Researchers develop new, more accurate computational tool for long-read RNA sequencing

January 20 2023

Credit: CC0 Public Domain

On the journey from gene to protein, a nascent RNA molecule can be cut and joined, or spliced, in different ways before being translated into a protein. This process, known as alternative splicing, allows a single

gene to encode several different proteins. Alternative splicing occurs in many biological processes, like when stem cells mature into tissue-specific cells. In the context of disease, however, alternative splicing can be dysregulated. Therefore, it is important to examine the transcriptome—that is, all the RNA molecules that might stem from genes—to understand the root cause of a condition.

However, historically it has been difficult to "read" RNA molecules in their entirety because they are usually thousands of bases long. Instead, researchers have relied on so-called short-read RNA sequencing, which breaks RNA molecules and sequence them in much shorter pieces—somewhere between 200 to 600 bases, depending on the platform and protocol. Computer programs are then used to reconstruct the full sequences of RNA molecules.

Short-read RNA sequencing can give highly accurate sequencing data, with a low per-base error rate of approximately 0.1% (meaning one base is incorrectly determined for every 1,000 bases sequenced). Nevertheless, it is limited in the information that it can provide due to the short length of the sequencing reads. In many ways, short-read RNA sequencing is like breaking a large picture into many jigsaw pieces that are all the same shape and size and then trying to piece the picture back together.

Recently, "long-read" platforms that can sequence RNA molecules over 10,000 bases in length end-to-end have become available. These platforms do not require RNA molecules to be broken up before they are sequenced, but they have a much higher per-base error rate, typically between 5% to 20%. This well-known limitation has severely hampered the widespread adoption of long-read RNA sequencing. In particular, the high error rate has made it difficult to determine the validity of novel, previously unknown RNA molecules discovered in a particular condition or disease.

To circumvent this problem, researchers at Children's Hospital of Philadelphia (CHOP) have developed a new computational tool that can more accurately discover and quantify RNA molecules from these error-prone long-read RNA sequencing data. The tool, called ESPRESSO (Error Statistics PRomoted Evaluator of Splice Site Options), was reported today in *Science Advances*.

"Long-read RNA sequencing is a powerful technology that will allow us to uncover RNA variation in rare genetic diseases and other conditions, like cancer," said Yi Xing, Ph.D., director of the Center for Computational and Genomic Medicine at CHOP and senior author of the study.

"We are probably at an inflection point in how we discover and analyze RNA molecules. The transition from short-read to long-read RNA sequencing represents an exciting technological transformation, and computational tools that reliably interpret long-read RNA sequencing data are urgently needed."

ESPRESSO can accurately discover and quantify different RNA molecules from the same gene—known as RNA isoforms—using error-prone long-read RNA sequencing data alone. To do so, the computational tool compares all long RNA sequencing reads of a given gene to its corresponding genomic DNA, and then uses the error patterns of individual long reads to confidently identify splice junctions—places where the nascent RNA molecule has been cut and joined—as well as their corresponding full-length RNA isoforms.

By finding areas of perfect matches between long RNA sequencing reads and genomic DNA, as well as borrowing information across all long RNA sequencing reads of a gene, the tool is able to identify highly reliable splice junctions and RNA isoforms, including those that have not been previously documented in existing databases.

The researchers evaluated the performance of ESPRESSO using simulated data and data on real biological samples. They found that ESPRESSO performs better than multiple currently available tools, both in terms of discovering RNA isoforms and quantifying them. The researchers also generated and analyzed over 1 billion long RNA sequencing reads covering 30 human tissue types and three human cell lines, providing a useful resource for studying human transcriptome variation at the resolution of full-length RNA isoforms.

"ESPRESSO addresses a long-standing problem of long-read RNA sequencing and could usher in new opportunities of discovery," Dr. Xing said. "We envision that ESPRESSO will be a useful tool for researchers to explore the RNA repertoire of cells in various biomedical and clinical settings."

More information: Yuan Gao et al, ESPRESSO: Robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data, *Science Advances* (2023). DOI: 10.1126/sciadv.abq5072. www.science.org/doi/10.1126/sciadv.abq5072

Provided by Children's Hospital of Philadelphia