on apoptosis and cell death[2]

sapiens," they reported, concluding that scientists should use independent validation when RNA-seq results are "critical to an experimental study"[4].

Enter Iso-Seq

The RGASP group predicted that third-generation, or long-read, sequencing instruments would eventually replace short-read sequencers for analysing genome-wide expression and generating intact transcripts. The most commonly used example of this is the Iso-Seq method, which uses Single Molecule, Real-Time (SMRT) Sequencing technology from PacBio to produce very long reads. Individual reads, averaging 10kb, often capture entire transcripts, including both the 5' end and the 3' end.

Scientists found the Iso-Seq method provides the ability to directly observe full transcripts, offering enormous advantages. The community now can access a complete view of transcript diversity, enabling discoveries in a broad number of research areas.

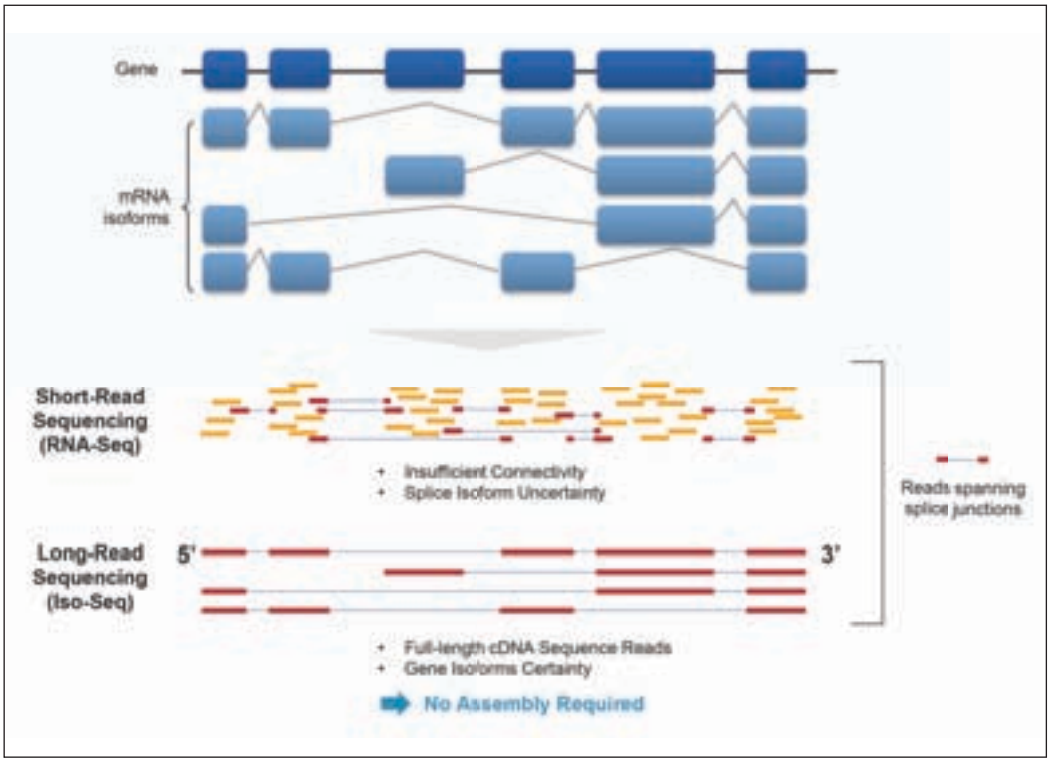
In the following sections, we look at how the Iso-Seq method has made a difference in studies relevant to the drug discovery process.

Disease mechanisms

Scientists at the University of California, Davis, School of Medicine have dramatically advanced the understanding of a range of conditions associated with the gene that causes Fragile X syndrome. Individuals differ in the numbers of a repeat sequence in FMR1; a full-blown case of Fragile X is characterised by more than 200 repeats, while smaller numbers of this repeat can lead to Fragile X-associated tremor/ataxia syndrome or Fragile X-associated primary ovarian insufficiency.

A team led by Flora Tassone used the Iso-Seq method to analyse gene activity related to these disorders, producing the first analysis of alternative splicing and full-length transcripts for the FMR1 gene in people with fewer than 200 repeats[5]. Unlike PCR-based methods, which detect individual splice sites but cannot show which ones are linked together in an RNA molecule, the Iso-Seq method "has allowed us to obtain a transcript map of all of the splice combinations within a single FMR1 transcript," the team reported in a *Journal of Medical Genetics* paper.

They were surprised by some results, including the detection of alternative splicing in an FMR1 exon not previously known to use this mechanism. In more recent work, the scientists analysed gene transcripts in various tissues, discovering even more isoforms along the way and finding a



Standard RNA-Seq methods, using short-read NGS sequencing, lack the read length required to span full length transcripts. This results in uncertainty regarding exon connectivity and alternative splicing events, with ambiguity around gene isoform expression. Long-read RNA sequencing methods (termed Iso-Seq) provide full-length reads across alternatively spliced transcripts, spanning from the 5' start site to 3' poly-A tail. This provides certainty regarding gene isoform expression, and obviates the need for transcript assembly

splice pattern that distinguishes between cases and controls.

Characterising cancer

Cancer research has been an early application for many cutting-edge genomics tools, and scientists followed the same path with the Iso-Seq method. A number of impressive studies have revealed novel information from subject samples and commonly used cancer cell lines.

In Sweden, scientists at Uppsala University applied the Iso-Seq method to samples gathered from individuals with chronic myeloid leukaemia. They focused on mutations in the BCR-ABL1 fusion gene, which is known to be important for determining when patients become resistant to tyrosine kinase inhibitors. According to their BMC Cancer paper, the new method is highly accurate, with a 0% false positive rate[6].

For all six individuals, the Iso-Seq approach detected mutations previously found using Sanger sequencing, and it also detected several low-frequency mutations that Sanger sequencing had missed. In one case, Iso-Seq was able to call a mutation four months earlier than Sanger sequencing could, suggesting that long-read technology may allow scientists to spot evidence for drug resistance much sooner. The team reported that the Iso-Seq method made it possible to distinguish multiple transcript isoforms from individual samples, providing a clearer view of alternative splicing. It also allowed them to differentiate compound mutations from independent ones, which may be useful for understanding drug response.

In a separate project, scientists at the Queensland Institute of Technology and several other institutes interrogated a prostate cancer cell line and discovered a novel fusion transcript between two relaxin genes with high sequence homology[7]. The team turned to the long-read Iso-Seq method because other technologies could not reliably distinguish between the RLN1 and RLN2 genes, the latter of which has been associated with promoting the progression of cancer.

Their work led to the finding that RLN1 is underrepresented in most prostate cancer cell lines; the LNCaP cells they studied are most reflective of the RLN1 expression seen in normal or cancerous tissue. A novel fusion transcript they uncovered incorporates long stretches of both RLN1 and RLN2 and appears to be inversely regulated by androgens.

Finally, research from Cold Spring Harbor Laboratory and other institutes characterised SK-BR-3, a commonly used breast cancer cell line featuring a Her2 amplification. The work revealed tens of thousands of novel isoforms and also helped explain complex fusion transcripts for which there previously had been no DNA evidence.

The scientists used the Iso-Seq method to validate more than a dozen gene fusions in the SK-BR-3 transcriptome and to detect several novel fusion events, including some that involved the Her2 oncogene. A number of these fusions could be seen in RNA but could not be traced back to the underlying DNA. The Iso-Seq study finally explained why: they were incredibly complex, requiring at least two variants to form the transcript. Without the full isoform sequence, they would not have been able to reconstruct the manner in which these fusions had formed.

Disease agents

Fungal infections affect hundreds of millions of people each year. While most infections are just a nuisance, some are far more serious: it is estimated that as many as one million people die from fungal infections each year. That makes fungi an important target for drug development.

A recent study of fungi shows how the Iso-Seq method can help elucidate their biology. While the project did not focus on medically relevant species, its results have clear implications for better understanding these complex organisms and show how the technique will be useful across many applications.

The project was described in *PLoS One* by scientists at the Joint Genome Institute, University of Minnesota and other institutions. By deploying long-read sequencing to the analysis of gene activity in four Agaricomycetes species, they found much more transcript diversity than expected[8]. In addition, they uncovered signs of polycistronic transcription units, which could be important for genetic manipulation of these organisms.

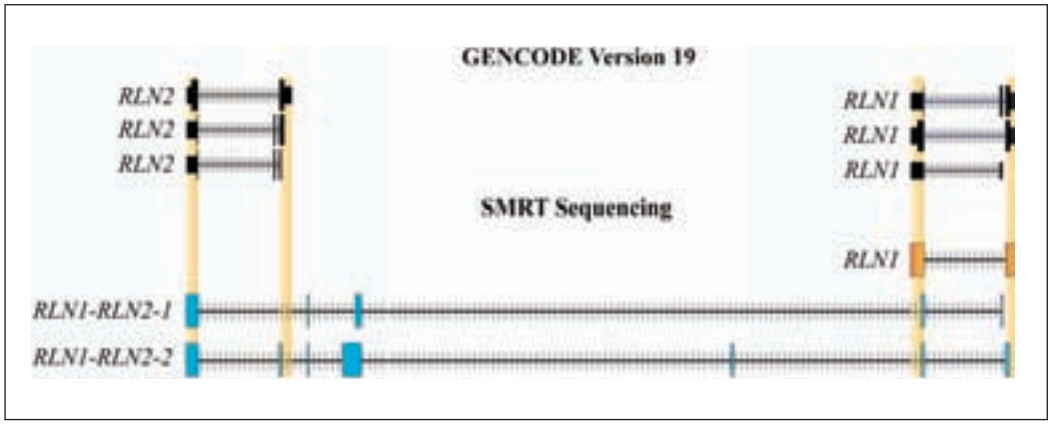
Conventional wisdom suggested that fungi exhibit less alternative splicing than other organisms – as little as 7%, in contrast to more than 95% in humans, the scientists reported. However, they found more than 25% of transcribed loci were alternatively spliced in the four species they examined, and suggest that this number may still underestimate how much splicing occurs since they only sampled two conditions. They also noted that a majority of the transcripts they discovered could not be fully reconstructed when assembled with short-read data.

Whole human transcriptomes

In addition to studying specific diseases or disease-causing organisms, the Iso-Seq method has been used for individual transcriptomes. Whether it is a personal transcriptome that sheds light on an individual's health or a population-specific transcriptome that can inform analyses of other members of that population, these projects are extremely useful for characterising the broad range of gene activity in humans.

One of the earliest efforts came from Michael Snyder's Stanford lab, where a team of scientists generated transcriptome-wide data from RNA representing 20 human organs and tissues. They found that long-read sequencing produced reads covering entire transcripts, yielding much more information than they had seen with other approaches. The scientists identified about 14,000 spliced GENCODE genes and predicted that about a tenth of the alignments represented novel transcripts; many of them included long, non-coding RNAs[9].

In more recent work from the Snyder lab, scientists produced transcriptome analyses for a family trio. As they reported in *PNAS*, they compared Iso-Seq and RNA-seq results for the lymphoblastoid transcriptomes of a child and two parents[10]. With the Iso-Seq method, the team "determined single-nucleotide variants (SNVs) in a *de novo* manner and connected them to RNA haplotypes, including HLA haplotypes, thereby assigning single full-length RNA molecules to their transcribed allele,



A novel gene fusion between the RLN1 and RLN2 genes was identified in prostate cancer cell lines using the Iso-Seq method. Full-length transcript sequencing reads spanning the two genes are aligned in the UCSC Genome Browser to show direct evidence for the fusion event[7]

References

1 NHGRI Online Education Kit: Understanding the Human Genome Project.
2 David, Charles J. II et al. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes & Dev.
3 Nawy, Tal. End-to-end RNA sequencing. Nature Methods.
4 Steijger, Tamara et al. Assessment of transcript reconstruction methods for RNA-seq. Nature Methods.
5 Pretto, Dalyir et al. Differential increases of specific FMR1 mRNA isoforms in premutation carriers. Journal of Medical Genetics.
6 Cavelier, Lucia et al. Clonal distribution of BCR-ABL1 mutations and splice isoforms by single-molecule long-read RNA sequencing. BMC Cancer.
7 Tevz, Gregor et al. Identification of a novel fusion transcript between human relaxin-1 (RLN1) and human relaxin-2 (RLN2) in prostate cancer. Molecular and Cellular Endocrinology.
8 Gordon, Sean et al. Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. PLoS One.
9 Sharon, Donald et al. A single-molecule long-read survey of the human transcriptome. Nature Biotechnology.
10 Tilgner, Hagen et al. Defining a personal, allele-specific, and single-molecule long-read transcriptome. PNAS.
11 Shi, Lingling et al. Long-read sequencing and de novo assembly of a Chinese genome. Nature Communications.
12 Fai Au, Kin et al. Characterization of the human ESC transcriptome by hybrid sequencing. PNAS.

and demonstrated Mendelian inheritance of RNA molecules", according to the paper. They also identified novel isoforms, which were used to produce personalised annotations for these individuals.

In separate work, scientists at Jinan University and other institutions produced a full transcriptome for a Chinese individual, using the Iso-Seq method to uncover a significant amount of gene and alternative splicing data that was not previously included in GENCODE annotations[11].

The team reported generating more than 58,000 isoforms across the transcriptome. Some of the data revealed a great deal of complexity, and follow-up studies validated these findings. The information produced could have tremendous utility in helping interpret transcriptome results for people of Chinese descent, for whom existing genetic databases may not fully represent the most common or important isoforms.

Another project from scientists at Stanford and other organisations focused on an extremely well-characterised human embryonic stem cell line, combining RNA-seq data with Iso-Seq data. The long-read data enabled direct detection of more than 8,000 full-length isoforms with RefSeq annotations, and prediction of several thousand more – including many novel isoforms[12]. In their *PNAS* paper, the team reported that long, non-coding RNAs were particularly likely to be overlooked with short-read data alone. The results "suggest that gene identification, even in well-characterised cell lines and tissues, is far from complete", the scientists wrote, noting that their approach could significantly enhance transcriptome studies.

Together, these results show that the Iso-Seq method routinely turns up biologically relevant isoform information that cannot be found easily through other means.

Drug discovery opportunity

In the drug discovery and development pipeline, there are any number of areas in which gene transcription information can prove quite useful. Whether it is a better understanding of the etiology of a particular disease, homing in on a biomarker that can signal which drug to use or whether a drug is working, or even targeting specific RNA molecules with a therapeutic, having confidence in predicting gene activity is essential. On the most basic level, accurate gene expression data is needed to ensure that cell lines are performing as expected, and that they reflect native biology.

Pharma and biotech scientists frequently use qPCR or microarrays to collect such data, and short-read sequencers have also made their way into these pipelines. But as the studies discussed above show, too often these techniques miss critical information. The only way to establish a comprehensive view of gene activity is to incorporate long-read data, covering whole transcripts and identifying the long, non-coding RNAs that so often escape detection with other methods.

With full-length isoform sequencing, drug discovery scientists now have the capability to explore previously unidentified genes and isoforms, dive into the complex world of alternative splicing and resolve elements that have so far been inexplicable, such as multi-variant gene fusion events. The Iso-Seq method could serve as a new lens through which to see biology more clearly, giving us a better chance to develop interventions that will keep people healthier even in the face of severe disease. DDW

Luke Hickey *is Senior Director of Biomedical Sciences at PacBio. He currently leads development of PacBio's single molecule long-read sequencing technology in the human genetics research market. In this role, he has contributed to the development of novel genetic analysis methods and applications of the technology, including the full-length transcript isoform sequencing (Iso-Seq) method.*