Call & Join vs. Join & Call

identifying isoforms across biological replicates

Genomics of Gene Expression Lab

Fabian Jetzinger[1,2,3], Stanley Cormack[2], Jorge Mestre-Tomás[2], Carolina Monzó[2], Stefan Götz[1], Alejandro Paniagua[2,3], Ana Conesa[2]

[1] BioBam Bioinformatics S.L., Valencia, Spain
[2] Genomics of Gene Expression Lab, I²SysBio, Valencia, Spain
[3] Department of Computer Science, University of Valencia, Spain

Introduction

In the rapidly advancing field of long read transcriptome analysis, many algorithms have been developed to identify and quantify transcripts from sequencing data. While the LRGASP[1] benchmark provided valuable insights, it did not compare strategies on how to handle multiple biological replicates. Isoforms which occur only rarely in one replicate may be more common in another. However, mixing replicates to call transcript models may result in artifacts, sample specific variants going undetected, or may simply be computationally costly.
We propose two strategies to investigate the best practices for analysing replicated data sets of long-read RNA sequencing. We compare multiple isoform identification tools commonly used in the community on these two strategies.

Materials & Methods

We examine two different strategies for the combination of data from biological replicates.

We perform this comparison on a novel data set of mouse brain and kidney tissue with 5 samples, sequenced with ONT PromethION, PacBio Sequel II, and Illumina NovaSeq. Additionally, Spike-In RNA-Variants (SIRVs) were added to each sample to serve as ground truth.

The sequencing depth of the ONT data is higher (~4,5-11,2M reads per sample) than of the PacBio data (~3,6-5,4M reads per sample).

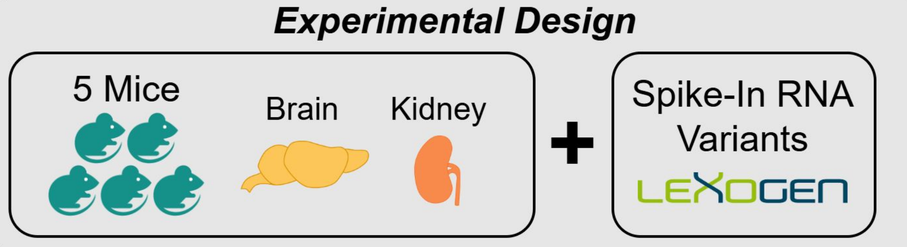
Isoform Identification Tools:
We evaluated six reconstruction tools: FLAIR[3], IsoQuant[4], TALON[5], Bambu[6], Mandalorion[7], and IsoSeq[8] + SQANTI3[9] Filter.

Join & Call:
Reads of all samples are concatenated per condition. Isoform identification creates a transcriptome for each condition (tissue).

Call & Join:
Isoform identification is performed individually on each sample. We then use TAMA Merge[2] to create a joint transcriptome for each condition.

Postprocessing:
SQANTI3 is used for quality control of the results to assess the structural categories of isoforms.

Experimental Design

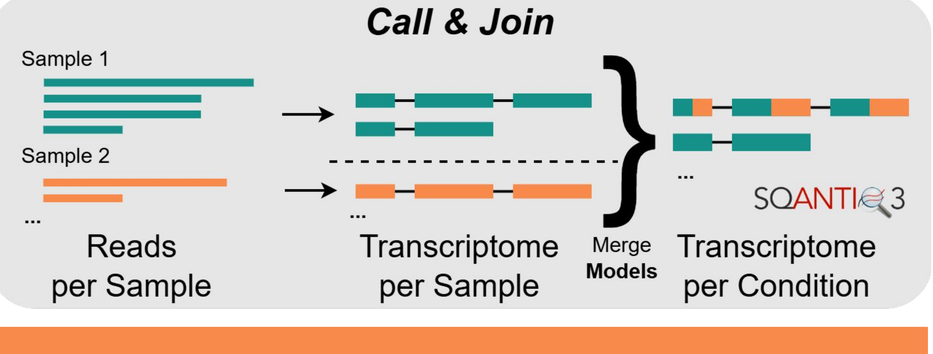


Sequencing

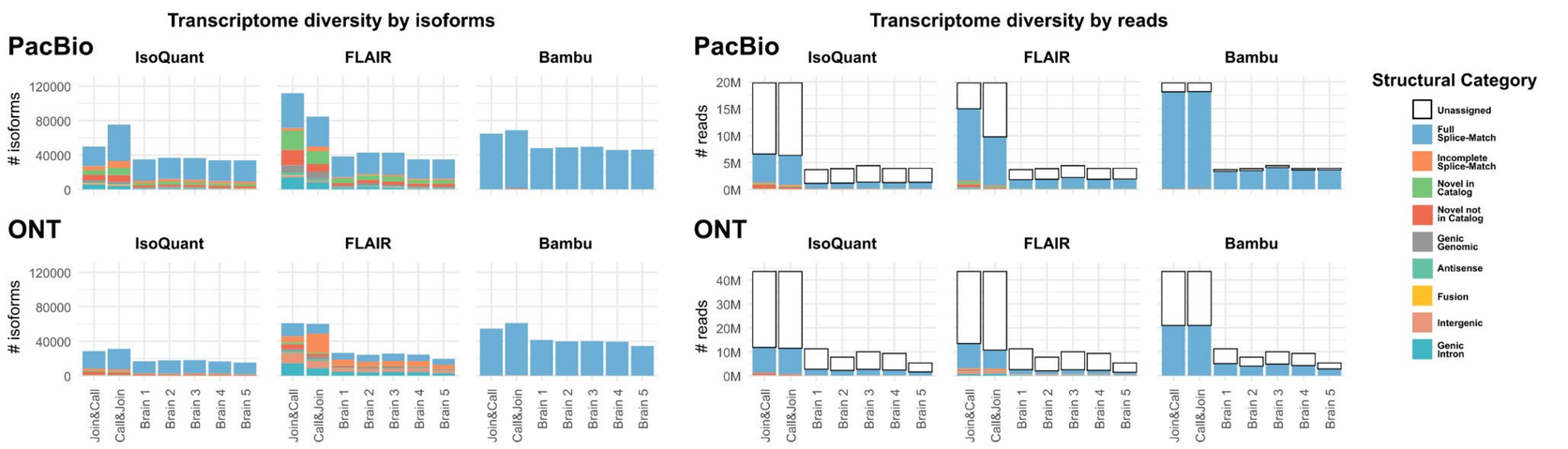
PacBio IsoSeq (Sequel II) — PacBio
illumina — illumina
ONT cDNA (PromethION P2) — Oxford NANOPORE Technologies

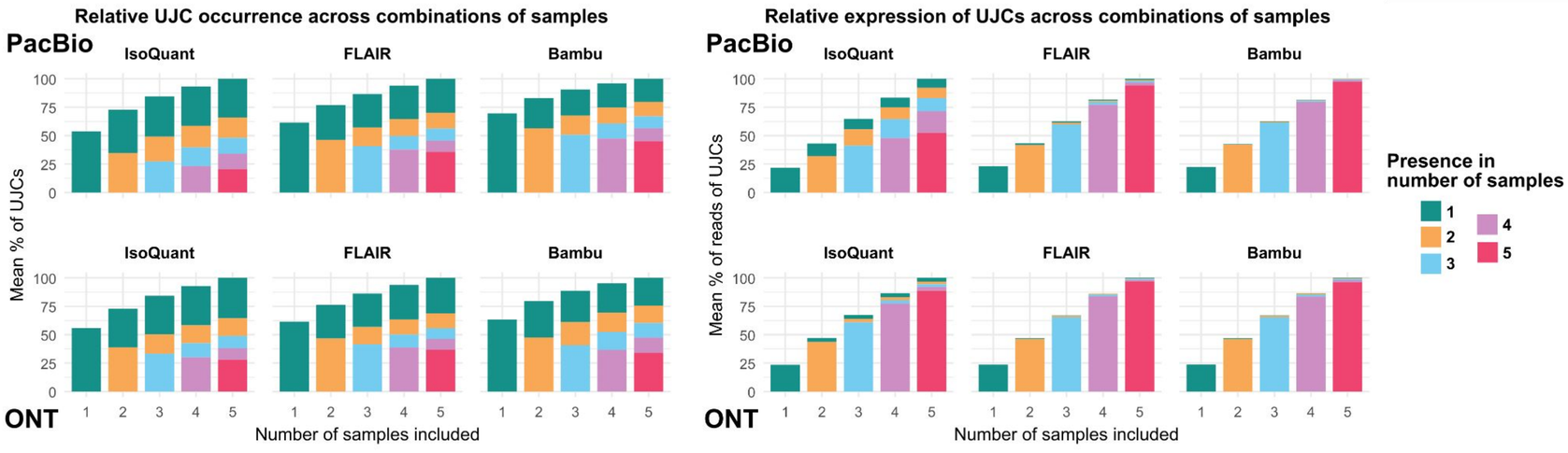
Join & Call

Call & Join



Results





Transcriptome diversity varies not just based on data type and choice of reconstruction tool, but also between the two combination strategies.
FLAIR (more permissive) **recovers more isoforms in the Join & Call strategy**, while **IsoQuant** and **Bambu** (more restrictive) **recover more isoforms in the Call & Join strategy.**

However, regardless of tool or strategy, the **majority of reads** is ultimately assigned to **Full Splice-Match** isoforms, which closely match the reference.

We further examine the discovery of **Unique Junction Chains (UJCs)** when considering varying numbers of samples. While we discover that a plurality of UJCs discovered is not present in every single sample, we also see that the majority of reads is assigned to UJCs that occur in all samples.

A notable outlier is IsoQuant, especially on PacBio data, which assigns a notable number of reads to more sample-specific UJCs.

Conclusions

1. While the capacity for novel isoform detection depends largely on the amount of data, it also varies widely between different algorithms, reflecting differences in transcript reconstruction strategies.
2. When the aim is to discover new isoforms, use the **Join & Call** strategy e.g. with FLAIR.
   Otherwise, **Call & Join** together with a reference-faithful method, e.g. Bambu, is a more time-effective approach.

References

1. Pardo-Palacios, Francisco J., et al. "Systematic assessment of long-read RNA-seq methods for transcript identification and quantification." Nature methods (2024): 1-15.
2. Kuo, Richard I., et al. "Illuminating the dark side of the human transcriptome with long read transcript sequencing." BMC genomics 21 (2020): 1-22.
3. Tang, Alison D., et al. "Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns." Nature communications 11.1 (2020): 1438.
4. Prjibelski, Andrey D., et al. "Accurate isoform discovery with IsoQuant using long reads." Nature Biotechnology 41.7 (2023): 915-918.
5. Wyman, Dana, et al. "A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification." Biorxiv (2019): 672931.
6. Chen, Ying, et al. "Context-aware transcript quantification from long-read RNA-seq data with Bambu." Nature methods 20.8 (2023): 1187-1195.
7. Volden, Roger, et al. "Identifying and quantifying isoforms from accurate full-length transcriptome sequencing reads with Mandalorion." Genome Biology 24.1 (2023): 167.
8. Pacific Biosciences. Iso-Seq - Scalable De Novo Isoform Discovery from Single-Molecule PacBio Reads. Pacific Biosciences (2024).
9. Pardo-Palacios, Francisco J., et al. "SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms." Nature Methods (2024): 1-5.

Contact: fjetzinger@biobam.com

Acknowledgements

This project has received funding from the European Union's programme Horizon Europe under the Marie Skłodowska-Curie grant agreement Number 101072892.