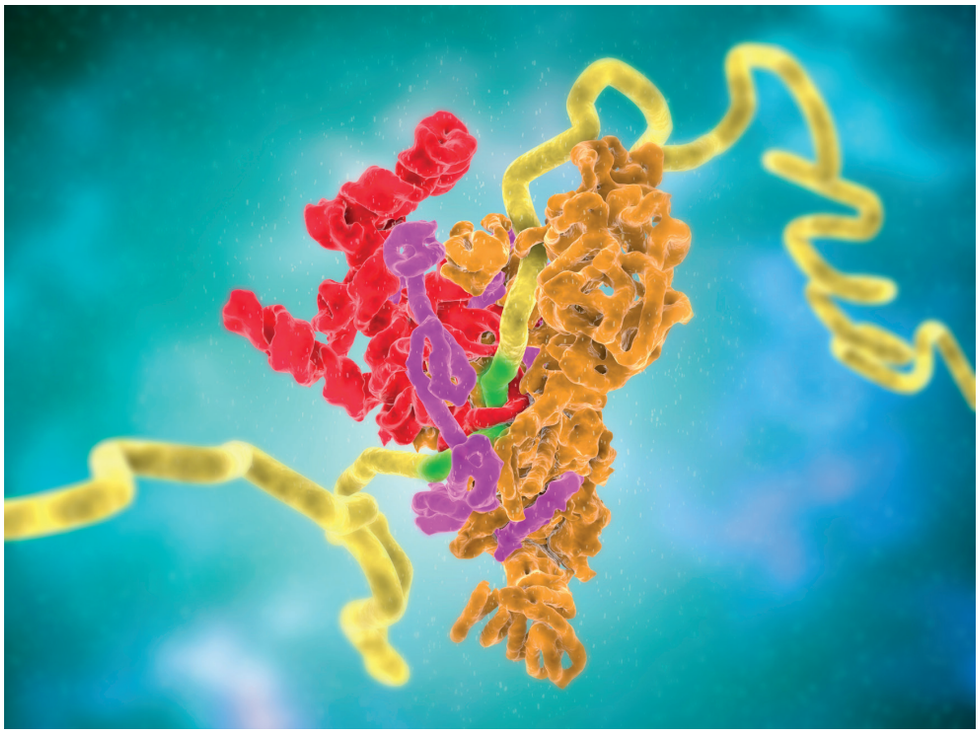


RAMON ANDRADE 3DCIENCIA / SCIENCE PHOTO LIBRARY

LONG-READ SEQUENCING REVEALS MORE OF THE TUMOUR TRANSCRIPTOME

As researchers move away from using only short-read sequencing, they are making new insights into the role of RNA DYSREGULATION in tumour biology.

Cancer is a disease of abnormalities: mutated DNA sequences give rise to atypical proteins and, ultimately, malignant behaviour. But beyond this generality, many details are less clear; how, exactly, do cancer cells change, and what impact do these changes have on cellular behaviour? One particular blind spot concerns alternative splicing, where one gene can be read in several ways to create different isoforms of messenger RNA (mRNA) — and hence different proteins with new functions.

▲ The spliceosome excises introns to create the mRNA. Dysregulation of this mechanism in cancer can create different forms of mRNA.

"Alternative splicing is a process that is very often dysregulated in tumours," says Olga Anczukow, an associate professor at The Jackson Laboratory of Genomic Medicine and Co-Program Leader at the Jackson Laboratory Cancer Center in Farmington, Connecticut. By detailing the various RNA isoforms present throughout a tumour's lifecycle, Anczukow hopes to gain "novel mechanistic insights into how tumours form, as well as identify novel targets that may lead to therapeutic strategies".

Abnormal alternative splicing in cancer can have many causes, including mutations at splice sites or in the splicing machinery. This, in turn, can produce protein isoforms that repress cell death, drive cell growth, or any number of malignant hallmarks. As Omar Abdel-Wahab, the director of Memorial Sloan Kettering's Center for Hematologic Malignancies in New York, explains: "Anything that might change gene expression or the constellation of proteins present in a cell, can be exploited by cancer — and splicing is a key part of that."

Researchers like Anczukow and Abdel-Wahab have struggled to make headway in the past. Sequencing RNA transcripts is a difficult task, especially if the aim is to differentiate alternatively spliced transcripts. This is, in large part, because most laboratories rely on short-read sequencing. But, recent innovations in sequencing technology give them hope that the field will quickly expand and reveal a host of fresh insights into tumour biology.

Short cuts and blind spots

Short-read sequencing works by fragmenting long genetic sequences into strands of approximately 100-300 nucleotides. Each fragment can then be read and processed in parallel, enabling rapid sequencing over large swathes of material.

Because the genetic material has been fragmented, bioinformatics software is needed to assemble the digital reads into a single sequence. For genomic DNA, assembly can be done with the aid of a reference genome against which fragments are compared to determine their order. For RNA, although there are reference transcriptomes, they are often incomplete — making the assembly job much harder.

"We take these short reads and then stitch them together to try to infer what isoforms are being expressed," explains Abdel-Wahab. "But, when you are inferring the isoforms, you are possibly missing some of the complexity."

And there is likely to be significant complexity. Approximately 94% of protein-coding genes undergo alternative splicing[1]. These genes are typically not a continuous sequence of bases, but rather an alternating series of protein-coding (exon) and

"WHEN YOU USE SHORT-READ SEQUENCING, IT'S MUCH HARDER TO RECONSTRUCT THE TRUTH FROM YOUR DATA."

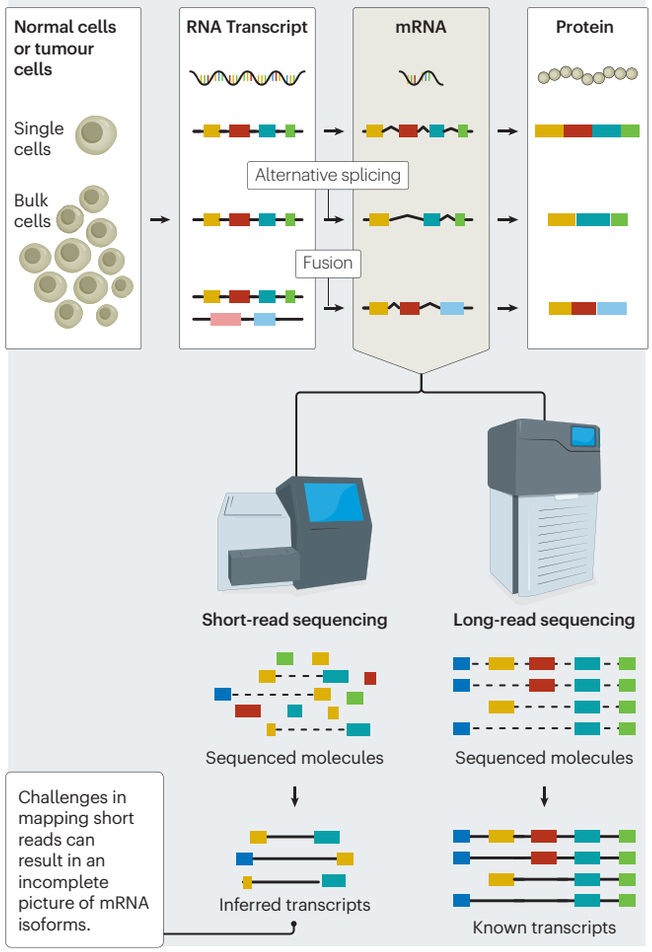
non-coding (intron) sections. Once transcribed into RNA, a protein complex known as the spliceosome will excise introns and splice together exons to form the mRNA, which is the true protein-coding template. However, the specific exons that are included in the mRNA can vary, leading to alternatively spliced transcripts. In this way, a single gene may give rise to many different protein isoforms, with varying functions.

The high degree of homology between alternatively spliced transcripts can make it extremely challenging to connect them to the original sequence (see 'RNA dysregulation in cancer, and how to read it'). "When you use short-read sequencing, it's much harder to reconstruct the truth from your data," says Jason Underwood, a principal scientist at genomic sequencing technology firm PacBio, based in Menlo Park, California.

Studies suggest that current short-read reconstruction methods may only assemble 20 to 40% of human transcriptomes[2,3], but advances in sequencing technology have given researchers an alternative: long-read sequencing.

RNA DYSREGULATION IN CANCER, AND HOW TO READ IT

Aberrant splicing and gene fusion events are common in cancer cells. Long-read sequencing provides a more complete picture of tumour cell transcriptomics, enabling knowledge – rather than inference – of RNA isoforms present.



A longer view of cancer

Whereas short-read sequencing goes up to 300 nucleotides, long-read sequencing can cope with a continuous sequence that is hundreds of thousands of nucleotides long, taking away the guesswork, often allowing researchers to sequence transcripts end-to-end. "That is very powerful," says Anczukow, "because it allows us to detect novel isoforms that we were not able to detect with short-read sequencing."

Anczukow, along with colleagues Christine Beck and Jacques Banchereau, conducted a study[4] to interrogate transcriptomic profiles in both breast cancer cell lines and patient samples. The team identified 142,514 unique full-length transcript isoforms from 16,772 annotated genes, with a mean isoform length of 2,600 nucleotides. Nearly 80% of the transcripts identified were novel, and most of these were found in patient samples, emphasizing the heterogeneity among patients.

These findings show the value of this approach, says Anczukow. "Long-read sequencing has really revolutionized the world of alternative splicing."

Abdel-Wahab agrees. And while short-read sequencing covers most research needs for studying gene expression, he says, "long-read sequencing gives us much more information about splice isoforms and complex structural variations that we can use to learn about cancer evolution and metastatic progression".

Long-read sequencing is particularly good at detecting structural variations in RNA transcripts, such as those formed from the fusion of two distinct genes. Fusions can drive carcinogenic phenotypes, and have attracted attention as both biomarkers and potential neoantigens in immunotherapy.

Identifying fusions requires finding reads that actually span the fusion point between transcripts. This is already a difficult task with short-read sequencing, made harder when attempting to identify fusions between aberrant transcripts. This was underscored in 2018 when long-read sequencing found several novel fusion transcripts in a tumour cell line that had previously undergone extensive sequencing[5]. Among the novel transcripts, the researchers discovered several sequences containing more than one fusion event. And a more recent long-read single cell study has confirmed a three-gene fusion event[6].

Many researchers have avoided long-read sequencing in the past, perceiving it to be slower and more expensive. But, improvements to throughput and accuracy are making it a more attractive option. A new method developed at the Broad Institute uses concatenation of cDNAs into single molecules to enhance long-read throughput 15-fold in single-cell RNA sequencing analyses[7].

Alternative splicing researchers are excited to see long-read sequencing advancing understanding of RNA dysregulation in cancer. "Right now we're still in a gene era," says Underwood. "I look forward to the 'RNA isoform' era. RNA is what makes each cell different." After all, he says, "the biology is in the transcripts". ■

REFERENCES

1. Zhang, Y., Qian, J., Gu, C. & Yang, Y. *Sig Transduct Target Ther* **6**, 78 (2021).
2. Sharon, D., Tilgner, H., Grubert, F. & Snyder M. *Nature Biotechnol* **31**(11):1009-14 (2013).
3. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M.P. *Proc Natl Acad Sci USA* **111**, 9869-74 (2014).
4. Veiga, D.F.T. *et al. Sci Advances* **8**, eabg6711 (2022).
5. Nattestad, M., *et al. Genome Res* **28**, 1126–1135 (2018).
6. Davidson, N.M., *et al. Genome Biol* **23**, 10 (2022).
7. Al'Khafaji, A.M. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2021.10.01.462818 (2021).

PacBio