PacBio

Application brief

Bioinformatics tools for full-length isoform sequencing

Introduction

Full-length isoform sequencing enables accurate characterization of alternative splicing events in eukaryotic species. Long-read RNA-Seq using PacBio® technology (the Iso-Seq® method) eliminates the need for transcript assembly by sequencing full-length cDNAs and enables new discoveries across many applications.

This application note details a wide array of bioinformatics tools for full length- isoform sequencing across isoform classification and annotation, fusion transcript discovery, variant calling, transcript assembly, differential transcript analysis, differential splicing usage, and transcript visualization.

102-326-593 REV01 OCT2023

PacBio

Isoform classification + annotation

The goal of long-read isoform classification and annotation tools is to identify novel isoforms or genes against an existing annotation for creating a master transcriptome. Different tools handle orthogonal data (short reads, CAGE peak, polyA site data, etc.) differently and use various methods for filtering common cDNA library artifacts.

Many of the long-read tools here also quantify the transcripts as part of the classification and annotation process, e.g., FLAIR, Mandolorian, IsoQuant, Bambu, while other tools such as SQANTI3 rely on the upstream Iso-Seq pipeline to provide read counts.

SQANTI3

SQANTI3 (Pardos-Palacios et al., 2023) is long-read transcript classification and filtering tool. It compares Iso-Seq transcripts against a reference annotation (e.g., Gencode) and classifies them as known/novel genes/isoforms. The nomenclature developed by SQANTI3 such as FSM, ISM, etc., is widely adopted by the long-read RNA community.

The PacBio official Iso-Seq bioinformatics workflow implements *pigeon*, which is based on SQANTI3 and integrates with the rest of the Iso-Seq tools that support both bulk & single-cell analysis.

TALON

TALON is a multi-sample long-read transcript classification and filtering tool. It compares Iso-Seq transcripts against a reference annotation and classifies them as known/novel isoforms. TALON further filters novel transcripts based on reproducibility and internal priming evidence. TALON outputs a transcriptome (GFF) and abundance information (TSV, AnnData). TALON supports both bulk and single-cell data.

Cerberus

Cerberus is an annotation tool that can be used to create a master transcriptome catalog of transcript start sites, end sites, and intron chains. It can be used to annotate existing transcriptomes as well as summarize and visualize transcriptome diversity.

LAPA

LAPA calls transcript start and end sites using long- and short- read data. It uses peak-calling and reproducibility filtering to call definitive ends of transcripts, and performs differential 5'/3' end usage testing between biological conditions.

FLAIR

Originally developed for long reads with higher error rates, FLAIR is a long-read annotation and quantification tool that will create a high-confidence isoform set from combining long- and short- read data. It also quantifies isoform usage by sample.

Mandalorion

Mandalorion is an isoform classification and quantification tool designed specifically for high-accuracy long reads. It generates isoform consensus sequences based on mapped Iso-Seq reads which are then re-aligned and filtered to generate a final transcript dataset.

IsoQuant

IsoQuant is a multi-sample long-read transcript annotation and quantification tool. It can predict transcript models both with and without a reference annotation, as wells as provide gene- and isoform-level quantification.

BAMBU

Bambu is a multi-sample long-read transcript classification and quantification tool that can be used both with and without initial training on an annotation model.

Case study brief: ENCODE long-read annotation pipeline

The ENCODE4 consortium uses TALON to define the initial transcript models from Iso-Seq data, followed by LAPA to refine transcript 5'/3' ends, and finally Cerberus to generate a master transcriptome.

102-326-593 REV01 OCT2023

Fusion transcript discovery

The goal of long-read fusion tools is to utilize the full-length transcript sequences in order to characterize not only fusion breakpoints, but also the splicing structure of the fusion gene, as multiple isoforms can emerge from the same fusion event.

pbfusion

pbfusion is a breakpoint-centric fusion gene caller. It uses mapped Iso-Seq reads and reference gene annotations to identify transcript breakpoints between genes and also provides visualization scripts.

CTAT-LR-fusion

CTAT-LR-fusion detects fusion transcripts from bulk and single-cell long-read data. While not required, it can additionally leverage short RNA-seq reads to further maximize detection of fusion isoform splice variants and fusion-expressing single cells. It also includes interactive visualizations to showcase read alignment evidence for identified fusions and isoform breakpoints.

FLAIR-fusion

FLAIR-fusion uses the FLAIR tool to first align and correct splice sites, then apply filters (e.g., genomic distance, breakpoint distance from splice sites) to distinguish true fusion events from library artifacts.

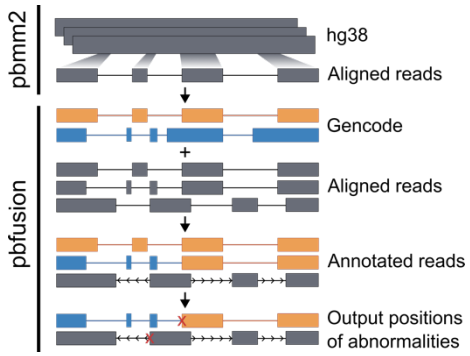

Figure 1. pbfusion workflow for detecting fusion transcripts in Iso-Seq data.

Variant calling

Many existing variant callers developed for genomic data have also been proven to work for transcriptome data as well. For example, de Souza et al. (2023) showed that DeepVariant, GATK, and Clair3 all can be modified to work on Iso-Seq data.

DeepVariant

DeepVariant is a deep learning-based variant caller that uses aligned reads to produce pileup images and classifies them with a convolutional neural network. DeepVariant has been trained on PacBio HiFi data and is suitable for both PacBio WGS and Iso-Seq data.

CTAT-Mutations

CTAT-Mutations is a variant calling pipeline for calling variants from transcriptome data that integrates GATK along with downstream steps to annotate, filter, and prioritize variants.

Transcript assembly

Long read-aware transcript assemblers can take advantage of matching long- and short- read data to generate transcripts. Some transcript assemblers do not need the reference genome, which are suitable especially in cases where high-quality reference genomes are not available.

StringTie2

StringTie2 is a reference-guided transcript assembler that works with both short- and long-read data.

RnaSPAdes

RnaSPAdes is a genome-free transcript assembly that works with both short- and long-read data.

Differential transcript analysis

Many existing differential expression analysis tools developed for short-read RNA-Seq can be used to detect differential gene expression (DGE) or differential transcript (isoform) expression (DTE). These tools use count information, which could be gene counts or isoforms counts. However, with Iso-Seq data, more can be done than simply detecting DE isoforms such as isoform switching events (see Terminology box).

tappAS

tappAS is a long-read differential analysis tool that can identify gene- and isoform-level expression and usage changes as well as its functional impact. It uses DEXSeq and maSigPro for DTU analysis. tappAS is part of the SQANTI3-IsoAnnot-tappAS framework for long-read RNA analysis.

DESeq2

DESeq2 is a R Bioconductor package for differential expression analysis. It can be used for both short and long reads. For example, Leung et al. (2023) used Iso-Seq read counts as input to DESeq2.

DRIMSeq

DRIMSeq is a R Bioconductor package for DTU analysis between different conditions and tuQTL analysis. It can also visualize the results in R.

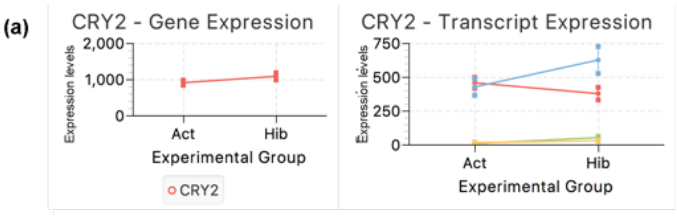

Figure 2. Example of differential transcript usage (DTU) with isoform switching from Tseng et al. (2021). The CRY2 gene has four isoforms and is not differentially expressed at the gene level (no DGE). Two of the isoforms (yellow, green) show no differential expression, while the other two (blue, red) show an isoform switching event from one condition to another.

Terminology: Different types of transcriptional changes

Differential gene expression (DGE): A gene that is expressed in different abundances across conditions. Identifying DGE is one of the most common analysis goals of short-read RNA-Seq.

Differential transcript expression (DTE): A transcript/isoform that is expressed in different abundances across conditions.

Differential transcript usage (DTU): Differences in the relative abundance of isoforms of the same gene. **DIU** (differential isoform usage) is sometimes written instead of DTU. Having at least one DTE is a necessary but not sufficient condition for DTU.

Isoform switching: A case of DTU where the dominantly expressed isoform switches from one condition to another.

Differential splicing usage

While long reads provide the benefit of full-length isoform resolution, exon- or splicing-based analyses can still reveal powerful information in a study.

DEXSeq

DEXSeq is a R Bioconductor package for differential exon usage analysis across multiple samples.

SUPPA2

SUPPA2 is a splicing analysis tool that can identify differential splicing and differential transcript usage (DTU) events across multiple conditions with replicates.

Case study brief: Isoform changes in hibernating bears

Tseng et al. (2021) created a master transcriptome by combining existing annotation with Iso-Seq data, then used the SQANTI3 → IsoAnnot → tappAS pipeline to identify DTUs such as the one shown in figure 2.

Transcript visualization

The visualization of full-length transcripts, especially when there are many isoforms per gene, can be a challenge. While multi-purpose genome browsers such as UCSC Genome Browser and IGV work well with the GFF/GTF formats most long-read tools produce, there are visualization tools that have been dedicated to visualizing complex isoform data.

Swan

Swan is a long-read visualization tool that can plot expression or percent isoform values of each transcript in a gene alongside the transcript's actual model. It also performs statistical tests to find isoform-switching genes.

ggtranscript

ggtranscript is an extension to the R package ggplot2 that can visualize long-read transcript structure and annotation.

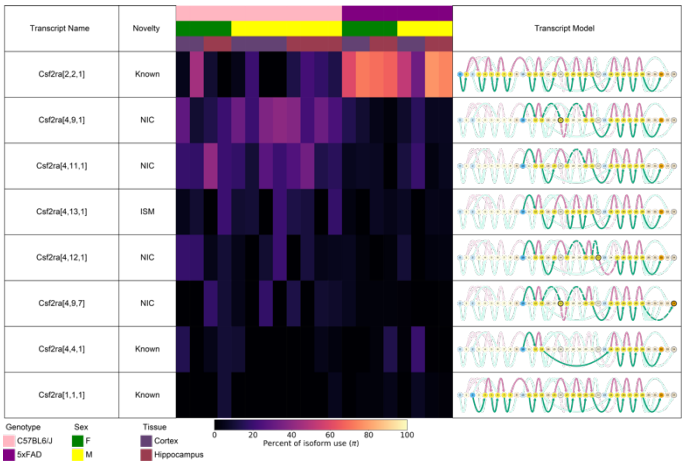

Figure 3. Example figure from Swan, a long-read isoform visualization tool.

Conclusion

The breadth of these computational tools highlight the availability of bioinformatic workflows that can easily capitalize on accurate full-length PacBio Iso-Seq data. The integration of these tools into the Iso-Seq workflow enables unambiguous isoform characterization that is critical for biological and disease studies.

Resources and references

Official PacBio Iso-Seq bioinformatics resources

Iso-Seq documentation

SMRT® Link & SMRT Analysis software

Additional resources

Long-read-tools.org aggregates an on-going collection of all long-read bioinformatics tools.

LRGASP consortium performed extensive testing of various long-read transcript annotation + quantification tools.

Dong et al. (2023) benchmarked StringTie2, Bambu, DESeq2, edgeR, and limma-voom for differential analysis.

References

Leung, S. K., et al., (2023). Long-read transcript sequencing identifies differential isoform expression in the entorhinal cortex in a transgenic model of tau pathology. *bioRxiv*, 2023-09. https://doi.org/10.1101/2023.09.20.558220

Pardo-Palacios, F. J., et al., (2023). Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *bioRxiv*, 2023-07. https://doi.org/10.1101/2023.07.25.550582

de Souza, V. B., Jordan, B. T., Tseng, E., Nelson, E. A., Hirschi, K. K., Sheynkman, G., & Robinson, M. D. (2023). Transformation of alignment files improves performance of variant callers for long-read RNA sequencing data. *Genome Biology*, *24*(1), 91. https://doi.org/10.1186/s13059-023-02923-y

Tseng, E., et al., (2022). Long-read isoform sequencing reveals tissue-specific isoform expression between active and hibernating brown bears (Ursus arctos). *G3*, *12*(3), jkab422. https://doi.org/10.1093/g3journal/jkab422

102-326-593 REV01 OCT2023