genomeweb

Broad Releases Benchmarking Data on New Long Read RNA-seq Toolkit for Fusions

Mar 15, 2024 | Andrew P. Han

Premium Save for later

NEW YORK – A new computational tool developed at the Broad Institute can better detect fusion transcripts with the help of long-read RNA sequencing data, especially from single-cell studies.

Trinity Cancer Transcriptome Analysis Toolkit-Long Read- fusion (CTAT-LR-fusion), developed by the Broad's Methods Development lab, detects RNA fusion transcripts in long reads "with or without companion short reads, with applications to bulk or single-cell transcriptomes," the lab members wrote in a preprint posted to *Biorxiv* last month. They said their tool performs better on fusion detection accuracy than alternative methods and applied it to bulk transcriptomes in nine tumor cell lines as well as single tumor cells from several cancer samples.

"It's hugely important because as we're getting into doing fusion transcript detection, with long read we do way better in single cells, which is important when you're looking at things like tumor heterogeneity where fusion transcripts might be differentially expressed in different populations of cells," said Brian Haas, a senior author of the preprint. "If you just have the short-read data, you might not detect the cells as expressing the fusion at all. And if you do, you might just get evidence of the breakpoint, not the full isoform information."

The new toolkit builds on Haas' work finding RNA fusions using short-read sequencing. Fusion transcripts are important because they can be drivers for oncogenesis and targets for therapy. Trinity, a project done in collaboration with researchers at the Hebrew University of Jerusalem and published in *Nature Biotechnology* in 2011, was able to extract full-length RNA splice isoforms from bulk RNA-seq data. STAR-Fusion, a short-read-based method for fusion transcript detection, was released in 2017 and benchmarked in a 2019 paper in *Genome Biology*. FusionInspector, published last year in *Cell Reports Methods*, helps model fusion contigs and quantify transcript isoforms.

CTAT-LR-fusion works with long reads from both Pacific Biosciences' HiFi and Oxford Nanopore Technologies' sequencing platforms; however, Haas suggested that the PacBio Kinnex kits, based on the MAS-seq method that concatenates single-cell cDNAs for analysis on the firm's sequencers, is crucial because it provides about 15 times more long-read data.

"In many cases, you're getting full-length fusion transcripts and multiple isoforms," he said.

The toolkit uses a customized version of the minimap 2 assembler to generate alignments four times faster than with the standard version. "Once we have candidates and reads providing evidence, we model fusion contigs," Haas said in a talk at the 2024 Advances in Genome Biology and Technology meeting last month in Orland, Florida. After that, the toolkit realigns those reads and defines fusion gene pairs and different isoforms.

The preprint includes benchmarking of CTAT-LR-fusion with other pipelines for long-read RNA isoform analysis, including JAFFAL, LongGF, FusionSeeker, and pbfusion. The authors compared their tool with the others on simulated long reads, isoform detection in a reference sample, in cancer cell lines with known oncogenic fusions, and in patient-derived single-cell tumor samples.

"The development of new long-read fusion tools, and in particular by groups with prior expertise in short-read fusion calling, such as the authors of CTAT-LR, is a welcome contribution," Nadia Davidson, a researcher at Australia's Walter and Eliza Hall Institute and first author on the paper introducing JAFFAL, said in an email. "What we learned from short-read sequencing is that no single method can detect all clinically relevant fusions, and so ensemble pipelines that utilize several tools are needed to achieve optimal sensitivity. No doubt this will also be true for long-read fusion detection."

In cell lines, using both long and short reads, CTAT-LR-fusion found 213 fusion genes with 288 fusion splicing isoforms. In the single-cell samples, long- and short-read methods found a total of 265 tumor cells with the NUTM2A-AS1::RP11-203L2.4 fusions. "Approximately 60 percent of the NUTM2A-AS1::RP11-203L2.4 containing tumor cells were solely identified by long-read evidence, another 20 percent by short reads only, and the remaining 20 percent by both short and long reads," the authors wrote.

There are some fusions where long reads clearly outperform short reads, Haas said, but the reverse is also true. For example, with BCR-ABL fusions, a clinically actionable hallmark of chronic myeloid leukemia, short reads are "one hundredfold better."

CTAT-LR-fusion has already been incorporated into LongSom, a bioinformatics pipeline for long-read single-cell RNA-seq described earlier this month in a preprint coauthored by Haas in collaboration with Niko Beerenwinkel of ETH Zurich and the Tumor Profiler

Consortium. LongSom calls de novo somatic SNVs, copy number alterations, gene fusions, and reconstructs tumor clonal heterogeneity. The pipeline also includes two other tools developed by Haas' team: CTAT-mutations, a variant calling pipeline that uses RNA-seq data and prioritizes variants that may be relevant to cancer biology, and InferCNV, which uses single-cell transcriptome data.

Overall, Haas predicted that long reads are the way of the future for detecting fusion transcripts, especially with higher throughput enabled by the PacBio Kinnex kit, launched in late 2023, and falling sequencing costs. Previously, Kinnex was known as MAS-Iso-Seq (multiplexed arrays isoform sequencing.)

"I absolutely think long read is going to take over and we'll all be doing long reads," he said. "At some point you have to ask yourself, 'Why do short reads when you can do long reads and get all this additional context?'"

Filed Under + Informatics + Sequencing + Gene Expression & RNA Sequencing + North America + Broad Institute + single-cell sequencing + single-cell gene expression analysis + Pacific Biosciences + Preprint + RNA-seq