Accurate Isoform Identification and Quantification from Long Read Isoform Sequencing with Applications to Bulk and Single Cell Transcriptomes

Brian Haas[1], Houlin Yu[1], Dan Bartlett[1], Emily White[1], Asa Shin[1], Christophe Georgescu[1], Can Kockan[1], Akanksha Khorgade[1], James T. Webber[1], Clotilde Lagier-Tourenne[2], Michael Ward[3], Paul Blainey[1], Victoria Popic[1], Niall Lennon[1], and Aziz Al'Khafaji[1]

1. Broad Institute of MIT and Harvard, Cambridge, MA; 2. MassGeneral Institute for Neurodegenerative Diseases, Charlestown, MA; 3. National Institute of Health, Bethesda MD

Abstract

Advances in long isoform read sequencing continue to yield unprecedented insights into the functional outputs of gene expression with ever-increasing resolution. While many long isoform read sequences are full-length and sequenced from the transcriptional start site to the poly-A tail, partial read sequences stemming from incomplete reverse transcription or RNA degradation pose major challenges for unambiguous isoform identification and quantification.

We present our new computational method Long Read Alignment Assembler (LRAA) for isoform identification and quantification based on spliced genome alignments of long read isoform sequences. LRAA leverages a novel splice graph implementation, read structure annotation, compatible read collapse and artifact filtering to yield isoform structures in both reference annotation-free and reference annotation-guided modalities. In benchmarking LRAA isoform identification and quantification in comparison to 11 alternative methods using both simulated and real PacBio or ONT isoform sequences, LRAA is demonstrated to achieve top performance.

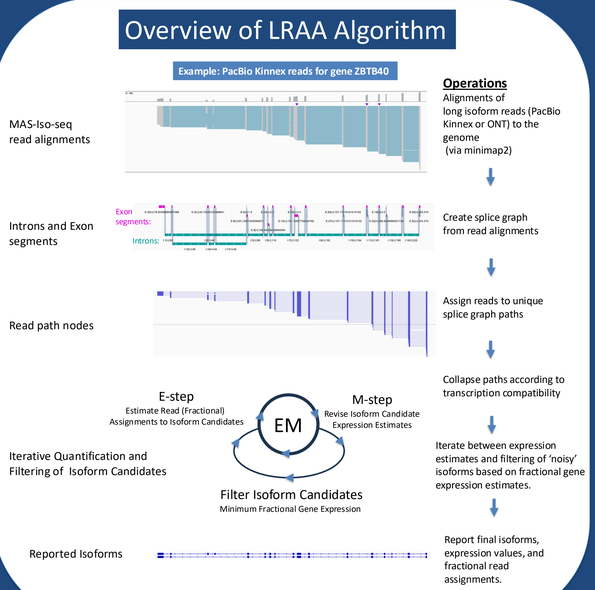
To address the inherent difficulty of benchmarking long read isoform quantification with real sequencing data regarding incomplete read sequences and lack of perfectly defined truth sets, we sequenced 2923 isoforms of 1566 human genes engineered with isoform-specific barcodes and expressed in cells. The barcoded isoforms enabled incompletely sequenced reads to be annotated with unambiguous isoform assignments. The barcodes were then trimmed, leaving the remaining reads with ground truth for quantification accuracy evaluation of tools. We find LRAA to be among the top-performing methods for long read isoform quantification and the most accurate method to enable both isoform identification and quantification capabilities together.

The computational efficiency of LRAA has been a great asset to us in processing and analyzing single cell transcriptomes. In our exploration of single nucleus sequencing of brain samples derived from patients with Amyotrophic lateral sclerosis (ALS), we identified disease-relevant isoforms involving cryptic splicing events evident in individual neurons.

Capabilities of Related Methods for Long Read-based Isoform Identification and Quantification

Tools	Isoform Identification	Reference Annotation-Guided	Reference Annotation-Free	Expression Quantification	Single Cell Read Assignments	Primary Input
LRAA	✔	✔	✔	✔	✔	bam
IsoQuant	✔	✔	✔	✔	✔	bam
Mandalorion	✔	✔	✔	✔		fastq
Bambu	✔	✔	✔	✔		bam
StringTie	✔	✔	✔	✔		bam
IsoSeq	✔	✔	✔	✔		fastq
Isosceles	✔	✔		✔	✔	bam
FLAIR	✔	✔		✔		fastq
FLAMES	✔	✔		✔		bam
ESPRESSO	✔	✔		✔		bam
TALON	✔	✔		✔		bam
Oarfish				✔		fastq

Overview of LRAA Algorithm



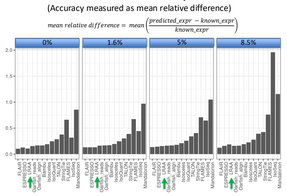
Benchmarking Isoform ID and Quantification Using Simulated PacBio IsoSeq

PacBio IsoSeq reads simulated using **IsoSeqSim** at varying levels of sequencing error: 0, 1.6, 0.5, and 8.5% error.

Mouse

35M long reads

Mean Relative Difference



Quantification of Expression

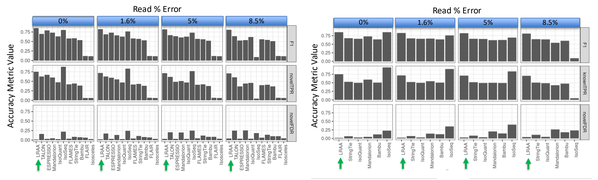
(Accuracy measured as mean relative difference)

Reference-Guided Isoform Discovery

(Single isoform provided per gene, assess discovery of remaining isoforms)

Annotation-Free Isoform Discovery

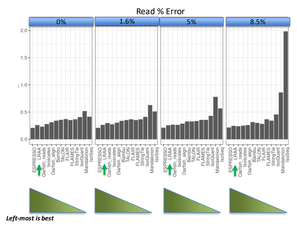
(Only genome-aligned reads provided as input)

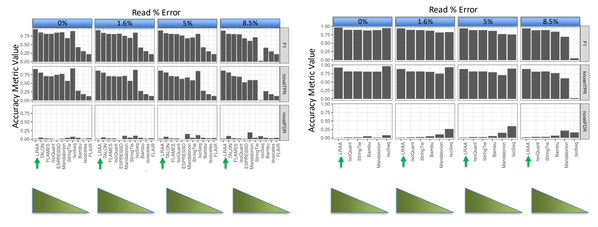


Arabidopsis

20M long reads

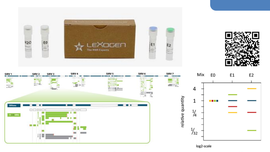
Mean Relative Difference



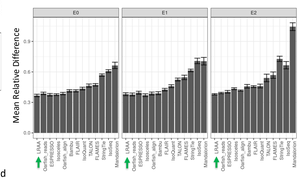


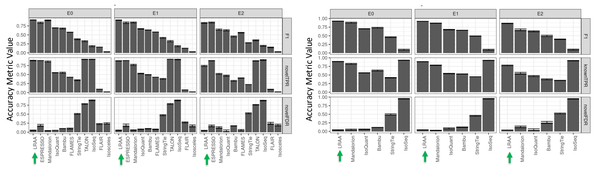
Benchmarking Isoform ID and Quantification Using Sequencing Standards

Lexogen Spike-In RNA Variant Control Mixes (SIRVs)

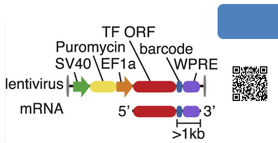


Spike-in RNA synthetic isoforms (SIRVs) are commercially available from **Lexogen** and include 69 RNA isoforms for 7 synthetic genes with structures modeled on known genes. They are available in different relative concentrations as shown below. We sequenced the SIRV sets using PacBio Kinnex with 4 replicates each and assayed quantification and isoform identification accuracy, 100-300k reads each replicate.



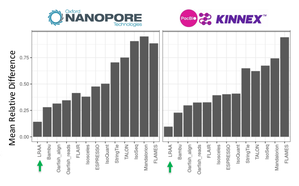


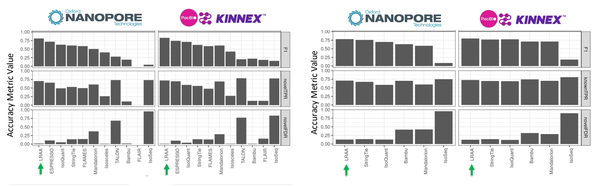
Multiplexed Overexpression of Regulatory Factors (MORF) Library



We sequenced 2923 isoforms of 1566 human genes engineered with isoform-specific barcodes and expressed in cells. The barcoded isoforms enabled incompletely sequenced reads to be annotated with unambiguous isoform assignments. The MORF library is commercially available from **addgene** and described in *Zhang et al. Cell. 2023. Lentivirus transfected into HEK293 cells followed by RNA-prep.*

4M PacBio and 1M ONT reads sequenced.





Isoform Identification From PacBio Kinnex Sequencing of Four Whole Transcriptomes

PacBio Kinnex sequencing was applied to transcriptomes corresponding to:

- two cell lines BT474 and K562
- the Universal Human Reference RNA standard (UHRR)
- the Genome in a Bottle (GIAB) RNA sample for HG002.

Accuracy statistics are shown as the mean and standard error across the three replicates for each sample.

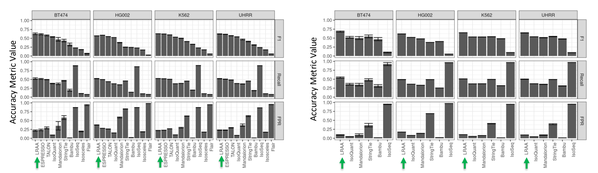
~13M reads sequenced each replicate

Reference-Guided Isoform Discovery

(Single isoform provided per gene, assess discovery of remaining isoforms)

Annotation-Free Isoform Discovery

(Only genome-aligned reads provided as input)



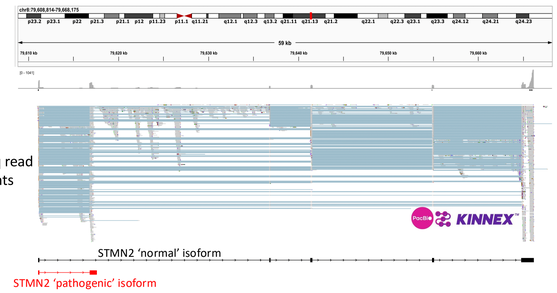
Application of LRAA to Kinnex snuc-Seq of a Human ALS Brain Tissue Sample and Detection of Disease-related STMN2 Aberrant Splicing

Abnormal splicing of the Stathmin 2 (STMN2) gene is a key finding in Amyotrophic Lateral Sclerosis (ALS) patients, particularly those with TDP-43 proteinopathy, where the abnormal splicing is caused by a loss of function of the TDP-43 protein, leading to reduced levels of functional STMN2 protein and contributing to motor neuron degeneration characteristic of ALS.

Aberrant STMN2 splicing leads to the inclusion of a "cryptic exon" in the STMN2 mRNA transcript, resulting in an abnormal, truncated STMN2 protein with reduced functionality. As STMN2 is crucial for axon stability and neuronal growth, aberrant splicing and concomitant reduction of normal STMN2 protein in neurons is thought to contribute to damage of neurons in ALS.

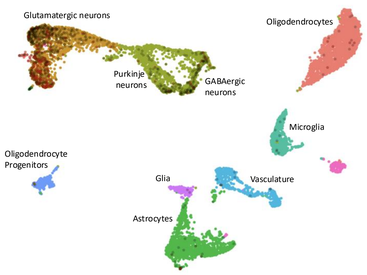
We applied single nucleus sequencing of a human ALS brain tissue sample using 10x Genomics Chromium coupled with PacBio Kinnex long isoform sequencing. Using LRAA for isoform identification and quantification, we identified both the normal and disease-associated STMN2 isoforms (as shown below).

Kinnex snuc-Seq read alignments



Using LRAA estimated expression levels per nucleus, we defined cell clusters and leveraged UMAP to visualize cell clusters (shown below).

From the cell barcodes corresponding to reads assigned to the the normal (**black**) and pathogenic (red) isoforms of STMN2, we identified those cells with evidence of expressing these isoforms and highlight them in the UMAP.



Pathogenic STMN2 isoforms are largely found expressed in glutamatergic neurons and consistent with cells expressing the normal isoform and more generally expressing the STMN2 gene.

Whether input data is single cell, single nucleus, or spatial data, long isoform sequencing and we expect LRAA isoform identification and quantification should prove highly effective.

Conclusion

LRAA:

- Provides quantification accuracy on par with best available methods
- Enables reference-guided or reference annotation-free isoform discovery
- Provides read-to-isoform assignments for extracting cell or spatial barcodes and readily interfaces with Seurat and related toolkits.
- Compatible with both PacBio and ONT isoform sequencing data

Software and documentation

https://github.com/MethodsDev/LongReadAlignmentAssembler

PacBio Pacific Biosciences supported this work