UNIVERSITÄT
LEIPZIG

BIO Leipzig
‹INFORMATICS›

Full-length Isoform Splice-Aware Synthetic RNA-Seq Benchmark of Transcript Assemblers and Quantifiers

Daniel Bandhu Nimptsch Guzmán
Msc. Biology student
Dr. Thomas Gatter, Prof. Peter Stadler

41st TBI Winterseminar 2026 in Bled

February 12, 2026

RNA-seq in Biological Research

- RNA-seq popular alternative to classic microarrays

  (Corchete et al., 2020; Deshpande et al., 2023)

  - Transcriptome-wide analysis
- Human reference databases $> 100,000$ alternative transcripts

  (Pozo et al., 2021)

  - Still incomplete
- Third-generation sequencing technologies
  - Novel transcripts and alternative splicing (AS) events

  (Park et al., 2018)

RNA-Seq Data Analysis

- **RNA-seq data analysis typically involves:**
  - Trimming
  - Alignment
  - Counting
  - Normalization
  - Differential gene expression

(Corchete et al., 2020)

RNA-Seq Quantification Approaches

Approach	**Category**	**Tools**
Alignment-based	Aligners	HISAT2, STAR, TopHat
	Assemblers	Scallop, StringTie, **Ryūtō**
	Traditional counting	HTSeq, featureCounts
Pseudoalignment-based	-	Sailfish, Salmon, Kallisto

Workflows:

- **Alignment-based:** Alignment $\rightarrow$ Assembly $\rightarrow$ Quantification
- **Traditional counting-based:** Alignment $\rightarrow$ Quantification
- **Pseudoalignment-based:** Direct quantification

Challenges in RNA-Seq Data Analysis

- Short reads vs long transcripts (Sarantopoulou et al., 2021; Deshpande et al., 2023)
  - Multi-mapping ambiguity
- Alternative splicing introduces quantification errors (Deshpande et al., 2023)

  (Pardo-Palacios et al., 2024)

Mission Statement

- Determine bias of transcript quantification tools
  - Benchmark of quantification accuracy
  - AS-specific annotation completeness
- Implement best-practice benchmarking pipeline facilitating continuous community updates (Brooks et al., 2024a; Beaulieu-Jones and Greene, 2017; Di Tommaso et al., 2017)

Problem Statement

- How does AS-specific annotation completeness affect transcript quantification accuracy?
- Which AS event types elicit vulnerability in quantification tools?
- Transcript quantification tools performance under AS-specific annotation gaps

Literature Review: Tool Performance Under Annotation Constraints

- **Annotation quality differentially impacts tools:**
  - Annotation incompleteness drives quantification error more than read accuracy (Pardo-Palacios et al., 2024)
  - Annotation quality affects tools differently (Raplee et al., 2019)
- **Tool performance:**
  - Winners: Salmon, Kallisto, RSEM, Cufflinks (Sarantopoulou et al., 2021)
  - Conservative methods underperform: HTSeq and featureCounts (Sarantopoulou et al., 2021)
- **Benchmark approaches:**
  - Synthetic benchmarking with random SNP introduction (Coxe et al., 2024)
- **Hint at AS annotation impact:**
  - Junction-level accuracy consistently lower than base-level accuracy (Coxe et al., 2024)

Literature Review: Critical Gaps

- **AS-specific annotation effects:**
  - Current benchmarks lack evaluation of AS-specific annotation or read-level error effects (Pardo-Palacios et al., 2024; Coxe et al., 2024)

AS Event types

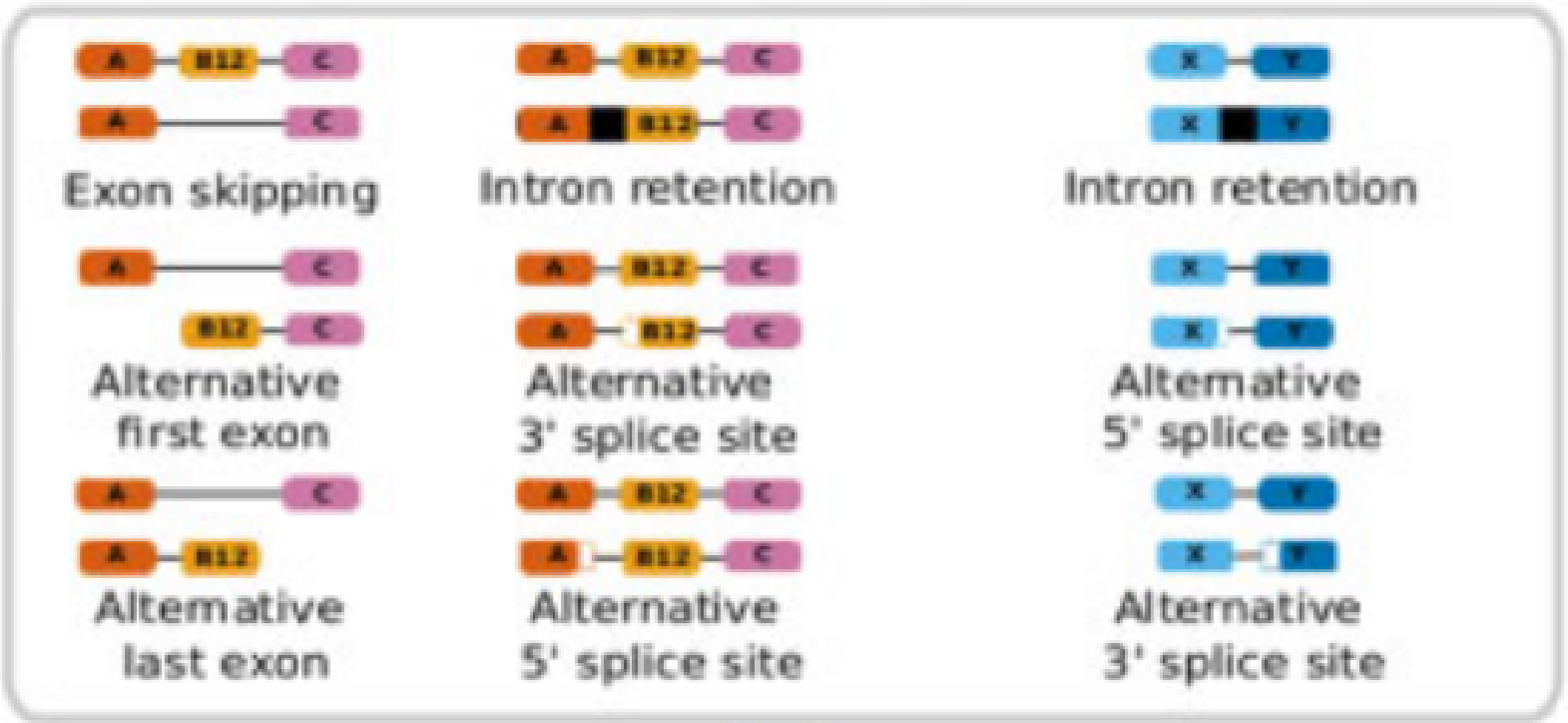

Figure 1: Alternative Splicing (AS) event types (Manz et al., 2021).

Literature Review: Short-Read RNA-Seq Simulation

Event Type	ASimulatoR	BEERS2	YASIM
ES (Exon skipping)	X	X	X
MES (Multiple exon skipping)	X	(X)	(X)
IR (Intron retention)	X	-	X
A3 (Alternative 3' splice site)	X	-	X
A5 (Alternative 5' splice site)	X	-	X
MEE (Mutually exclusive exons)	X	-	-
ALE (Alternative last exon)	X	-	-
AFE (Alternative first exon)	X	-	-

Table 1: Comparison of RNA-Seq simulators ASimulatoR (Manz et al., 2021), BEERS2 (Brooks et al., 2024b) and YASIM (Wan, 2024) with controllable AS event types.

- **Simulator capabilities:**
  - Most simulators do not model specific AS event types
  - FLUX simulator lacks complexity compared to BEERS2

    (Sarantopoulou et al., 2021)

Methodology: Experimental Design

- Isolate AS-annotation errors via **three scenarios**:
  1. **Control:** Transcript present in reads & annotation
  2. **Incomplete annotation:** Transcript simulated in reads, but some removed from annotation
  3. **False annotation:** Transcript present in annotation, but some have no expression in reads

Alternative Splicing (AS)

- 8 AS event types
- Manipulation frequency gradient: 0.05 – 0.95 rel. freq.

Tools

- Aligners: STAR, HISAT2
- Assemblers: Scallop, Ryūtō, StringTie
- Traditional: Featurecounts
- Pseudoaligners: Salmon, Kallisto

Simulation Process Overview

1. **Genomic AS Relative Frequency**
   - Reference Annotation (GFF) $\rightarrow$ Gene Splice Graphs $\rightarrow$ AS Event Relative Frequencies of Reference Genome

2. **ASimulatoR** (Manz et al., 2021)
   - Reference Annotation (GFF), Chromosome FASTAs, AS Event Relative Frequencies $\rightarrow$ Transcript Variants with AS Events and Counts

3. **Polyester** (Frazee et al., 2014)
   - Transcripts FASTA + Counts $\rightarrow$ Paired-End Reads with Library Preparation and Sequencing Biases

Theoretical AS Frequency Inference

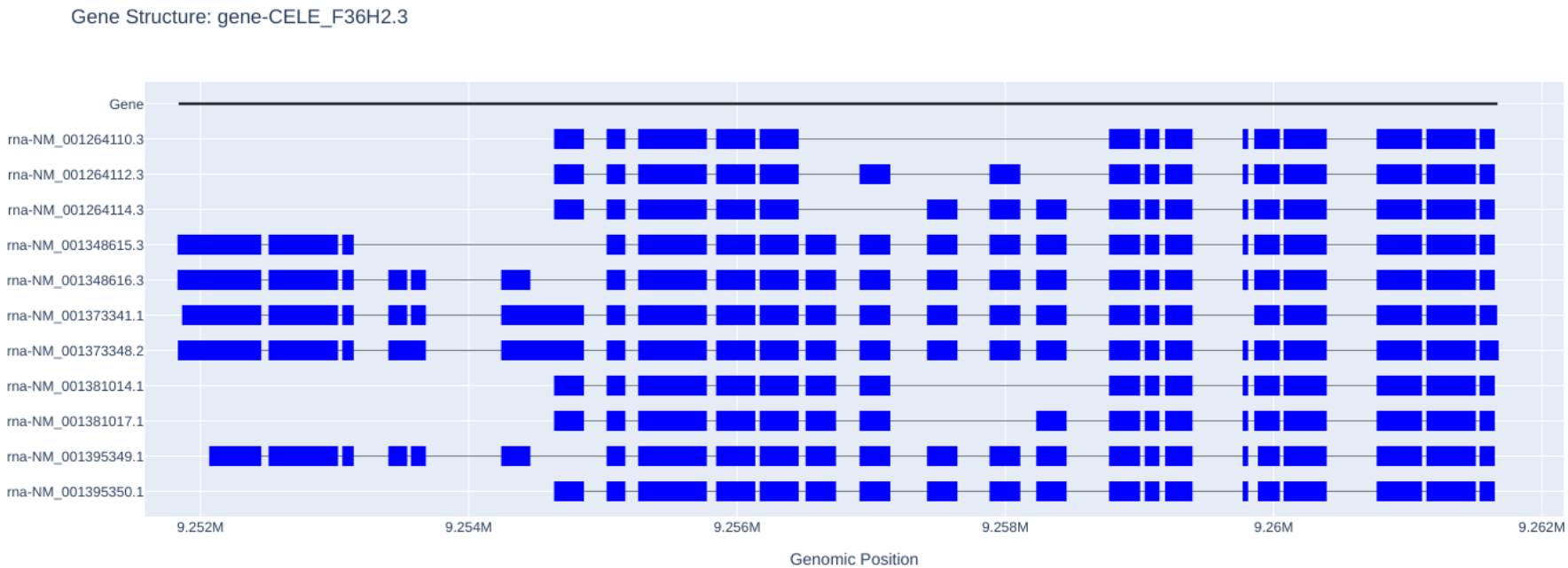

Figure 2: Visualization of *C. elegans* genomic gene structure. The plot illustrates the exon-intron organization of gene gene-CELE_F36H2.3 and its associated isoforms along genomic coordinates (x-axis). Individual isoforms are arranged vertically. Exons are represented by blue rectangular blocks, while intronssplice junctions are depicted as thin grey connecting lines. The vertical stacking facilitates the identification of alternative splicing events by comparing exon alignment across transcripts.

AS Frequency: Splice Graph Construction

Splice graph: $\mathcal{G} = (V, E)$

- $V = V_{\text{positions}} \cup \{\text{START}, \text{END}\}$: genomic coordinates + virtual nodes
- Edges represent exons (type 'E') and junctions (type 'J')
- **Edge weight:** $w(e) = |T(e)|$ = number of transcripts supporting edge e

For multi-transcript genes ($|T_g| > 1$):

- Construct $\mathcal{G}(g)$ from all transcripts in T_g for gene g
- Accumulate edge weights across shared paths

Theoretical AS Frequency Inference: Splice Graph Visualization

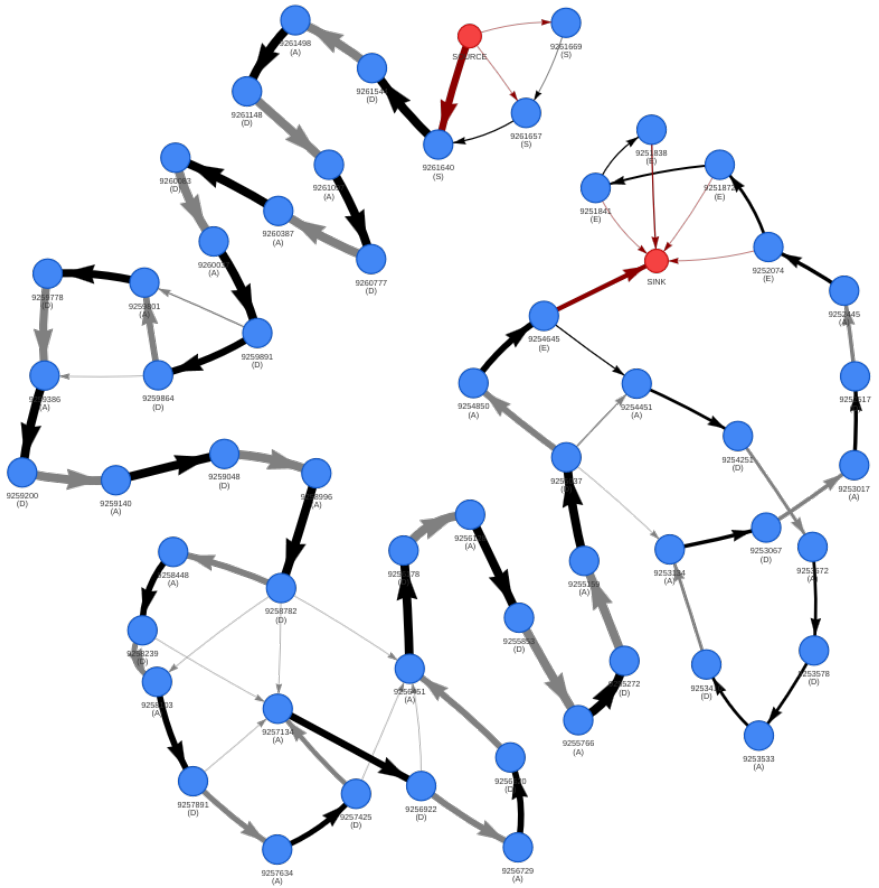

Figure 3: Splice graph visualization of *C. elegans* gene-CELE_F36H2.3 structure. Vertices: represent genomic splice sites labeled with coordinates and types. Red vertices indicate virtual Start and End points used to anchor the graph traversal. Edges: represent structural connections, color-coded by feature key: black arrows denote exonic regions (type E), gray arrows denote splice junctions (type J), and dark red arrows represent virtual start/end connections. Weights: The thickness of an edge is proportional to its weight, representing the number of transcripts supporting that specific connection.

Dominant Exon Chain & AS Event Detection

Dominant exon chain π^* (most common path):

$$\pi^* = \max_{\pi \in \mathcal{P}(\mathcal{G})} W(\pi) = \max_{\pi} \sum_{e \in \pi} w(e)$$

Alternative splicing event detection:

- For each transcript t, extract junction chain $\mathcal{J}(t)$
- Compare $\mathcal{J}(t)$ to π^* in subgraph $\mathcal{G}^{(\pi^*, \mathcal{J})}(t)$ by traversal
- Identify **branch points** ($B = \{v : \text{outdegree}(v) > 1\}$)
- Identify **convergence points** ($C = \{v : \text{indegree}(v) > 1\}$)

For each pair $(b, c) \in B \times C$**:**

- Extract exactly two paths: dominant $\pi_{\text{dec}}(b, c)$ and alternative $\pi_{\text{alt}}(b, c)$
- Classify event type via pattern matching

Event Frequency Aggregation

AS event categories: $\mathcal{E} = \{\text{ES}, \text{MES}, \text{RI}, \text{A5SS}, \text{A3SS}, \text{AFE}, \text{ALE}, \text{MXE}\}$

Event count for transcript t:

$$e_c(t) = |\{e \in \mathcal{E}(t) : \text{type}(e) = c\}|, \quad \forall c \in \mathcal{E}$$

Relative frequency for event type c:

$$rf_c = \frac{|\{t : e_c(t) \geq 1\}|}{N_t}$$

where N_t = total number of transcripts

Output: $\{rf_{\text{es}}, rf_{\text{mes}}, rf_{\text{ir}}, rf_{\text{a5}}, rf_{\text{a3}}, rf_{\text{mee}}, rf_{\text{afe}}, rf_{\text{ale}}\}$

Expected Outcomes

Outcome:

- Absolute Error per Transcript
  - Annotation completeness $>$ read accuracy
- Differential effects
  - AS event type
  - Tool

NRMSE Quantification Accuracy Across Experimental Scenarios

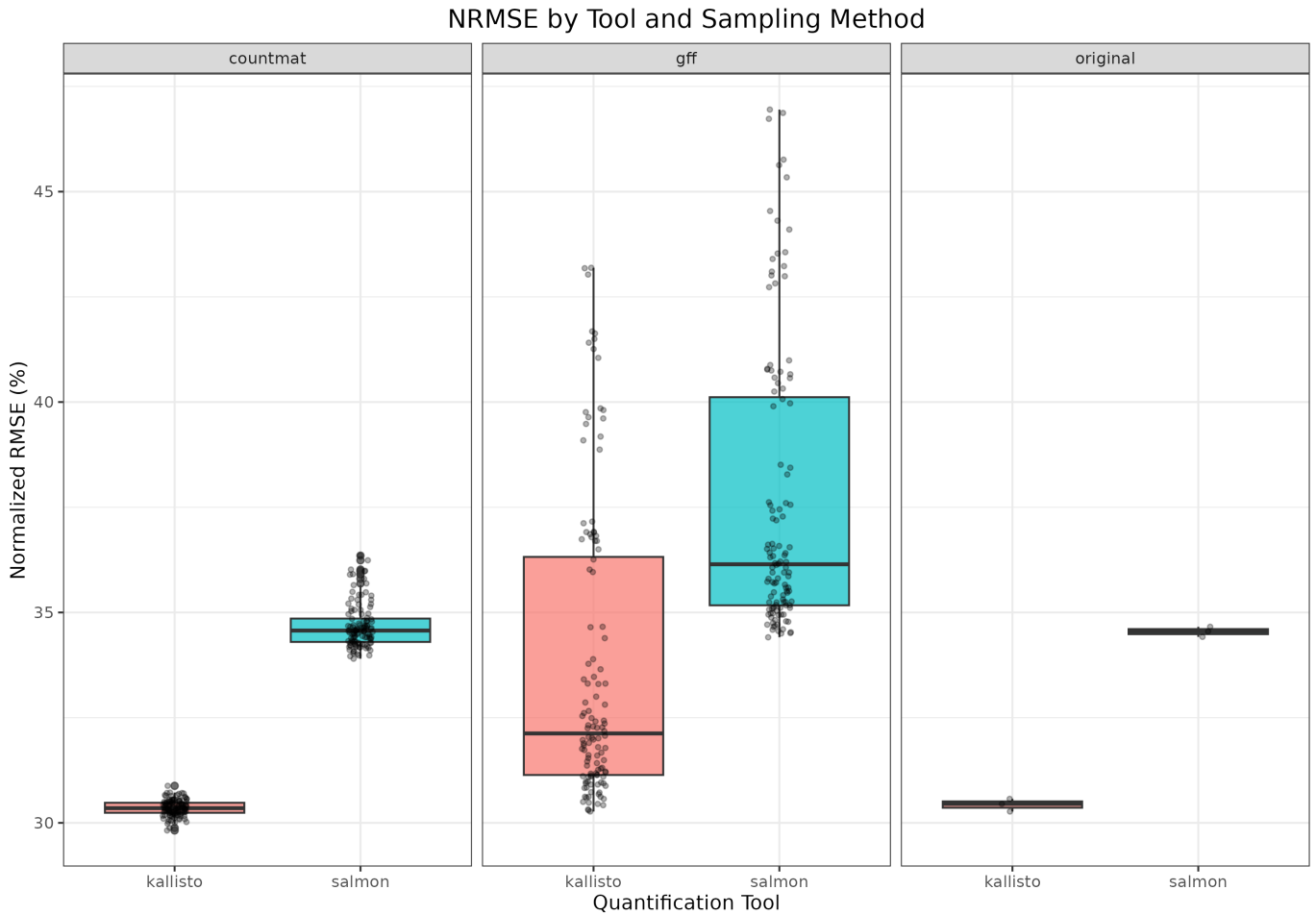

Figure 4: Normalized Root Mean Squared Error (NRMSE) distributions of transcript quantification accuracy. The boxplot displays error distributions across tools under three experimental manipulation scenarios: (1) Original (control), (2) Count-Matrix (false annotation), and (3) GFF (annotation incompleteness).

Three-Way Interaction Experimental Design

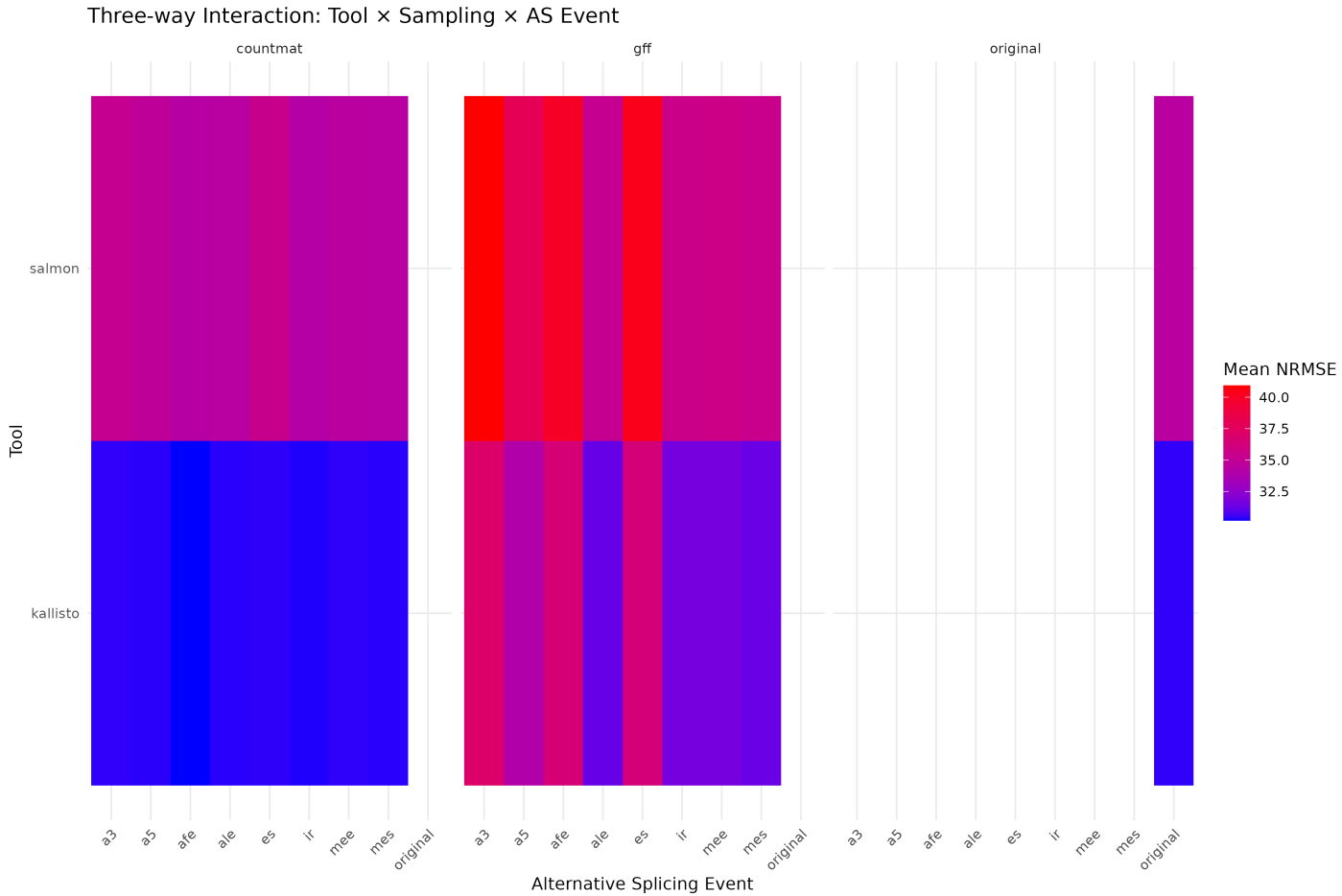

Figure 5: Three-way manipulation heatmap of Normalized Root Mean Squared Error (NRMSE) of transcript quantification accuracy across three scenarios: three experimental manipulation scenarios: (1) Original (control), (2) Count-Matrix (false annotation), and (3) GFF (annotation incompleteness).

Thank You!

References I

WanluLiuLab/yasim. WanluLiuLab, Aug. 2024.

B. K. Beaulieu-Jones and C. S. Greene. Reproducibility of computational workflows is automated using continuous analysis. *Nature Biotechnology*, 35(4):342–346, Apr. 2017. ISSN 1546-1696. doi: 10.1038/nbt.3780.

T. G. Brooks, N. F. Lahens, A. Mrčela, and G. R. Grant. Challenges and best practices in omics benchmarking. *Nature Reviews Genetics*, 25(5):326–339, May 2024a. ISSN 1471-0064. doi: 10.1038/s41576-023-00679-6.

T. G. Brooks, N. F. Lahens, A. Mrčela, D. Sarantopoulou, S. Nayak, A. Naik, S. Sengupta, P. S. Choi, and G. R. Grant. BEERS2: RNA-Seq simulation through high fidelity in silico modeling. *Briefings in Bioinformatics*, 25(3):bbae164, May 2024b. ISSN 1477-4054. doi: 10.1093/bib/bbae164.

L. A. Corchete, E. A. Rojas, D. Alonso-López, J. De Las Rivas, N. C. Gutiérrez, and F. J. Burguillo. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Scientific Reports*, 10(1):19737, Nov. 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-76881-x.

T. Coxe, D. J. Burks, U. Singh, R. Mittler, and R. K. Azad. Benchmarking RNA-Seq Aligners at Base-Level and Junction Base-Level Resolution Using the Arabidopsis thaliana Genome. *Plants*, 13(5):582, Jan. 2024. ISSN 2223-7747. doi: 10.3390/plants13050582.

D. Deshpande, K. Chhugani, Y. Chang, A. Karlsberg, C. Loeffler, J. Zhang, A. Muszyńska, V. Munteanu, H. Yang, J. Rotman, L. Tao, B. Balliu, E. Tseng, E. Eskin, F. Zhao, P. Mohammadi, P. P. Łabaj, and S. Mangul. RNA-seq data science: From raw data to effective interpretation. *Frontiers in Genetics*, 14, Mar. 2023. ISSN 1664-8021. doi: 10.3389/fgene.2023.997383.

P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, Apr. 2017. ISSN 1546-1696. doi: 10.1038/nbt.3820.

A. C. Frazee, A. E. Jaffe, B. Langmead, and J. T. Leek. Polyester: Simulating RNA-Seq Datasets With Differential Transcript Expression, Dec. 2014.

Q. Manz, O. Tsoy, A. Fenn, J. Baumbach, U. Völker, M. List, and T. Kacprowski. ASimulatoR: Splice-aware RNA-Seq data simulation. *Bioinformatics*, 37(18):3008–3010, Sept. 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab142.

References II

F. J. Pardo-Palacios, D. Wang, F. Reese, M. Diekhans, S. Carbonell-Sala, B. Williams, J. E. Loveland, M. De María, M. S. Adams, G. Balderrama-Gutierrez, A. K. Behera, J. M. Gonzalez Martinez, T. Hunt, J. Lagarde, C. E. Liang, H. Li, M. J. Meade, D. A. Moraga Amador, A. D. Prjibelski, I. Birol, H. Bostan, A. M. Brooks, M. H. Çelik, Y. Chen, M. R. M. Du, C. Felton, J. Göke, S. Hafezqorani, R. Herwig, H. Kawaji, J. Lee, J.-L. Li, M. Lienhard, A. Mikheenko, D. Mulligan, K. M. Nip, M. Pertea, M. E. Ritchie, A. D. Sim, A. D. Tang, Y. K. Wan, C. Wang, B. Y. Wong, C. Yang, I. Barnes, A. E. Berry, S. Capella-Gutierrez, A. Cousineau, N. Dhillon, J. M. Fernandez-Gonzalez, L. Ferrández-Peral, N. Garcia-Reyero, S. Götz, C. Hernández-Ferrer, L. Kondratova, T. Liu, A. Martinez-Martin, C. Menor, J. Mestre-Tomás, J. M. Mudge, N. G. Panayotova, A. Paniagua, D. Repchevsky, X. Ren, E. Rouchka, B. Saint-John, E. Sapena, L. Sheynkman, M. L. Smith, M.-M. Suner, H. Takahashi, I. A. Youngworth, P. Carninci, N. D. Denslow, R. Guigó, M. E. Hunter, R. Maehr, Y. Shen, H. U. Tilgner, B. J. Wold, C. Vollmers, A. Frankish, K. F. Au, G. M. Sheynkman, A. Mortazavi, A. Conesa, and A. N. Brooks. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nature Methods*, 21(7):1349–1363, July 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02298-3.

E. Park, Z. Pan, Z. Zhang, L. Lin, and Y. Xing. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *American Journal of Human Genetics*, 102(1):11–26, Jan. 2018. ISSN 0002-9297. doi: 10.1016/j.ajhg.2017.11.002.

F. Pozo, L. Martinez-Gomez, T. A. Walsh, J. M. Rodriguez, T. Di Domenico, F. Abascal, J. Vazquez, and M. L. Tress. Assessing the functional relevance of splice isoforms. *NAR Genomics and Bioinformatics*, 3(2):lqab044, June 2021. ISSN 2631-9268. doi: 10.1093/nargab/lqab044.

I. Raplee, A. Evsikov, and C. Evsikova. Aligning the Aligners: Comparison of RNA Sequencing Data Alignment and Gene Expression Quantification Tools for Clinical Breast Cancer Research. *Journal of Personalized Medicine*, 9:18, Apr. 2019. doi: 10.3390/jpm9020018.

D. Sarantopoulou, T. G. Brooks, S. Nayak, A. Mrčela, N. F. Lahens, and G. R. Grant. Comparative evaluation of full-length isoform quantification from RNA-Seq. *BMC Bioinformatics*, 22(1):266, May 2021. ISSN 1471-2105. doi: 10.1186/s12859-021-04198-1.

Milestones

- **Complete**
  - Pipeline implementation: Nextflow workflow with three-way scenario strategy and basic error data analysis
  - Tool benchmarking: Pseudoalignment based quantifiers (Salmon, Kallisto), Alignment-Assembler-Quantifier (Ryūtō, Scallop, StringTie) (STAR, HISAT2) (Salmon) workflows and Alignment-Traditional (Featurecounts) counting workflow
- **Open**
  - Advanced data analysis: Correlation, statistical testing
- **Out of scope**
  - Experimental data based AS-frequency inference
  - AS-frequency module: Support for complex splicing patterns
  - Additional tools: Cufflinks, nrp, sailfish, eXpress, rsem, htseq, tophat, BbMap2
  - Long-read simulations

ASimulatoR

1. **Exon Superset Creation**
   - Merge all exons from annotated transcripts per gene

2. **Event Assignment to Genes**
   - Allocate events respecting exon count constraints
   - $n_g(c) = \lfloor rf_c \cdot N_g \rfloor$ (frequency mode)
     - ★ $n_g(c)$: number of genes assigned to event c
     - ★ rf_c: relative frequency for event c
     - ★ N_g: total genes available

3. **Event-to-Exon Mapping**
   - Recursively place events on non-overlapping exon blocks

4. **Variant Transcript Construction**
   - Apply splicing rules (es, mes, ir, a3, a5, mee, afe, ale)
   - Recalculate transcriptomic coordinates

Polyester Pipeline: Count Matrix to FASTQ

- **Fragment Generation Model**: Generate R_{tj} fragments per transcript
  - R_{tj}: Number of reads to simulate from transcript t in sample j
- **Fragment Length Distribution**: $L_{tj} \sim$ Empirical(logspline)
- **GC Content Bias**: Retain fragment with probability $\phi_j(g_f)$ where $g_f \in [0, 1]$
- **PCR Amplification Bias**: Duplicate with rate $p_{PCR} = 0.1$,
  $D_f \sim \text{Poisson}(\lambda = 1) + 1$
- **Strand Orientation**: Reverse complement with probability 0.5
- **Paired-end Reads**: Mate 1: $f[1 : R]$, Mate 2: $\text{rc}(f[L_f - R + 1 : L_f])$, $R = 100$
- **Illumina5 Error Model**: Position- and base-dependent errors $P_{b,k}(b'|b)$

Simulation Parameters

- AS-event rel-freq grouping: individual
- Sequencing depth: 200000
- Read length: 100
- Error model: illumina5
- Fragment length distribution: empirical
- Dup rate: 0.1
- Dup rate lambda: 1
- Adapter contamination: false
- Error rate (SNP): 0.1

Simplifying Assumptions in Simulation

- Complex AS-Events
- No alternative transcription start and transcription end