Supplementary information for "Benchmarking RNA-seq with the Quartet Reference Materials to establish Best Practices for Accurate Alternative Splicing Detection"

Content

1. Supplementary Methods

1.1 Construction of the Quartet junction reference datasets

Long-read sequencing data. Seven high-quality batches of long-read RNA-seq (lrRNA-seq) data from the Quartet samples were used to construct junction reference datasets. These included two batches of multiplexed arrays sequencing (MAS-seq) data generated using the PacBio platform, as well as two batches of direct RNA (dRNA) sequencing and three batches of PCR-cDNA sequencing generated using the Oxford Nanopore Technologies (ONT). Detailed descriptions of these data are available in our companion study[1].

lrRNA-seq data processing and mapping. For PCR-cDNA ONT sequencing data, primer sequences were first removed using Pychopper (v2.7.10) (https://github.com/epi2me-labs/pychopper), followed by Poly(A) tail trimming using the trim_isoseq_polyA script from the official PacBio GitHub repository. MAS-seq data were preprocessed using the standard Iso-Seq toolkit (v4.3.0) workflow (https://isoseq.how/), which includes primer removal with lima and Poly(A) tail trimming with isoseq refine. dRNA ONT sequencing data were used without preprocessing. All seven batches of long-read data were aligned to the GRCh38 genome assembly using Minimap2 (v2.28)[2], with gene annotations from Ensembl release-109. The ONT data was aligned with Minimap2 using the parameters -ax splice, -uf, and -k14. For the PacBio data, the alignment was performed using the parameters -ax splice:hq and -uf. Junctions were extracted from BAM files using the sjFromSAMcollapseUandM_inclOverlaps.awk script provided by STAR (v2.7.10b)[3].

Construction of the junction reference datasets. Building the reference datasets mainly involved three main steps: identifying high-confidence junctions, selecting candidate gene sets with sufficient coverage, and generating the final reference datasets (Figure S4).

(i) First, junctions were defined as high-confidence positives if, in each sample, at least two out of three replicates had ≥3 supporting reads, and this criterion was met in at least 5 out of the 7 lrRNA-seq batches. Conversely, junctions with no supporting reads in

any of the 3 replicates across all 7 batches were defined as high-confidence negative junctions. For novel junctions, only high-confidence positive junctions were retained.
(ii) Next, genes were selected as candidates for inclusion in the reference datasets if, in each sample, at least 2 out of 3 replicates had ≥3 supporting reads, and this criterion was met in at least 5 out of the 7 batches. Genes that did not meet this threshold and genes containing any low-confidence junctions were excluded.
(iii) Finally, junctions from the candidate gene sets were retained for inclusion in the final junction reference dataset. For annotated junctions, we ensured that all junctions within each gene were known. For novel junctions, only a reliable set of high-confidence positive junctions was included.

1.2 Annotation of features for junctions, isoforms, and alternative splicing (AS) events

Features of junctions. The length and exon number of isoforms containing each junction were directly extracted from the gene annotation (GTF). Read coverage for each junction was obtained from STAR-derived junction files (SJ.out.tab). Isoform coverage uniformity was calculated using the geneBody_coverage.py script from RSeQC (v5.0.4)[4]. For each isoform, 100 positions were uniformly sampled, and the coverage at each position was divided by the maximum sequencing depth. The proportion of positions with relative coverage greater than 0.75 was used as the final measure of isoform coverage uniformity. GC content was calculated by extracting the 50 bp upstream of the junction start and 50 bp downstream of the junction end, and computing the proportion of guanine and cytosine bases within this sequence. Mappability was assessed using GenMap (v1.3.0)[5] to compute genome-wide mappability scores for each base, and the mean mappability of the 50 bp flanking sequence was used as the junction's mappability measure.
Features of isoforms. Isoform length, the number of isoforms per gene, the number of exons per isoform, and exon length were directly extracted from the gene annotation (GTF). Isoform GC content was calculated by concatenating all exon sequences of the

isoform and computing the proportion of guanine and cytosine bases. Isoform expression levels were obtained directly as FPKM values from the respective quantification pipelines. The K value, which measures the complexity of exon-isoform structures for each gene, was defined in previous studies[6]. The calculation procedure was conducted as follows:

First, for a given gene harboring I distinct transcript isoforms and E exons, we constructed a binary matrix $A \in \{0,1\}^{I \times E}$. Each element a_{ie} of this matrix was assigned a value of 1 if the i-th isoform included the e-th exon, and 0 otherwise. This matrix A thus provides a complete and precise representation of the exon composition of all annotated transcript isoforms for the gene.

Subsequently, the singular value decomposition (SVD) of the matrix A was performed. From this decomposition, we obtained the singular values of A, denoted $\sigma 1, \sigma 2, \ldots, \sigma r$, where r is the rank of A. The maximum singular value, $\sigma_{max}(A)$, and the minimum non-zero singular value, $\sigma_{min}(A)$, were identified.

Finally, the K-value was calculated as the ratio of these two singular values:

$$\text{K value} = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \quad (1)$$

Features of AS events. The number of transcript isoforms associated with each AS event was directly obtained from the .ioe files generated by the generateEvents command in SUPPA2 (v2.4) [7]. The number of neighboring AS events was calculated by selecting 150 bp upstream and downstream of each event and counting all event types present within these regions. For sequence-based features, we extracted the 75 bp upstream and downstream of the relevant junctions for SE, A5, A3, MX, AF, and AL events, while the entire intronic region was used for RI events. GC content was computed as the proportion of guanine and cytosine bases in these event-specific sequences, and mappability was assessed using GenMap as described above.

1.3 Construction of a unified format for different AS event types

Skipping exon (SE). For SE events, four coordinates were used, including the start and end coordinates of the alternative exon, as well as the end coordinate of the upstream

exon and the start coordinate of the downstream exon. For the rMATS output, these four coordinates correspond to “exonStart_0base”, “exonEnd”, “upstreamEE”, and “downstreamES”, respectively. For the PSI-Sigma output, the coordinates of the alternative exon correspond to the “Target Exon,” while the end coordinate of the upstream exon and the start coordinate of the downstream exon are derived from the “Event Region” coordinates. In MAJIQ output files, SE events are divided into four local splicing variations (LSVs). For LSVs with junction names C1_C2 or C2_C1, the “junction_coord” corresponds to the end coordinate of the upstream exon and the start coordinate of the downstream exon. For LSVs with junction name C1_A, the “spliced_with_coord” corresponds to the start and end coordinates of the alternative exon. PSI represents the mean PSI value across all LSVs that support inclusion of the alternative exon. For the SUPPA2 output, the four coordinates were directly extracted from the “event_id” field.

Alternative 5' splice site (A5SS). For A5SS, three coordinates were used: the end coordinate of the alternative exon, the splice site coordinate on the alternative exon, and the start coordinate of the downstream exon. These three coordinates correspond to the junctions supporting inclusion or exclusion of the 5′ splice site. For the rMATS output, using + strand genes as an example, the three coordinates are “longExonEnd”, “shortEE”, and “flankingES”. For the PSI-Sigma output, using + strand genes as an example, the three coordinates correspond to the end coordinate of the “Target Exon”, and the start and end coordinates of the “Event Region”. For the MAJIQ output, an A5SS event is represented by two LSVs. The “junction_coord” with junction name *Proximal* corresponds to the splice site of the alternative exon and the start coordinate of the downstream exon, while the “junction_coord” with junction name *Distal* corresponds to the end coordinate of the alternative exon and the start coordinate of the downstream exon. The LSV with junction name *Proximal* represents the isoform that includes the alternative exon, and its PSI value is used. For the SUPPA2 output, the three coordinates were directly extracted from the “event_id” field.

Alternative 3' splice site (A3SS). For A3SS, similar to A5SS, three coordinates are

used to define the event, essentially mirroring the strand orientation. The outputs from the different tools are processed in the same way as described above.

Alternative first exon (AF). For AF, two alternative exons are involved, with the one at the start of the isoform designated as alternative exon 1 and the other as alternative exon 2. AF is defined using five coordinates: the start and end coordinates of alternative exon 1, the start and end coordinates of alternative exon 2, and the start coordinate of the downstream exon. rMATS and PSI-Sigma do not detect this event type. For the MAJIQ output, AF events are also represented as two LSVs. The "spliced_with_coord" of the junction with the name *Proximal* corresponds to the start and end coordinates of alternative exon 1, while that of the junction with the name *Distal* corresponds to the start and end coordinates of alternative exon 2. In addition, the start coordinate of the downstream exon can be extracted from the "junction_coord". The LSV with junction name *Distal* represents the isoform that includes the alternative exon 1, and its PSI value is used. For the SUPPA2 output, the five coordinates were directly extracted from the "event_id" field.

Alternative last exon (AL). For AL, similar to AF, five coordinates are used to define the event, essentially mirroring the strand orientation. The outputs from the different tools are processed in the same way as described above.

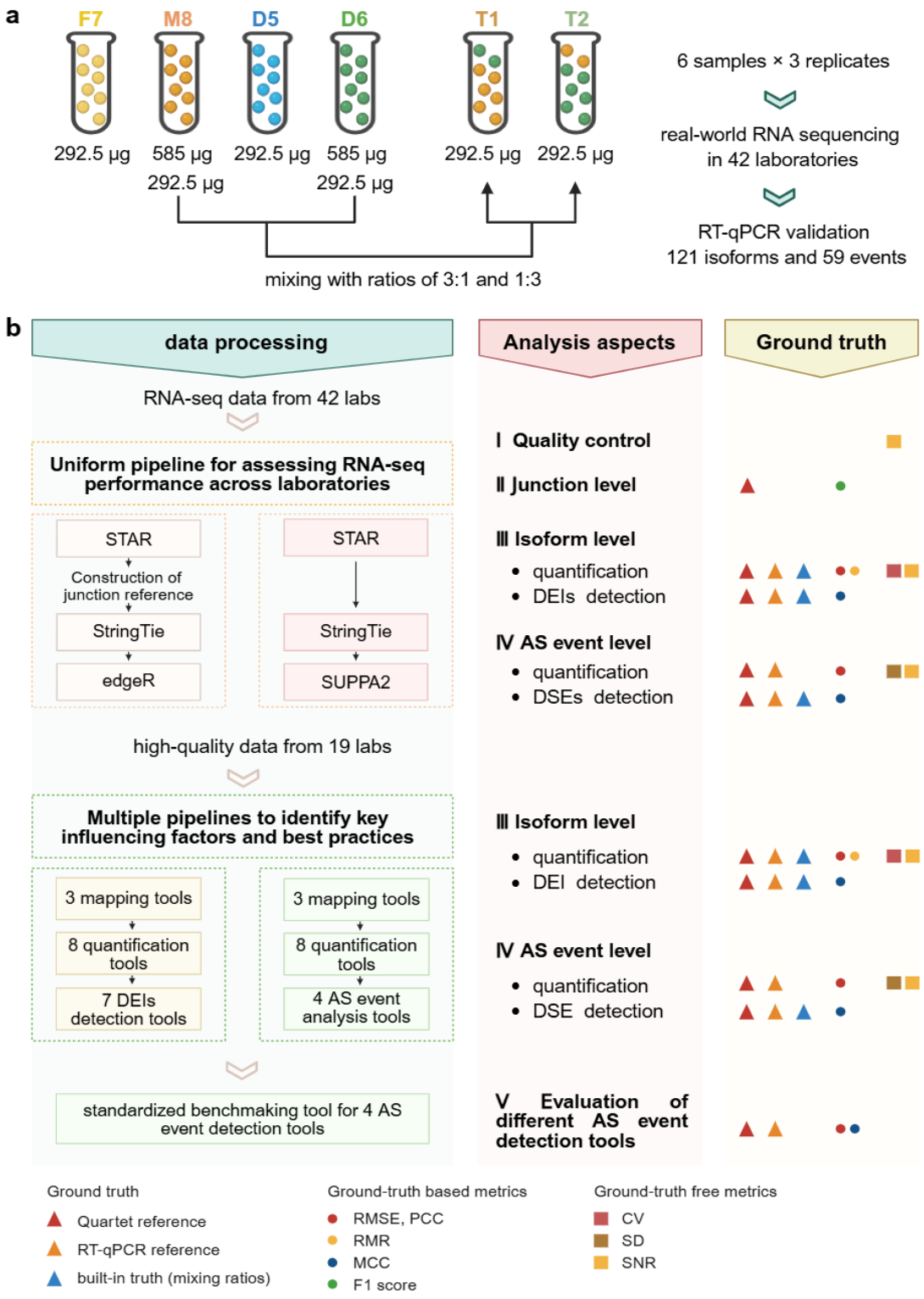
Mutually exclusive exon (MX). MX events involve two alternative exons. If the strand is +, the 5′ exon is defined as alternative exon 1; if the strand is -, the 3′ exon is defined as alternative exon 1. MX is defined using six coordinates: the start and end coordinates of alternative exon 1, the start and end coordinates of alternative exon 2, the end coordinate of the upstream exon, and the start coordinate of the downstream exon. For rMATS output, using the + strand as an example, the six coordinates correspond to "1stExonStart_0base", "1stExonEnd", "2ndExonStart_0base", "2ndExonEnd", "upstreamEE", and "downstreamES". For PSI-Sigma output, MX events are split into two separate events corresponding to the two alternative exons. The coordinates of the "Target Exon" represent the alternative exons, while the "Event Region" spans the end coordinate of the upstream exon and the start coordinate of the downstream exon. The

PSI value of the first alternative exon is used for analysis. For MAJIQ output, MX events are represented by four LSVs with junction names *C1_A1, C2_A2, C2_A1, and C2_A2*. The "spliced_with_coord" of LSVs with "spliced_with" equal to *A1* and *A2* correspond to the coordinates of alternative exon 1 and alternative exon 2, respectively. Additionally, the "junction_coord" of *C1_A1* and *C2_A2* contains the end coordinate of the upstream exon and the start coordinate of the downstream exon. The PSI value of the LSV with "spliced_with" equal to *A1* is used as the PSI for the MX event. For the SUPPA2 output, the six coordinates were directly extracted from the "event_id" field.

Retained intron (RI). For RI events, four coordinates are used, including the start and end of the retained intron, as well as the start coordinate of the upst ream exon and the end coordinate of the downstream exon. For rMATS output, these correspond to "upstreamES", "upstreamEE", "downstreamES", and "down streamEE". For PSI-Sigma output, the "Target Exon" corresponds to the coordi nates of the retained intron, while the "Event Region" includes the end coordin ate of the upstream exon and the start coordinate of the downstream exon. For MAJIQ output, RI events are represented by two LSVs, with "junction_name" as *C1_C2_intron/C2_C1_intron* and *C1_C2_spliced/C2_C1_spliced*. The "juncti on_coord" of *C1_C2_intron/C2_C1_intron* corresponds to the coordinates of the retained intron, while "spliced_with_coord" and "reference_exon_coord" provide the start coordinate of the upstream exon and the end coordinate of the downst ream exon. The PSI of the RI event is taken from the LSV with "junction_na me" *C1_C2_intron/C2_C1_intron*. For the SUPPA2 output, the four coordinates were directly extracted from the "event_id" field.

2. Supplementary Figures

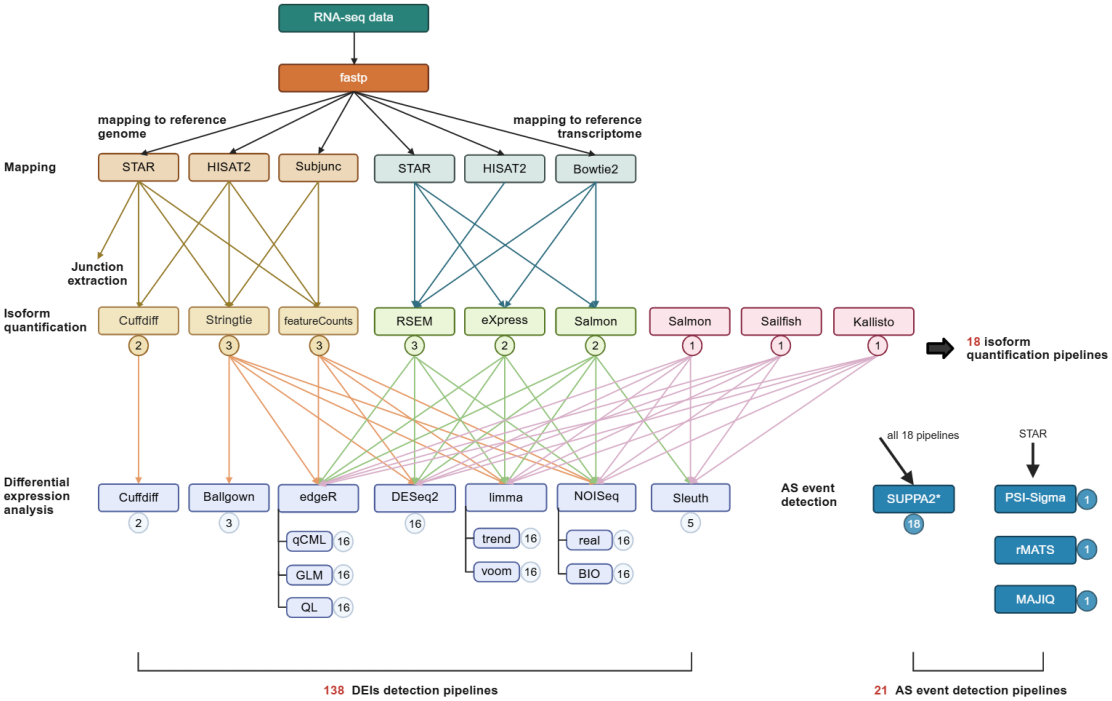
Figure S1. Experimental and analytical design of this study.



a, Preparation and mixing scheme of the Quartet RNA reference materials. **b,** Overview of data analysis strategies and their objectives. A unified isoform and alternative

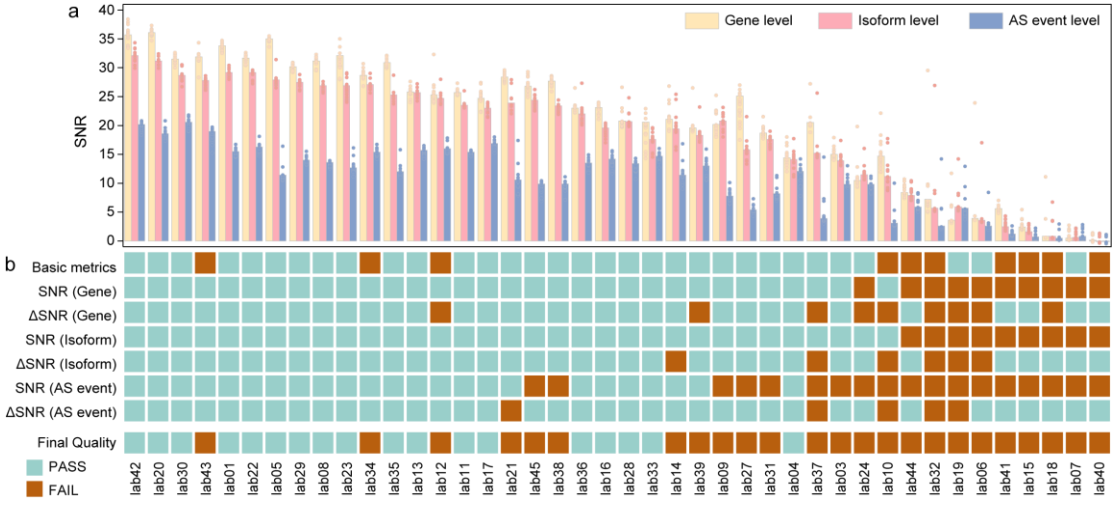
splicing (AS) event analysis pipeline was first applied to all RNA-seq data from 42 laboratories to evaluate overall detection performance. Based on signal-to-noise ratio (SNR) and other quality metrics, high-quality laboratories were selected for further analysis. Multiple combinations of mapping tools, isoform quantification tools, differentially expressed isoforms (DEIs) detection methods, and AS event callers were included to investigate the influence of experimental procedures, analysis pipelines, and intrinsic features of isoforms and AS events. This allowed identification of best practices for isoform and AS event analysis. Finally, a standardized benchmarking tool was developed specifically for evaluating widely used AS event detection tools, including SUPPA2 (v2.4), rMATS (v4.3.0) [8], MAJIQ (v2.5.6)[9], and PSI-Sigma (v2.3)[10]. Throughout the study, three types of ground truth were employed, complemented by both ground-truth based and ground-truth free performance metrics to comprehensively assess RNA-seq performance in AS profiling. RMSE, root mean square error; PCC, Pearson correlation coefficient; MCC, Matthews correlation coefficient; RMR, recovery of mixing ratios; CV, coefficient of variation; SD, standard deviation. Created in BioRender. Zhang, R. (2025) https://BioRender.com/tgay6bs.

Figure S2. Overview of bioinformatics design.



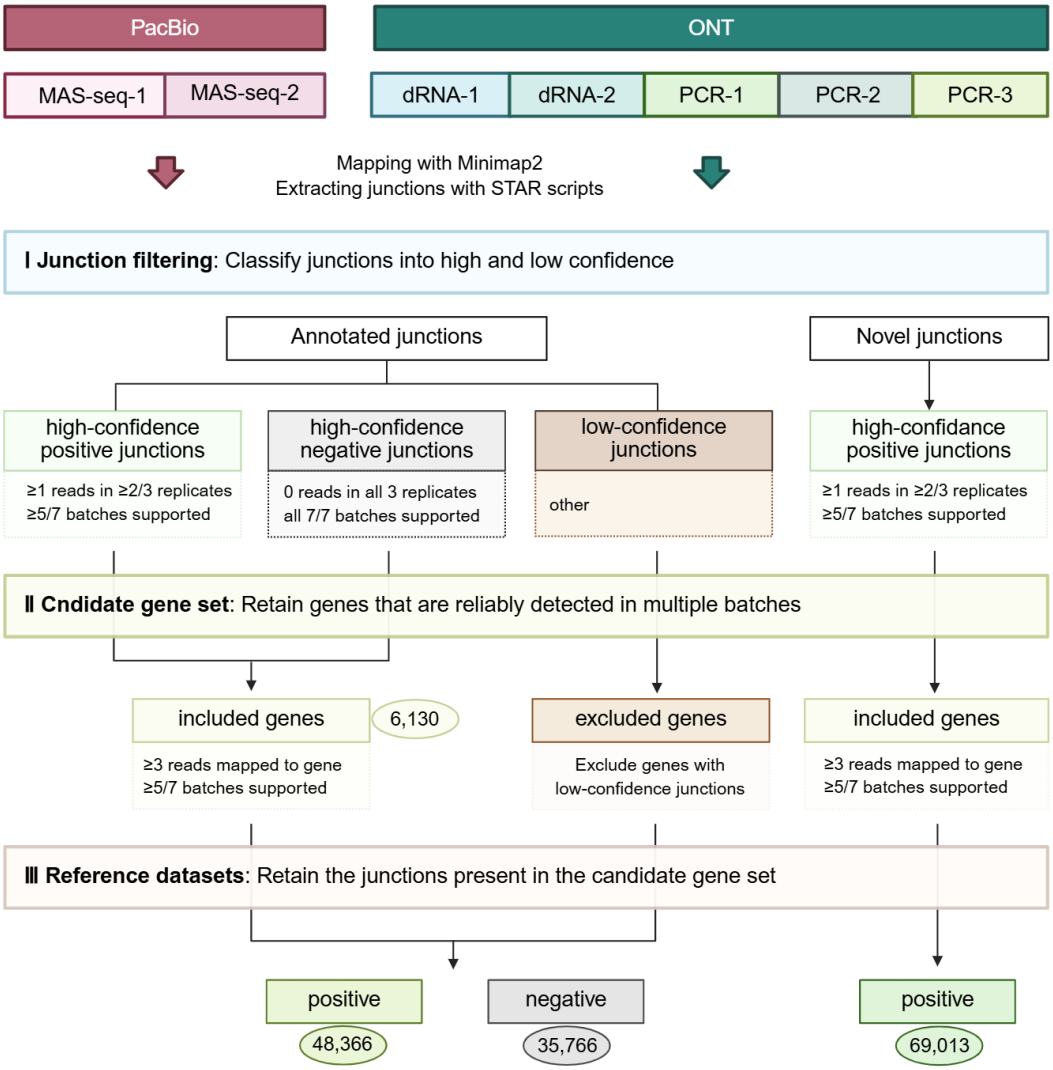
RNA-seq data analysis was performed using four mapping tools, eight isoform quantification tools, seven DEI detection tools, and four AS event detection tools, resulting in a total of 159 distinct analysis pipelines. Created in BioRender. Zhang, R. (2025) https://BioRender.com/lststz9.

Figure S3. Quality control of RNA-seq data.



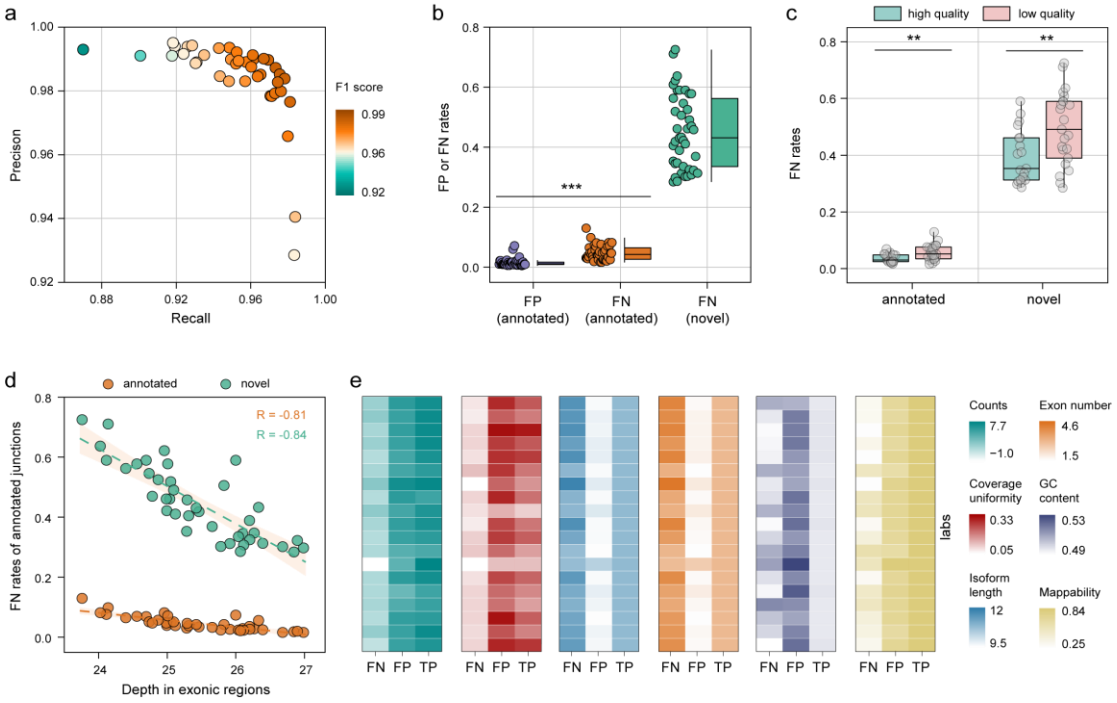
a, Signal-to-noise ratio (SNR) values to measure the quality of quantification data at gene, isoform, and AS event levels for 42 laboratories based on the Quartet and mixed samples (18 samples). Dots represented SNR values based on any 17 of the 18 samples. The red dashed line represents the SNR cutoff of 10. **b**, Multiple quality metrics were combined to assign quality tags for each laboratory. The basic metrics included the number of sequencing reads, base quality (Q30), mapping rate, gene body bias (5′–3′ bias), the percentage of mapped reads in the intergenic region, duplicate rates, cross-contamination, and sample-identity check (Supplementary Data 1). ΔSNR represents the difference between the SNR computed from any subset of 17 out of 18 samples and that computed from all 18 samples, serving to identify low-quality outliers.

Figure S4. Stepwise workflow for constructing junction reference datasets.



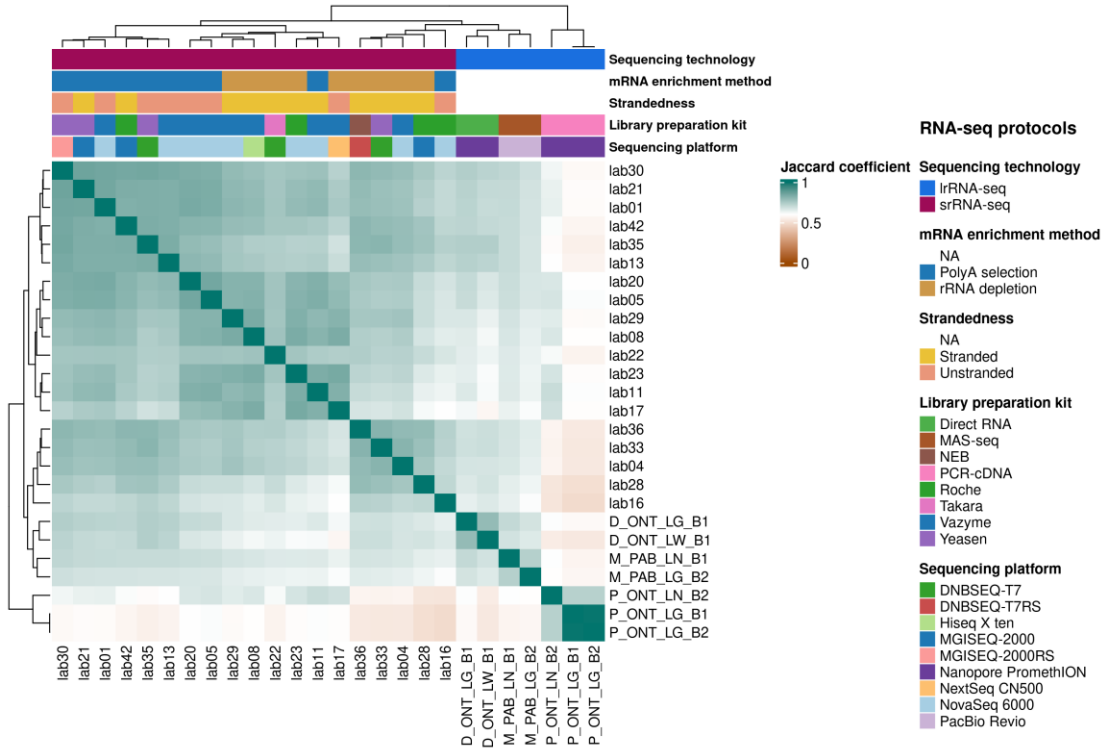
Seven batches of high-quality long-read sequencing data from our companion study were used to construct reference datasets (details in Supplementary Methods). First, positive and negative junctions supported by long-read sequencing data across multiple batches were selected. Second, a candidate gene set with sufficient read coverage was identified. Finally, positive and negative junctions located within the candidate gene set were further screened, resulting in a reference dataset in which all annotated positive and negative junctions within these genes were known, as well as an additional dataset containing novel junctions. Created in BioRender. Zhang, R. (2025) https://BioRender.com/5f2hh7w.

Figure S5. Evaluation of junction-level detection performance of RNA-seq.



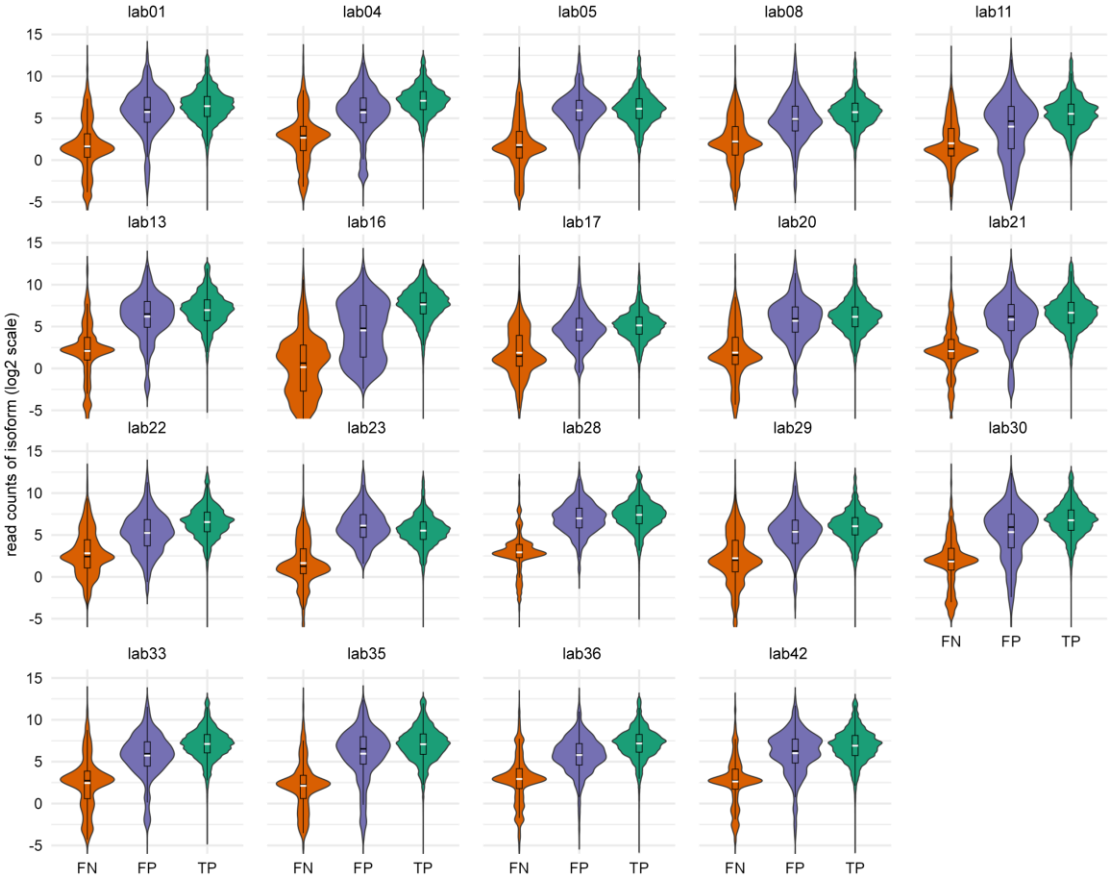
a, Recall, precision and F1 score of junction detection across 42 laboratories based on the junction reference datasets. **b**, Comparison of false positive (FP) and false negative (FN) rates of annotated and novel junctions across 42 laboratories. Significance testing among groups was conducted using paired t-tests. **c**, Comparison of FN rates of annotated and novel junctions between high-quality and low quality laboratories. Data are presented as median values (center lines) and the upper and lower quartiles (box limits). Significance testing among groups was conducted using Mann-Whitney U test. **d**, Correlation between FN rates of annotated and novel junctions and the sequencing depth in exonic regions. **e**, Comparison of several intrinsic features among true positive (TP), FP, and FN junctions, including read counts, coverage uniformity, isoform length and exon number of the associated isoforms, as well as GC content and mappability in the regions flanking the junctions. ** indicates a two-sided *p-value* < 0.01, and *** indicates a two-sided *p-value* < 0.001.

Figure S6. Pairwise consistency in junction detection across short- and long-read RNA-seq.



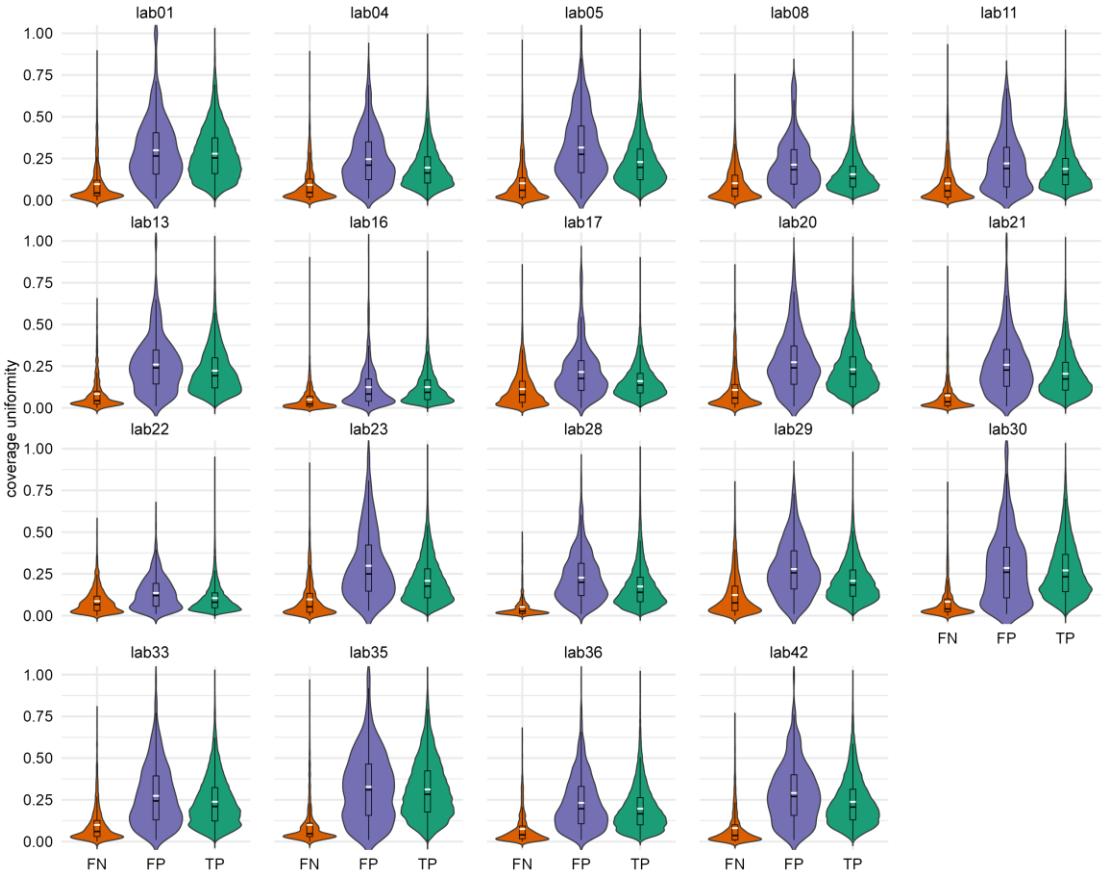
Heatmap showing the Jaccard coefficients for junction detection results between each pair of 19 high-quality short-read RNA-seq batches from this study and 7 high-quality long-read RNA-seq batches from our companion study. Only junctions supported by ≥3 reads in at least two out of three replicates were included in the Jaccard coefficient calculation. Laboratories employing short-read RNA-seq clustered mainly according to the mRNA enrichment methods and strandedness, suggesting these are key factors influencing detection performance.

Figure S7. Read counts distribution of isoforms associated with FN, FP, and TP junctions.



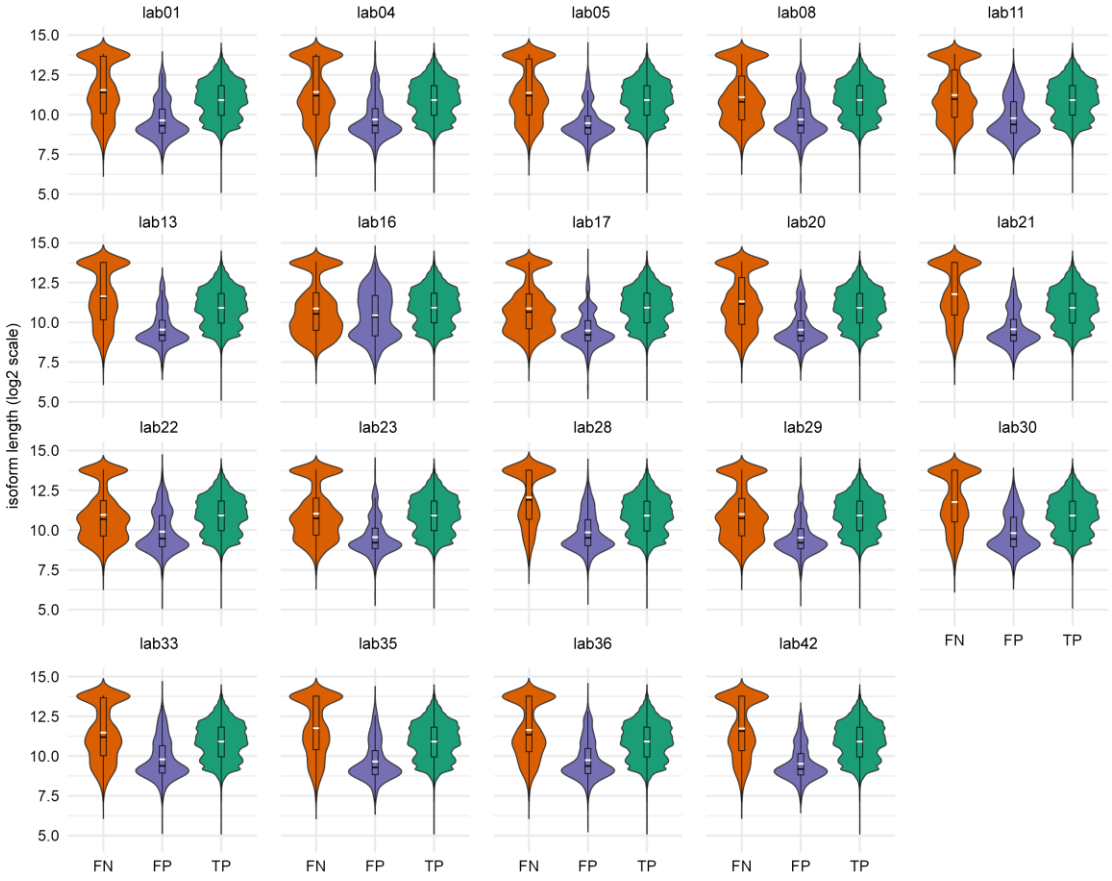
Violin plots illustrate the distribution of sequencing read counts for isoforms corresponding to annotated junctions flagged as FN, FP, and TP in 19 high-quality laboratories. Embedded boxplots indicate the mean (white line), median (black line), and interquartile range (box limits) of the read count distribution.

Figure S8. Coverage uniformity distribution of isoforms associated with FN, FP, and TP junctions.



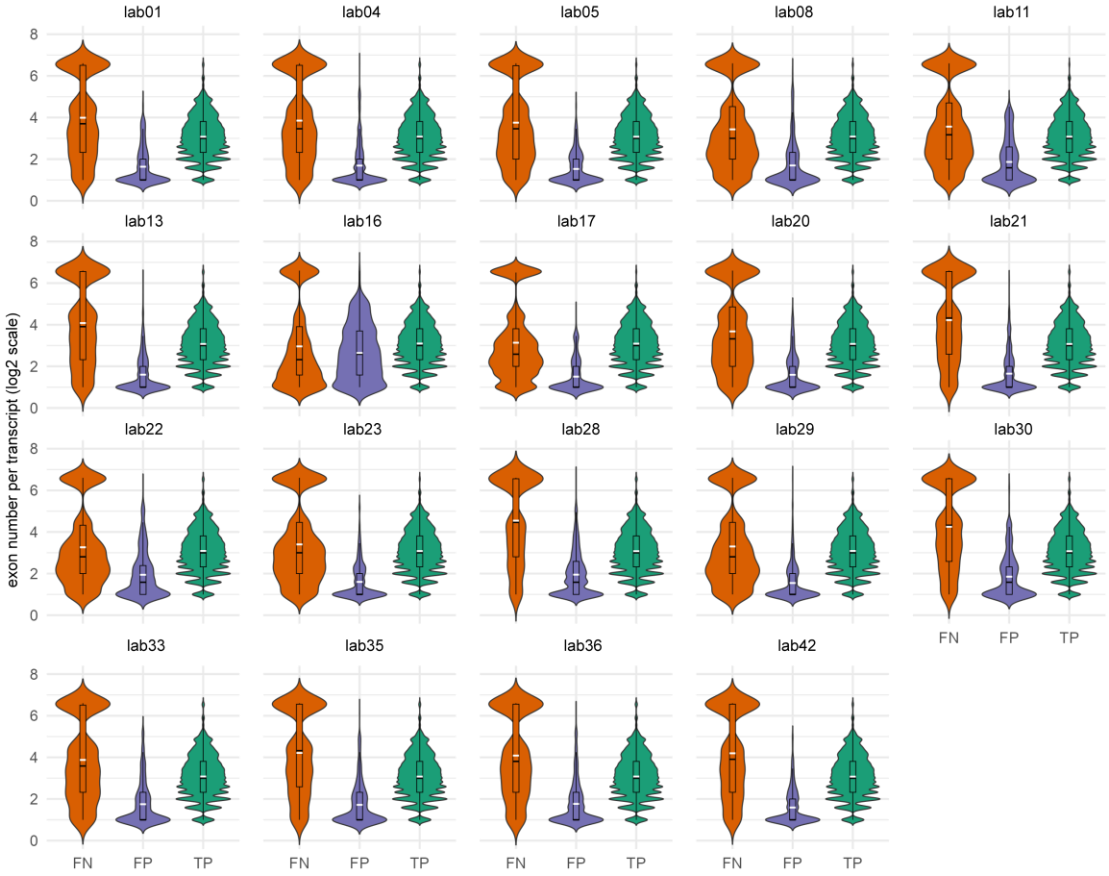
Violin plots illustrate the coverage uniformity distribution for isoforms corresponding to annotated junctions flagged as FN, FP, and TP in 19 high-quality laboratories. Coverage uniformity for each isoform was defined as the proportion of bases with coverage depth greater than 75% of the maximum. The embedded boxplots depict the mean (white line), median (black line), and the interquartile range (box limits) of the coverage uniformity distribution.

Figure S9. Length distribution of isoforms associated with FN, FP, and TP junctions.



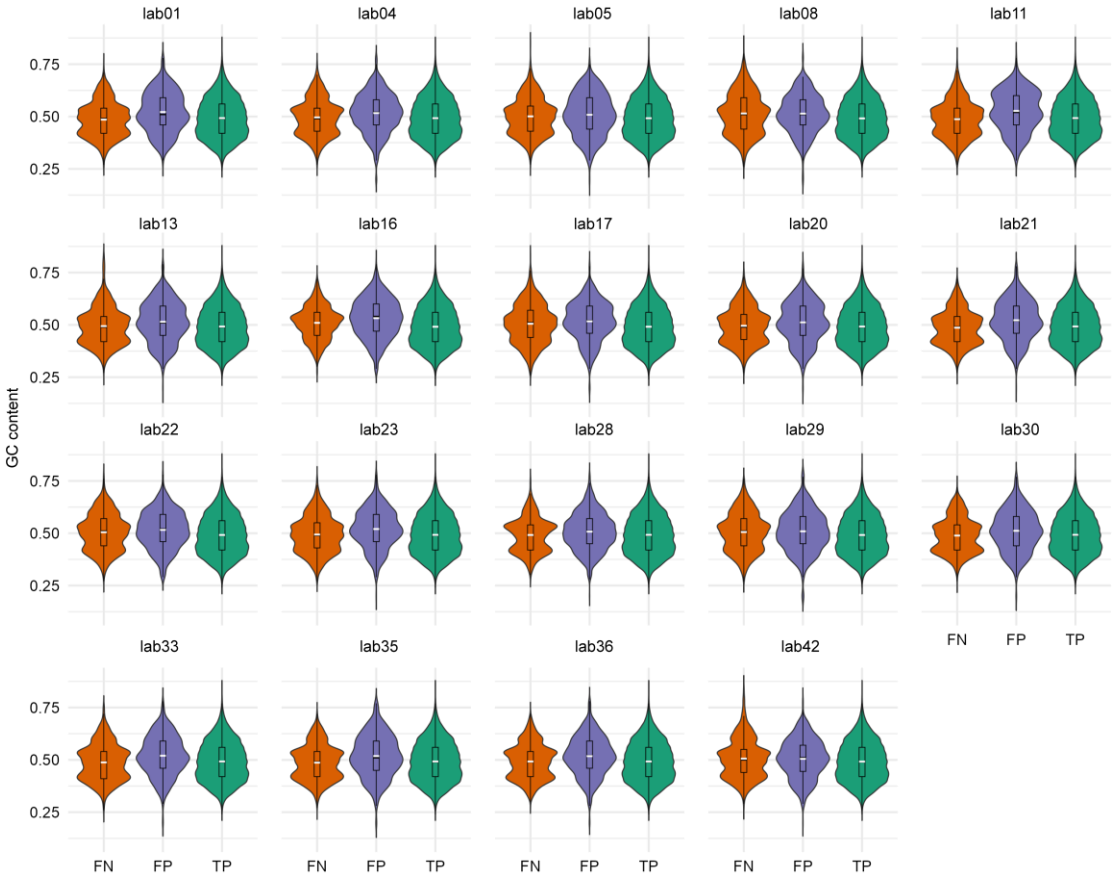
Violin plots illustrate the length distribution for isoforms corresponding to annotated junctions flagged as FN, FP, and TP in 19 high-quality laboratories. The embedded boxplots depict the mean (white line), median (black line), and the interquartile range (box limits) of the length distribution.

Figure S10. Exon count distribution of isoforms associated with FN, FP, and TP junctions.



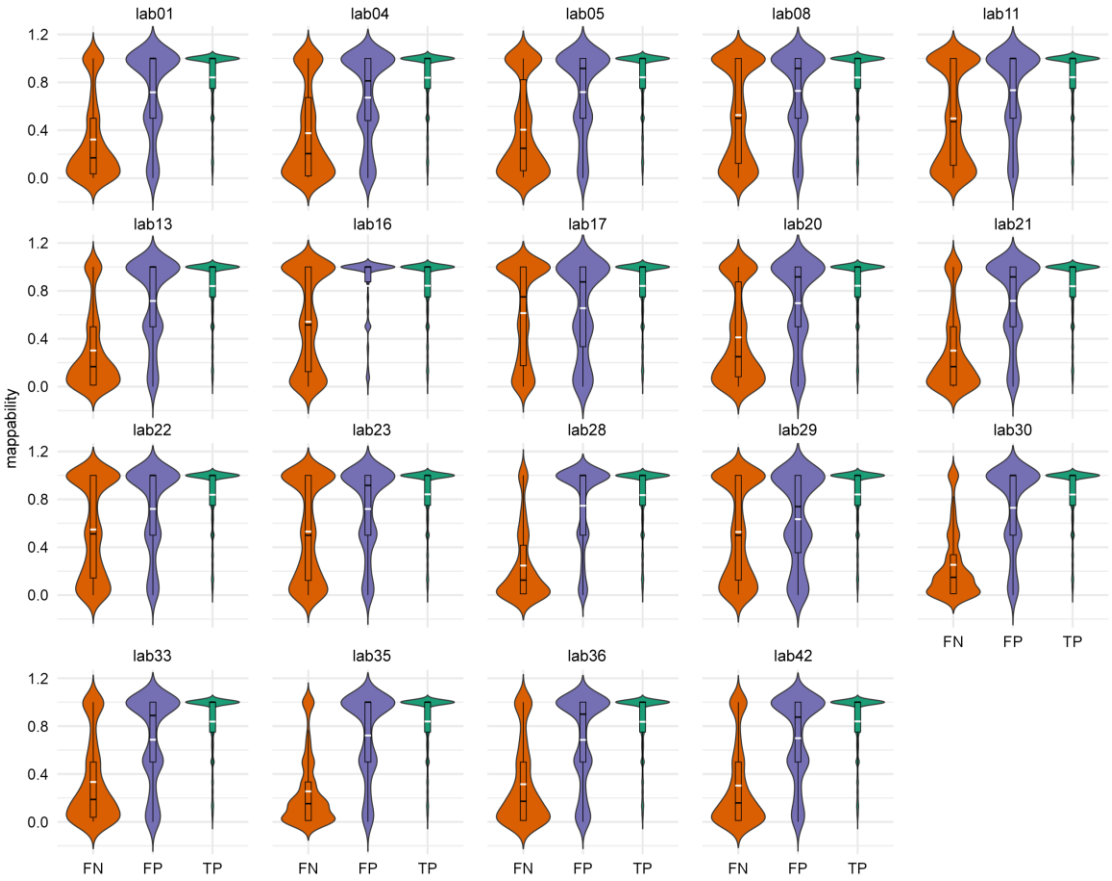
Violin plots illustrate the exon number distribution of annotated junctions flagged as FN, FP, and TP in 19 high-quality laboratories. The embedded boxplots depict the mean (white line), median (black line), and the interquartile range (box limits) of the exon number distribution.

Figure S11. GC content distribution of FN, FP, and TP junctions.



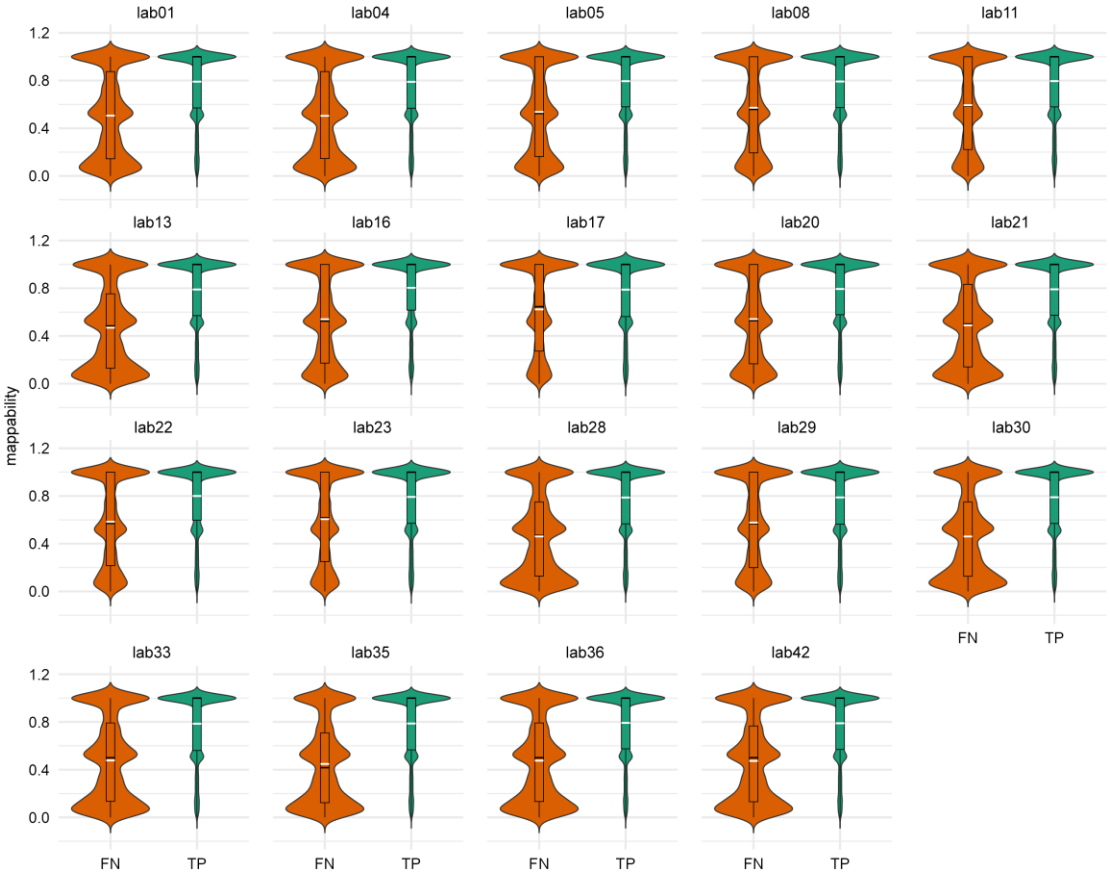
Violin plots illustrate the GC content distribution of annotated junctions flagged as FN, FP, and TP in 19 high-quality laboratories. GC content was calculated using a 100 bp window flanking each junction, comprising 50 bp upstream of the start coordinate and 50 bp downstream of the end coordinate. The embedded boxplots show the mean (white line), median (black line), and interquartile range (box limits) of the GC content.

Figure S12. Mappability distribution of FN, FP, and TP junctions.



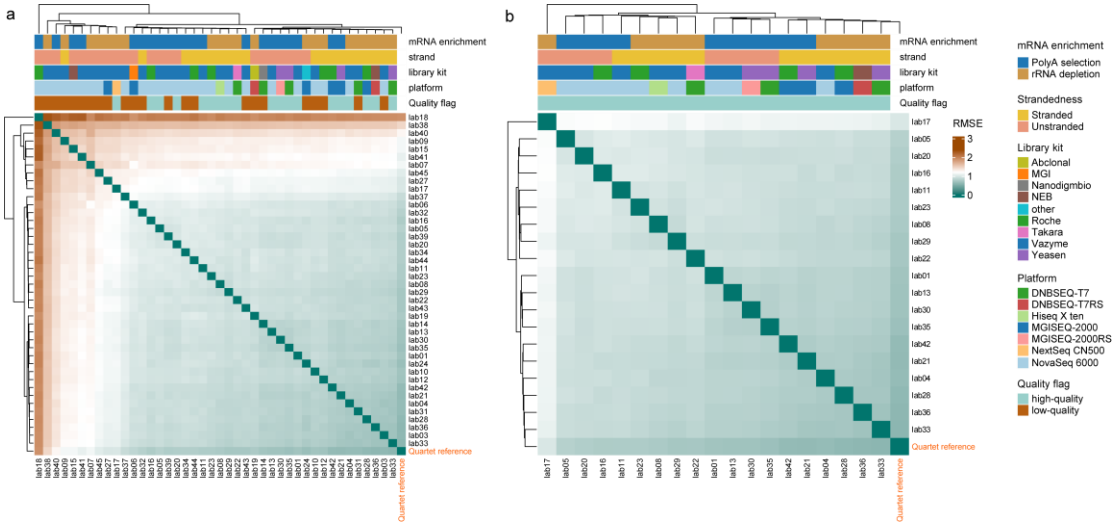
Violin plots illustrate the mappability distribution of annotated junctions flagged as FN, FP, and TP in 19 high-quality laboratories. Mappability was calculated using GenMap (v1.3.0) [5] in a 50 bp window flanking each junction, comprising 25 bp upstream of the start coordinate and 25 bp downstream of the end coordinate. The embedded boxplots depict the mean (white line), median (black line), and the interquartile range (box limits) of the mappability distribution.

Figure S13. Mappability distribution of novel junctions flagged as FN and TP.



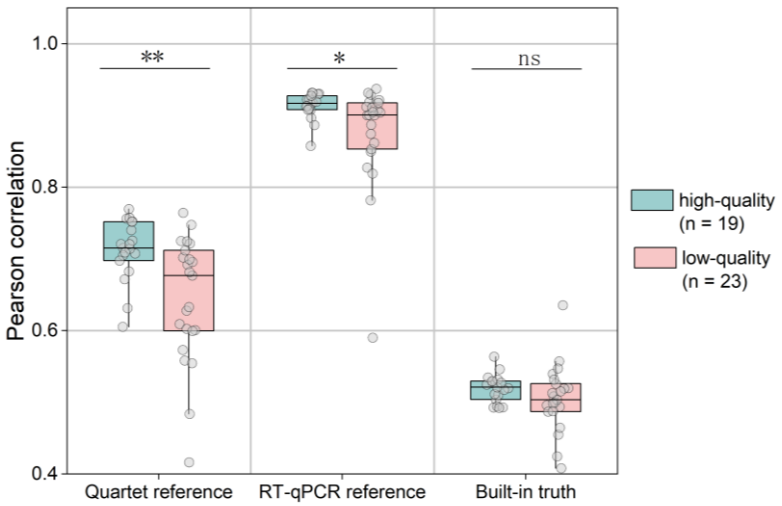
Violin plots illustrate the mappability distribution of novel junctions flagged as FN and TP in 19 high-quality laboratories. Mappability was calculated using GenMap (v1.3.0) [5] in a 50 bp window flanking each junction, comprising 25 bp upstream of the start coordinate and 25 bp downstream of the end coordinate. The embedded boxplots depict the mean (white line), median (black line), and the interquartile range (box limits) of the mappability distribution.

Figure S14. Pairwise consistency in isoform quantification across laboratories.



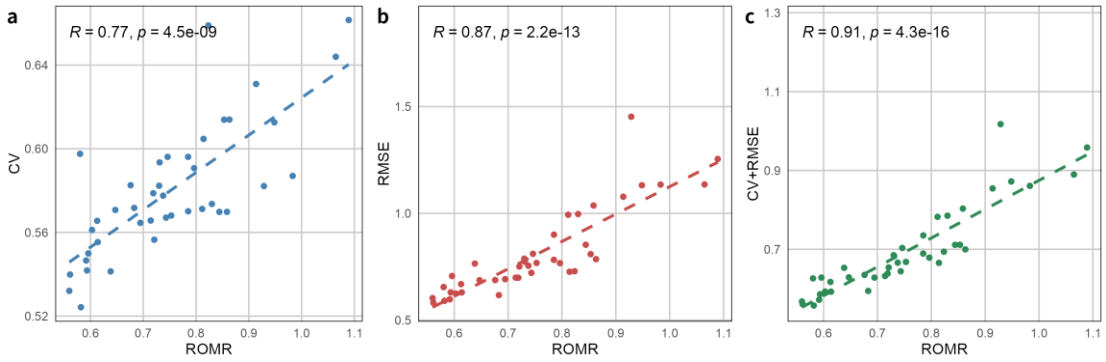
a, Heatmap showing the root mean square error (RMSE) of isoform-level log2FC values between each pair of the 42 participating laboratories. **b**, Heatmap showing RMSE values for log2FC comparisons among the 19 high-quality laboratories. Rows and columns are clustered based on hierarchical clustering of RMSE values.

Figure S15. Impact of data quality on isoform quantification accuracy.



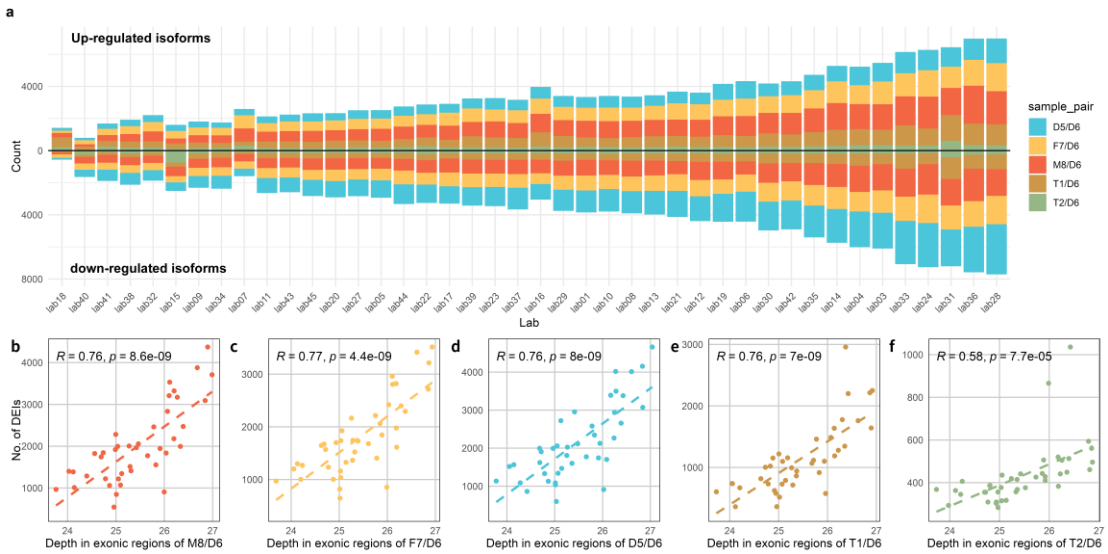
Isoform quantification accuracy was assessed using the Pearson correlation coefficient. Statistical significance was evaluated using the Mann–Whitney U test. ** and * indicate *p-values* < 0.001 and < 0.05, respectively; ns, not significant.

Figure S16. Comparison of three performance metrics.



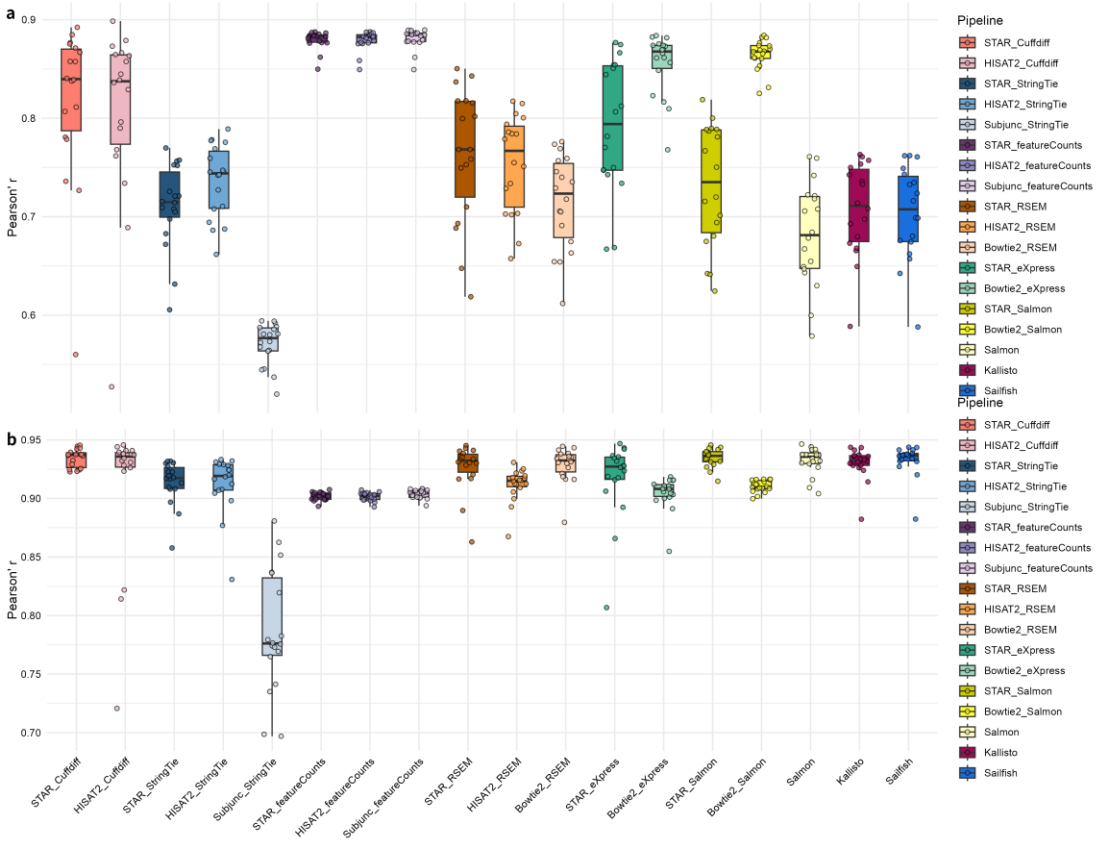
ROMR (Recovery of Mixing Ratios) measures how well RNA-seq expression data reflect the known mixing ratios of mixed samples. Correlation of ROMR with (**a**) coefficient of variation (CV), (**b**) root mean square error (RMSE), and (**c**) the average of CV and RMSE. ROMR correlates most strongly with the combined accuracy and reproducibility metric (average of CV and RMSE), suggesting its utility as an integrated performance indicator. *R* represents the Pearson correlation coefficient.

Figure S17. Number of DEIs detected by laboratories.



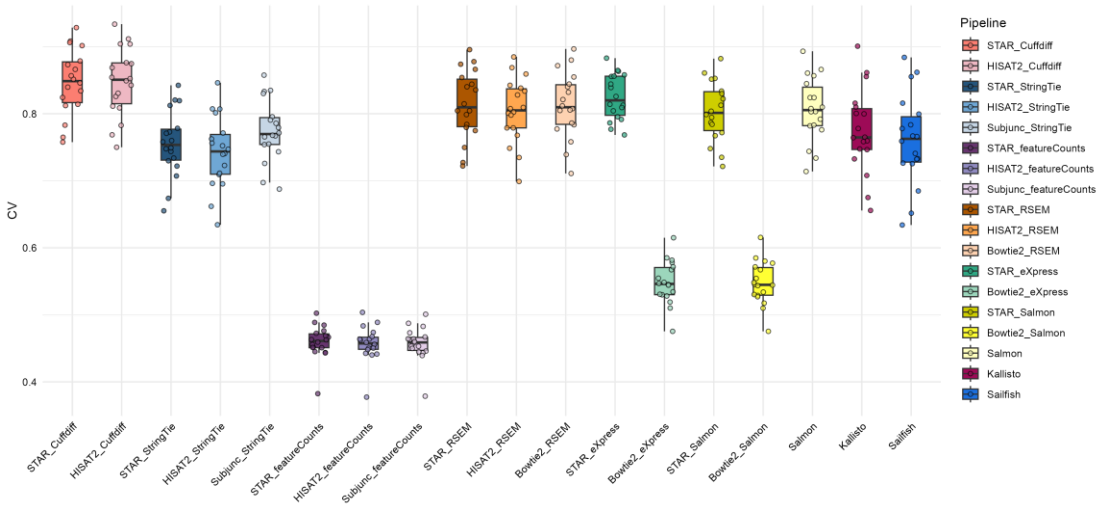
a, Total number of DEIs identified by each laboratory, including up-regulated and down-regulated isoforms. **b–f,** A positive correlation was observed between sequencing depth in exonic regions and the number of detected DEIs for the (b) M8 vs. D6, (c) F7 vs. D6, (d) D5 vs. D6, (e) T1 vs. D6, (f) T2 vs. D6 sample pairs. *R* represents the Pearson correlation coefficient.

Figure S18. PCCs of 18 isoform quantification pipelines.



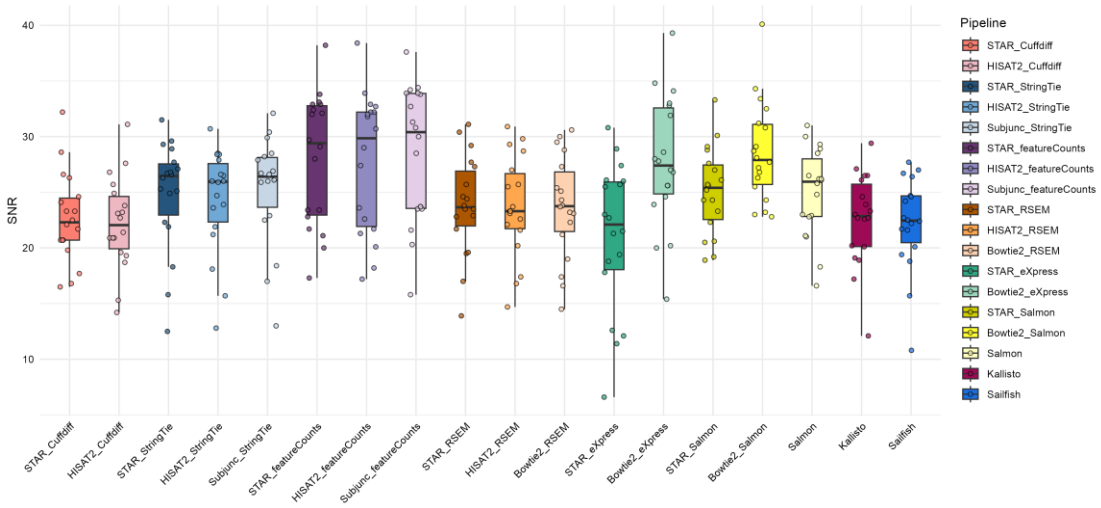
Boxplots show the distribution of Pearson correlation coefficients (PCCs) across 19 high-quality datasets for each isoform quantification pipelines based on the (a) Quartet and (b) RT-qPCR reference datasets. Data are presented as median values (center lines) and the upper and lower quartiles (box limits).

Figure S19. Comparison of CV across 19 quantification pipelines.



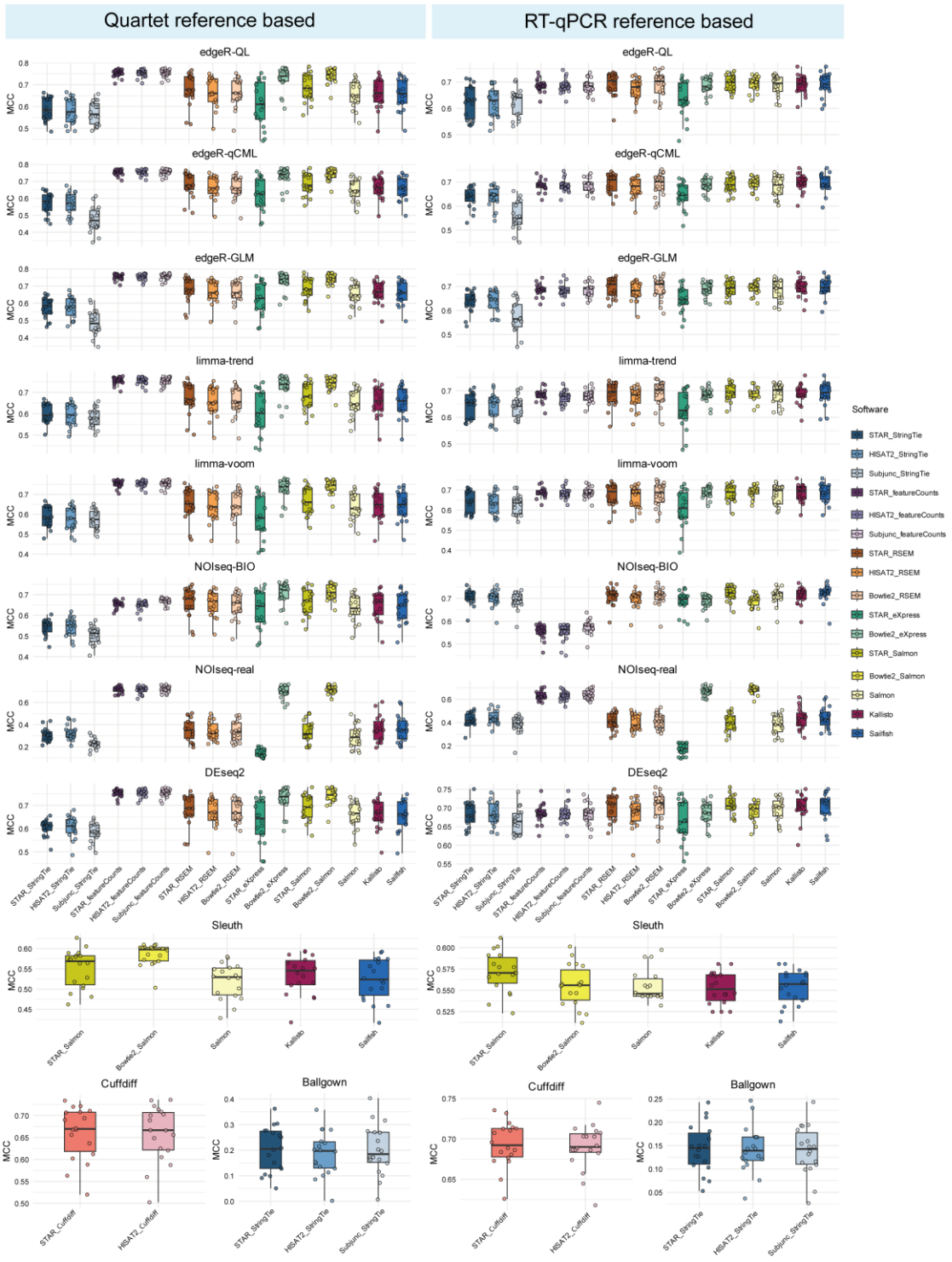
Boxplots show the distribution of average coefficient of variation (CV) values of isoform expression levels among replicates in the Quartet samples across 19 quantification pipelines. Individual points represent CV values from high-quality laboratories. Data are presented as median values (center lines) and the upper and lower quartiles (box limits).

Figure S20. Comparison of SNR across 18 quantification pipelines.



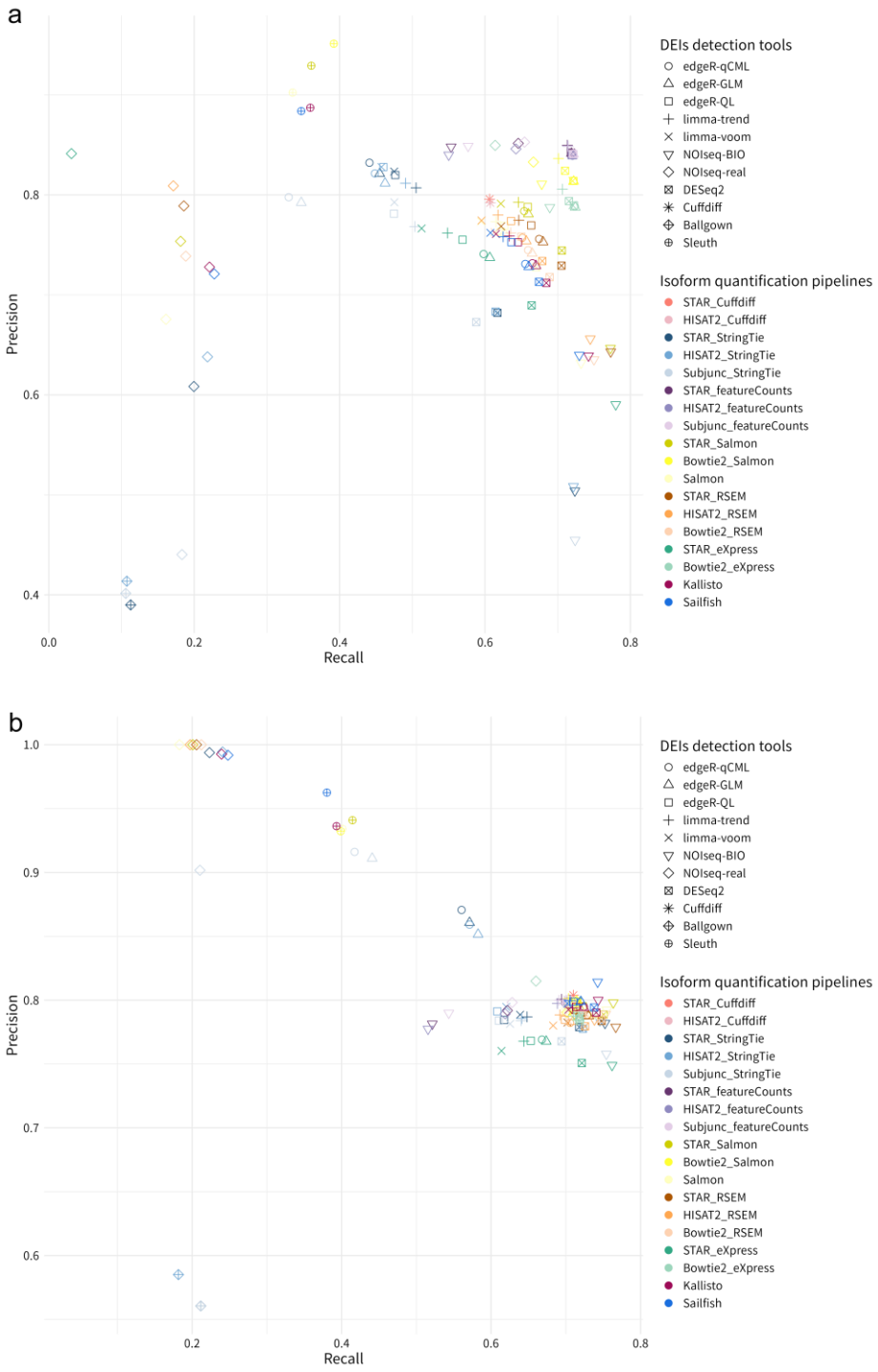
Boxplots display the distribution of SNR values for isoform expression data from the Quartet and mixed samples across 18 quantification pipelines. Individual points represent SNR values for high-quality laboratories. Data are presented as median values (center lines) and the upper and lower quartiles (box limits).

Figure S21. Impact of isoform quantification tools on DEI detection.



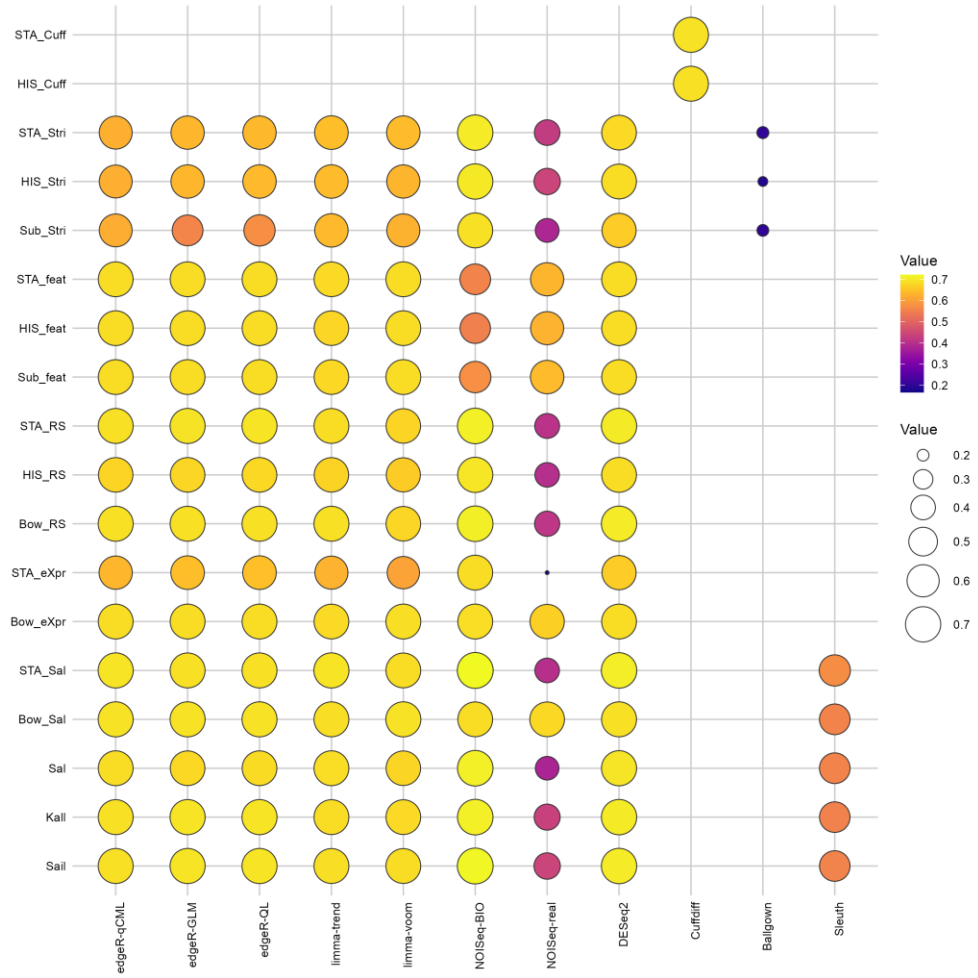
Boxplots show the distribution of Matthews correlation coefficients (MCC) across 19 high-quality datasets for each differential expression analysis software, stratified by different isoform quantification tools. Data are presented as median values (center lines) and the upper and lower quartiles (box limits).

Figure S22. Recall and precision of differential expression analysis pipelines.



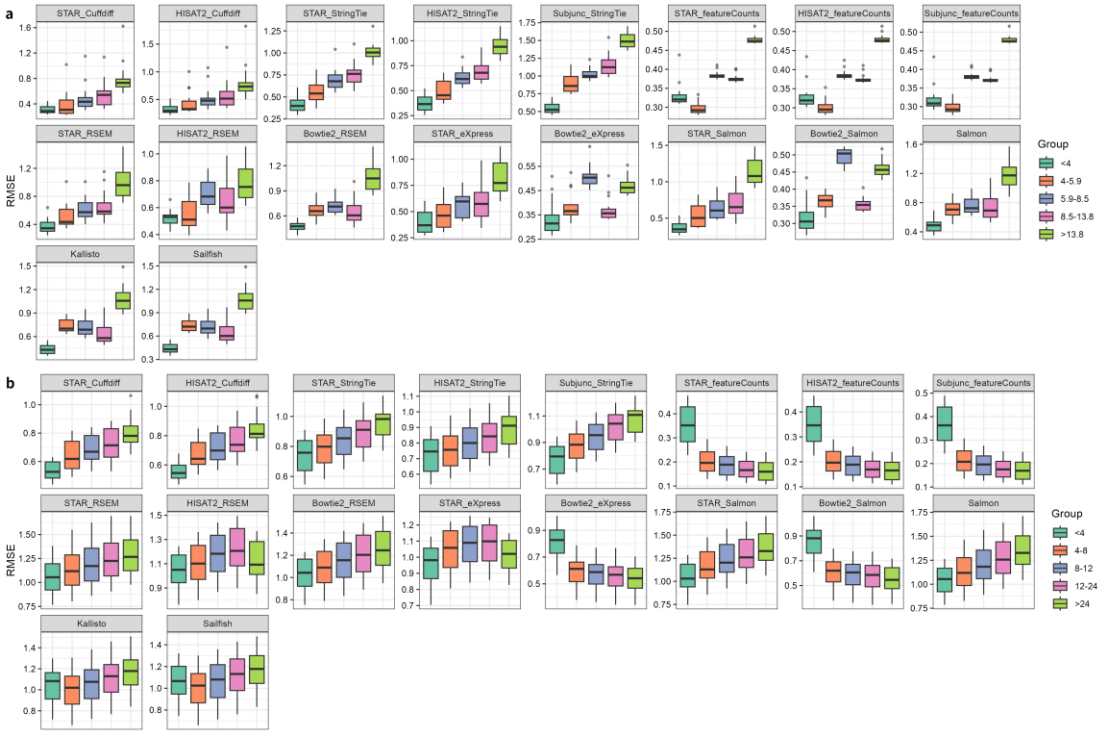
The scatter plot illustrates the recall and precision of 138 differential expression analysis pipelines based on the (a) Quartet and (b) RT-qPCR reference datasets. Each point represents the mean values calculated from the analysis results of 19 high-quality RNA-seq data. Different colors indicate distinct quantification pipelines, whereas different shapes denote different differential expression analysis tools.

Figure S23. Performance of DEI detection tools combined with isoform quantification pipelines.



Bubble plots show the mean Matthews correlation coefficient (MCC) across 19 high-quality datasets for each combination. Both the size and the color intensity of the bubbles indicate the magnitude of MCC.

Figure S24. Impact of isoform complexity (K-value) on isoform quantification pipelines.



Boxplots display the performance of 18 isoform quantification pipelines across 19 high-quality datasets in relation to K-value of isoforms based on the (a) Quartet reference datasets and (b) the recovery of mixing ratios against the built-in truth. Data are presented as median values (center lines) and the upper and lower quartiles (box limits).

Figure S25. Impact of isoform length on isoform quantification pipelines.

Boxplots display the performance of 18 isoform quantification pipelines across 19 high-quality datasets in relation to length of isoforms based on the (a) Quartet reference datasets and (b) the recovery of mixing ratios against the built-in truth. Data are presented as median values (center lines) and the upper and lower quartiles (box limits).

Figure S26. Impact of isoform number per gene on isoform quantification pipelines.

Boxplots display the performance of 18 isoform quantification pipelines across 19 high-quality datasets in relation to isoform number per gene based on the (a) Quartet reference datasets and (b) the recovery of mixing ratios against the built-in truth. Data are presented as median values (center lines) and the upper and lower quartiles (box limits).

Figure S27. Impact of exon number per isoform on isoform quantification pipelines.

Boxplots display the performance of 18 isoform quantification pipelines across 19 high-quality datasets in relation to exon number per isoform based on the (a) Quartet reference datasets and (b) the recovery of mixing ratios against the built-in truth. Data are presented as median values (center lines) and the upper and lower quartiles (box limits).

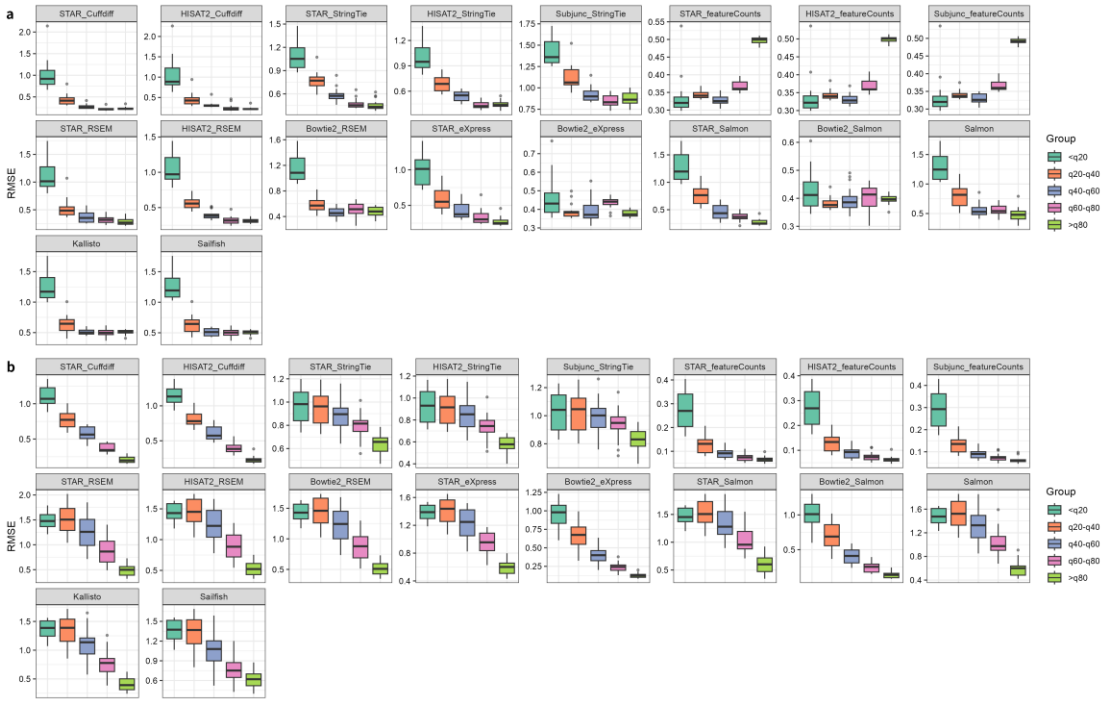
Figure S28. Impact of exon length on isoform quantification pipelines.

Boxplots display the performance of 18 isoform quantification pipelines across 19 high-quality datasets in relation to mean exon length of each isoform based on the (a) Quartet reference datasets and (b) the recovery of mixing ratios against the built-in truth. Data are presented as median values (center lines) and the upper and lower quartiles (box limits).

Figure S29. Impact of GC content on isoform quantification pipelines.

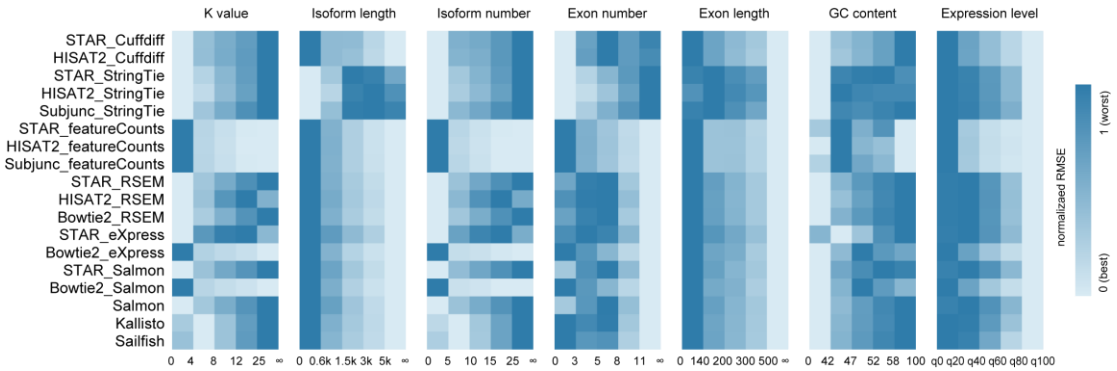
Boxplots display the performance of 18 isoform quantification pipelines across 19 high-quality datasets in relation to GC content based on the (a) Quartet reference datasets and (b) the recovery of mixing ratios against the built-in truth. Data are presented as median values (center lines) and the upper and lower quartiles (box limits).

Figure S30. Impact of isoform expression on isoform quantification pipelines.



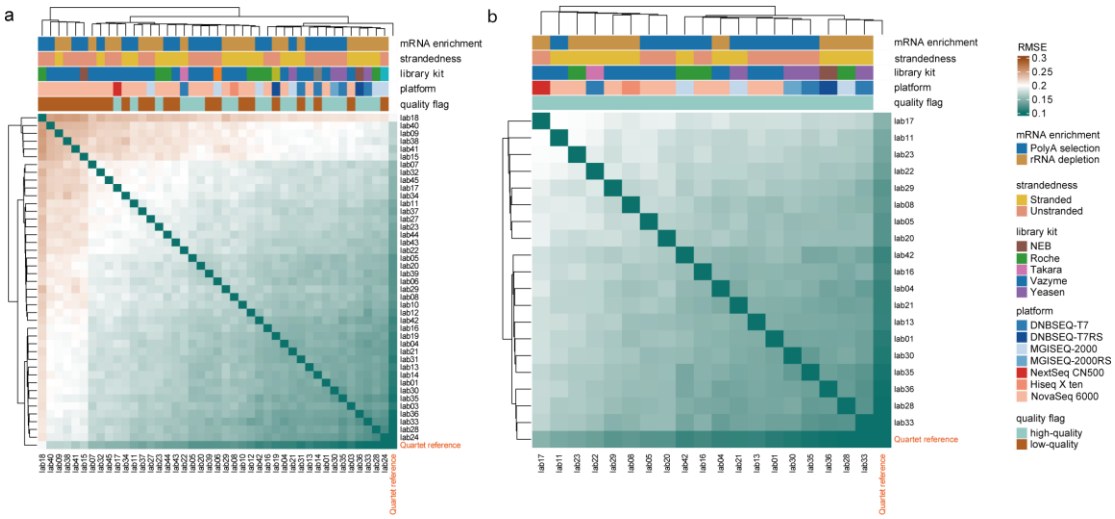
Boxplots display the performance of 18 isoform quantification pipelines across 19 high-quality datasets in relation to isoform expression level based on the (a) Quartet reference datasets and (b) the recovery of mixing ratios against the built-in truth. Data are presented as median values (center lines) and the upper and lower quartiles (box limits).

Figure S31. Impact of isoform features based on the built-in truth.



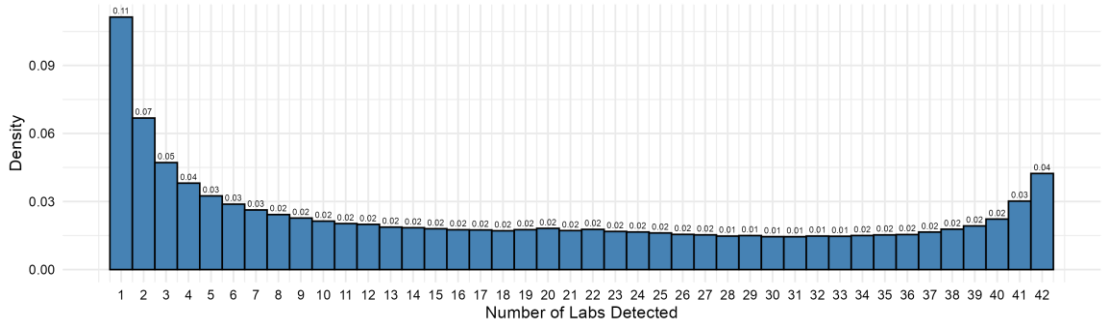
Evaluation of quantification pipelines with respect to multiple isoform features based on the built-in truth, including the isoform length, the number of isoforms per gene, the average length of exons per isoform, the number of exons per isoform, GC content, K-value (exon-isoform structural complexity), and expression level (FPKM).

Figure S32. Pairwise consistency in AS event quantification across laboratories.



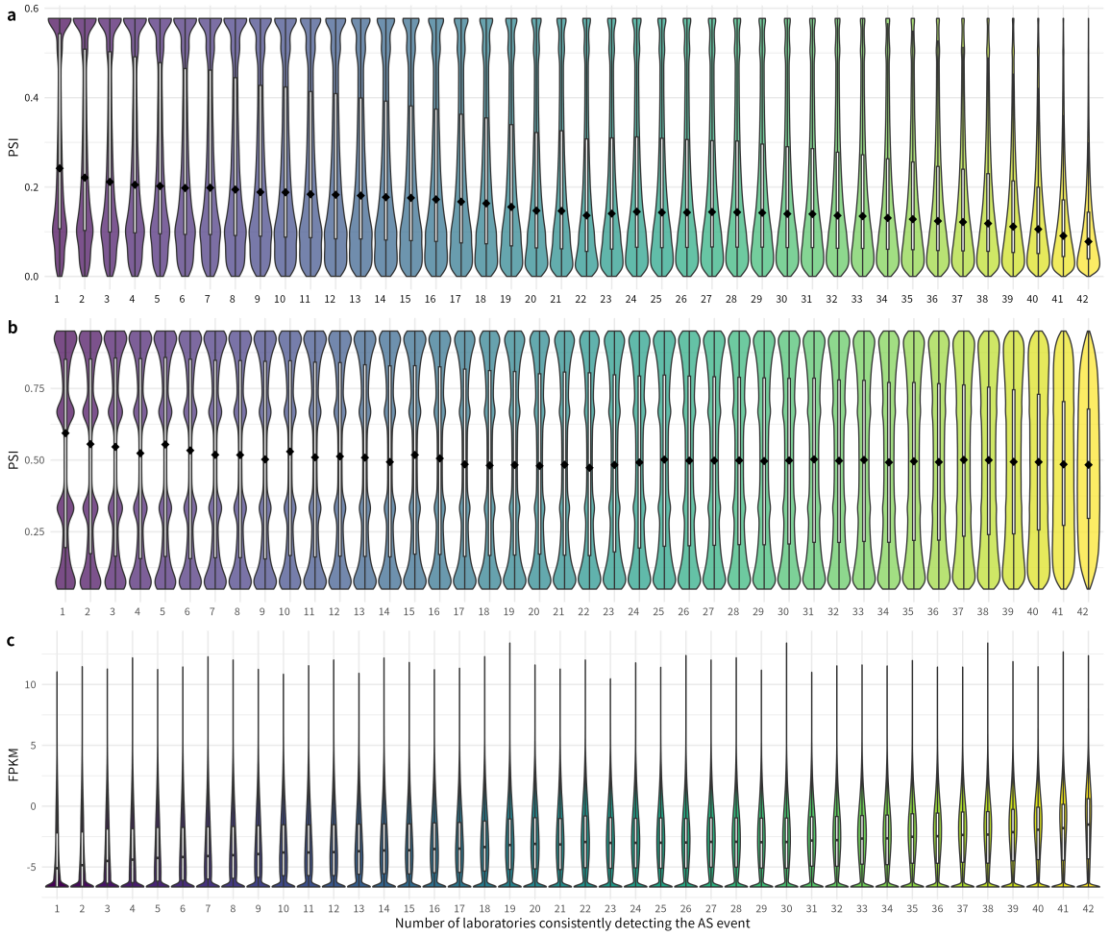
a, Heatmap showing the root mean square error (RMSE) of alternative splicing (AS) event-level delta percent spliced-in (dPSI) values between each pair of the 42 participating laboratories. **b**, Heatmap showing RMSE values for dPSI comparisons among the 19 high-quality laboratories. Rows and columns are clustered based on hierarchical clustering of RMSE values.

Figure S33. Consistency of ASE detection across 42 laboratories.



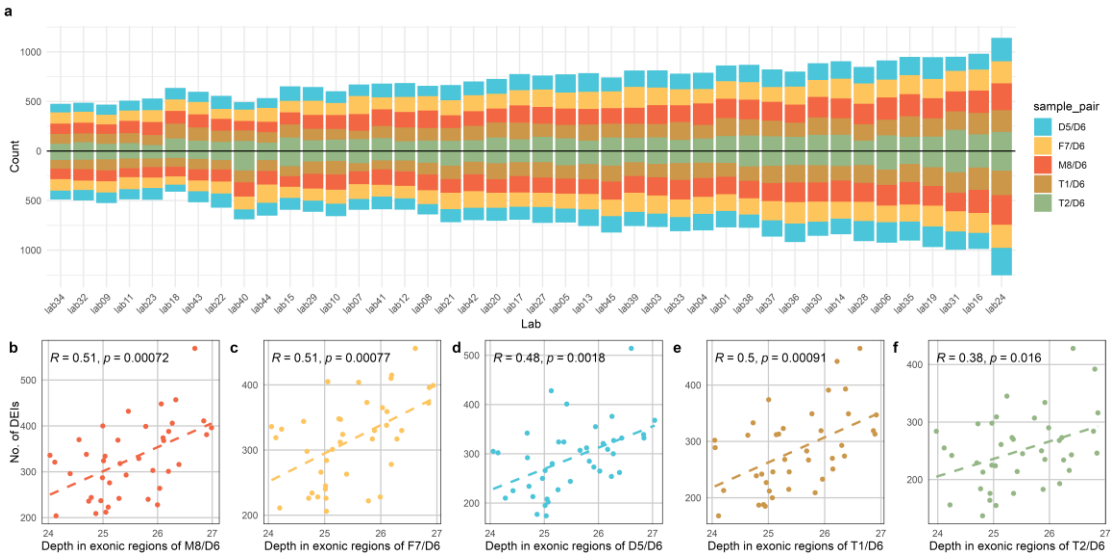
The density plot illustrates the distribution of alternative splicing events (ASEs) with respect to the number of laboratories that commonly detected them.

Figure S34. Characteristics of ASEs with different detection consistency.



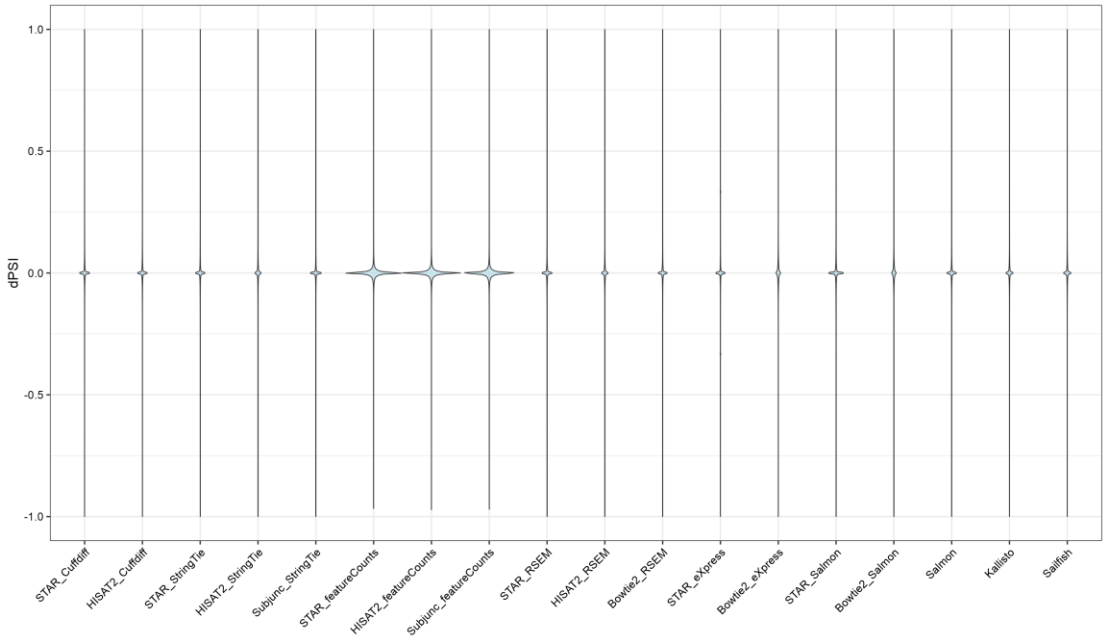
a, Percent spliced-in (PSI) levels, **b**, PSI reproducibility (standard deviation), **c,** and expression levels of isoforms associated with ASEs, respectively, stratified by their detection consistency across 42 laboratories from low to high.

Figure S35. Number of DSEs detected by laboratories.



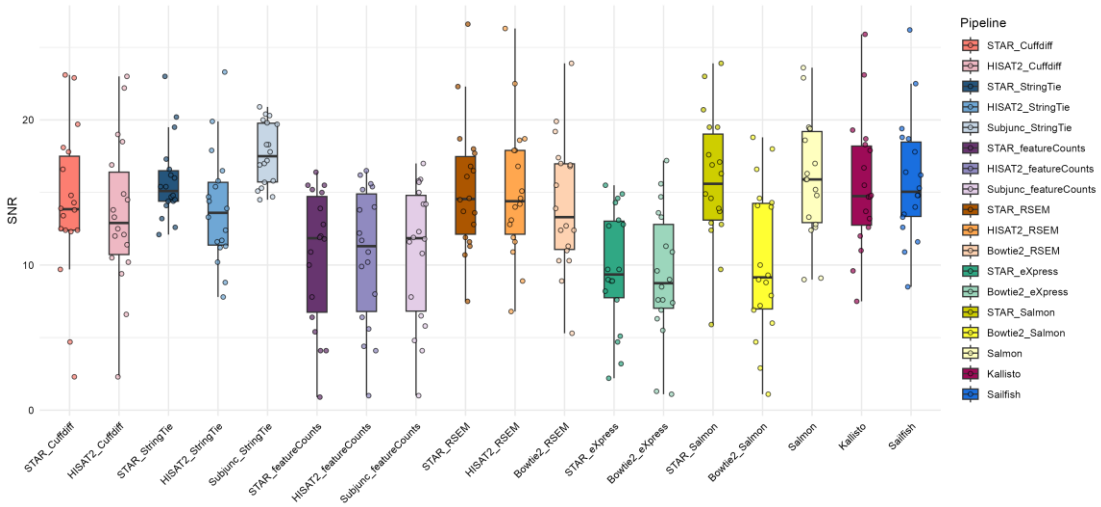
a, Total number of differential splicing events (DSEs) identified by each laboratory, including up-regulated and down-regulated events. **b–f,** A positive correlation was observed between sequencing depth in exonic regions and the number of detected DSEs for the (b) M8 vs. D6, (c) F7 vs. D6, (d) D5 vs. D6, (e) T1 vs. D6, and (f) T2 vs. D6 sample pairs. *R* represents the Pearson correlation coefficient.

Figure S36. Distribution of dPSI values detected by SUPPA2 combined with 18 isoform quantification pipelines.



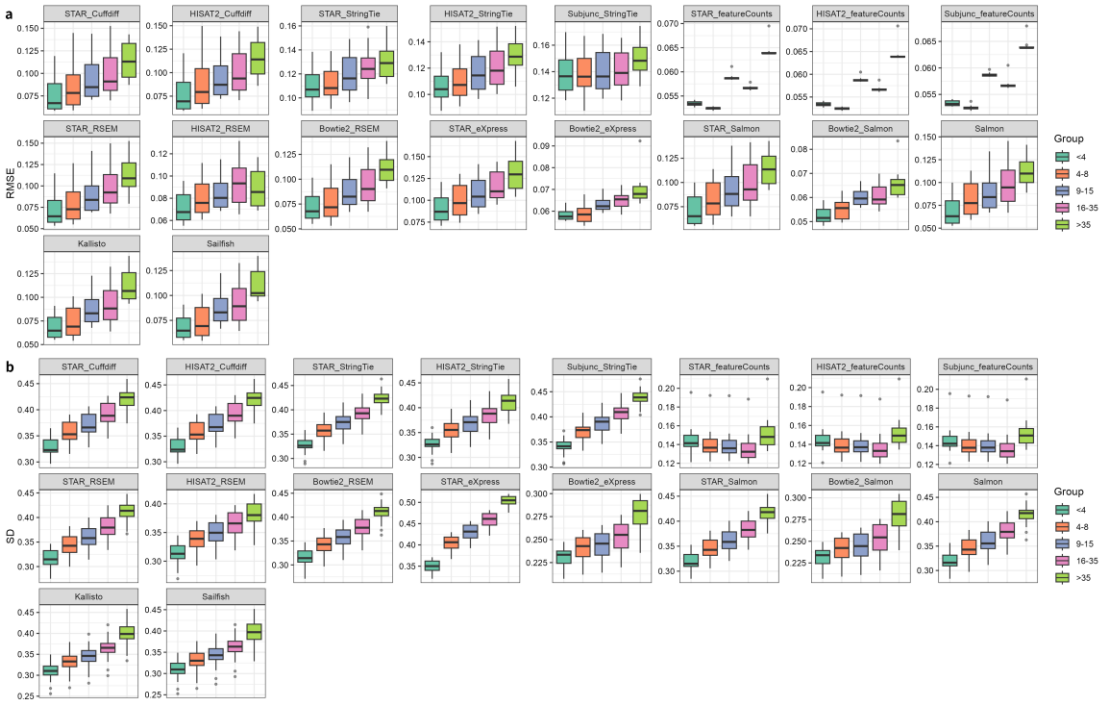
Ridgeline Plot show the delta percent spliced-in (dPSI) distributions analyzed by SUPPA2 using TPM data from 18 different isoform quantification pipelines for all detected alternative splicing (AS) events. The dPSI distribution of AS events included in the Quartet reference datasets are showed in Fig.5d.

Figure S37. SNR for SUPPA2 combined with 18 isoform quantification pipelines.



Violin plots show the SNR distributions of quantification data (PSI) analyzed by SUPPA2 using TPM data from 18 different isoform quantification pipelines.

Figure S38. Impact of the number of adjacent AS events on AS detection pipelines.



Boxplots showing the (a) accuracy and (b) reproducibility of SUPPA2 combined with 18 isoform quantification pipelines as a function of the number of AS events in adjacent regions. Reproducibility was assessed using the standard deviation (SD). Data are shown as median values (center lines) with upper and lower quartiles (box boundaries).

Figure S39. Impact of AS event-associated isoform number on AS detection pipelines.

Boxplots showing the (a) accuracy and (b) reproducibility of SUPPA2 combined with 18 isoform quantification pipelines as a function of the isoform number associated with AS events. Reproducibility was assessed using the standard deviation (SD). Data are shown as median values (center lines) with upper and lower quartiles (box boundaries).

Figure S40. Impact of the GC content on AS detection pipelines.

Boxplots showing the (a) accuracy and (b) reproducibility of SUPPA2 combined with 18 isoform quantification pipelines as a function of the GC content. The GC content of each AS event was calculated from two 75 bp regions flanking its splice junctions. Reproducibility was assessed using the standard deviation (SD). Data are shown as median values (center lines) with upper and lower quartiles (box boundaries).

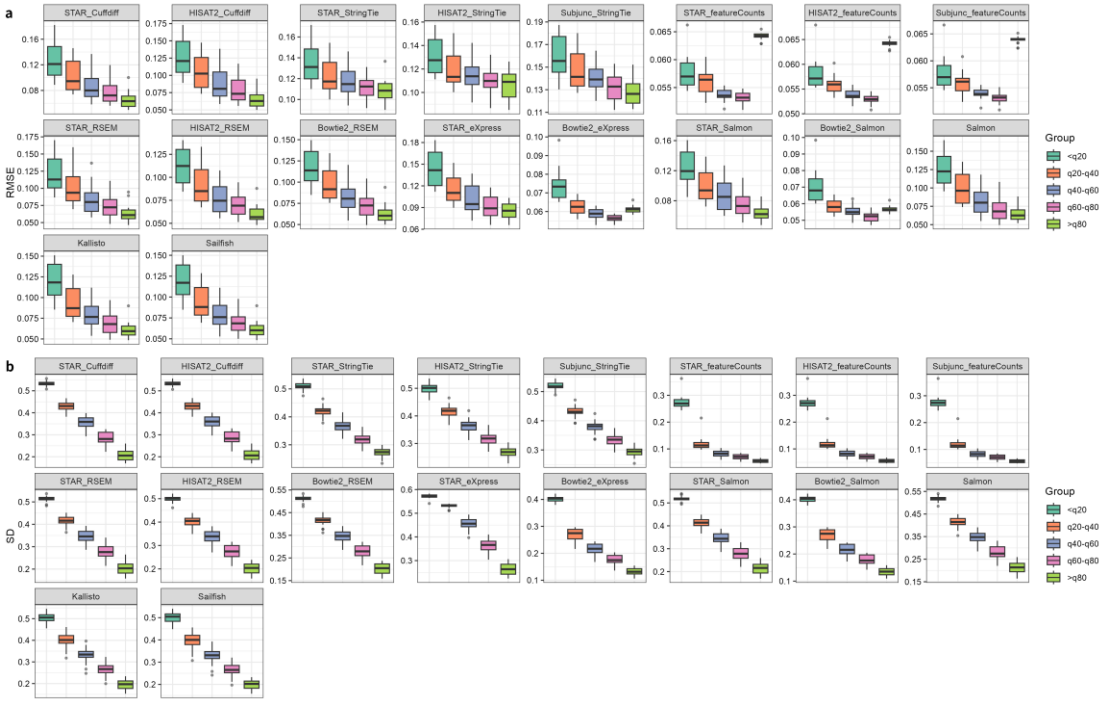
Figure S41. Impact of the mappability on AS detection pipelines.

Boxplots showing the (a) accuracy and (b) reproducibility of SUPPA2 combined with 18 isoform quantification pipelines as a function of the mappability. The mappability of each AS event was calculated from two 75 bp regions flanking its splice junctions. Reproducibility was assessed using the standard deviation (SD). Data are shown as median values (center lines) with upper and lower quartiles (box boundaries).

Figure S42. Impact of PSI levels on AS detection pipelines.

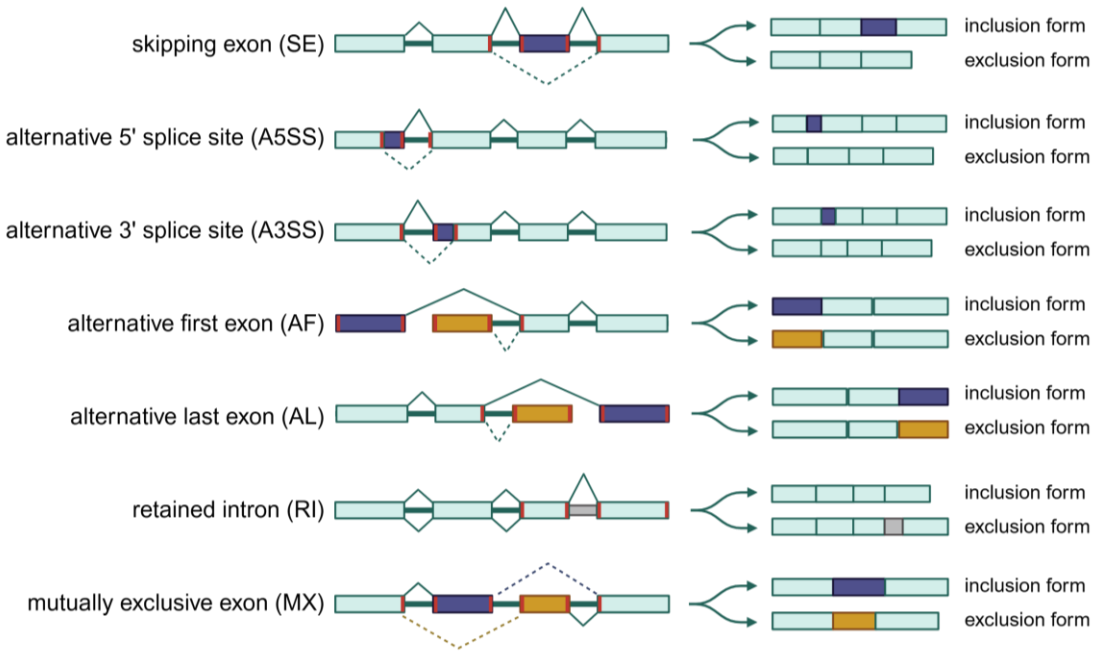
Boxplots showing the (a) accuracy and (b) reproducibility of SUPPA2 combined with 18 isoform quantification pipelines as a function of AS event PSI levels in 19 high-quality RNA-seq datasets. Reproducibility was assessed using the standard deviation (SD). Data are shown as median values (center lines) with upper and lower quartiles (box boundaries).

Figure S43. Impact of AS event-associated isoform expression levels on AS detection pipelines.



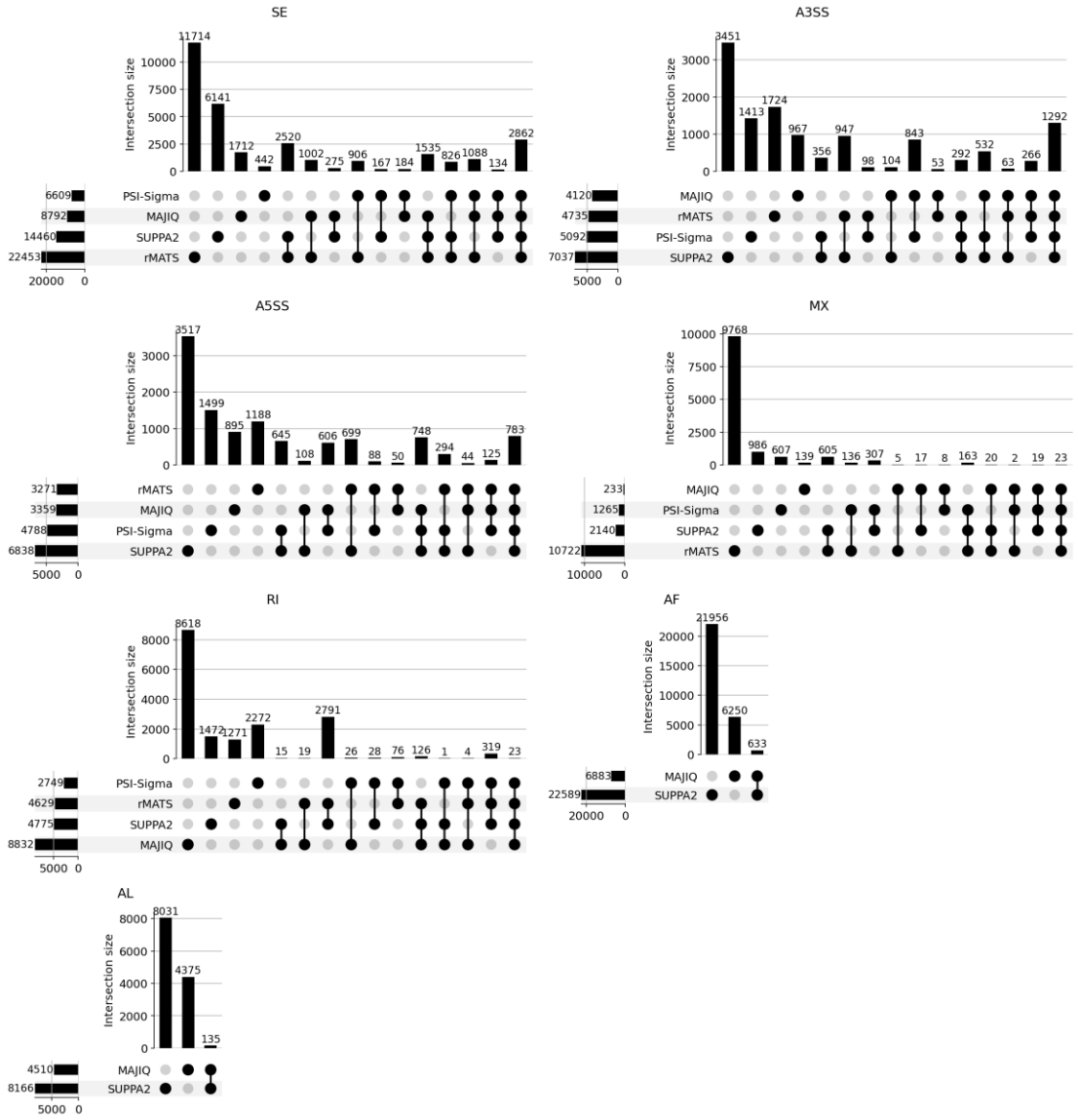
Boxplots showing the (a) accuracy and (b) reproducibility of SUPPA2 combined with 18 isoform quantification pipelines as a function of AS-related isoform expression levels in 19 high-quality RNA-seq datasets. Isoform expression levels were defined as the average FPKM of all isoforms related to each AS event. Reproducibility was assessed using the standard deviation (SD). Data are presented as median values (center lines) with upper and lower quartiles (box boundaries).

Figure S44. A unified format for different AS event types.



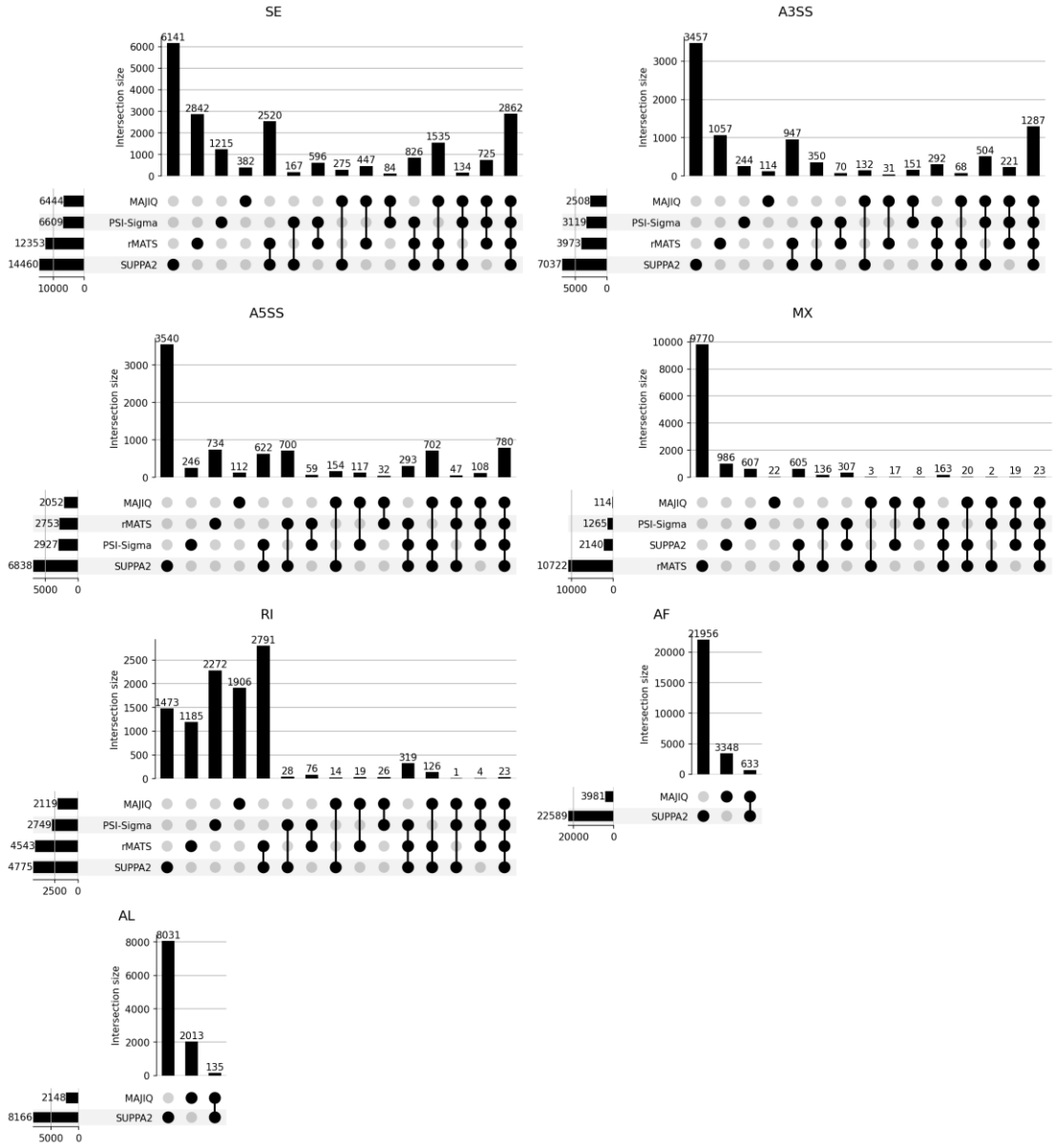
The coordinates of seven AS events defined in the unified format output are shown as red lines in the corresponding splice graph. Created in BioRender. Zhang, R. (2025) https://BioRender.com/uvu00yo.

Figure S45. The intersections of annotated and novel AS events by four AS event detection tools.



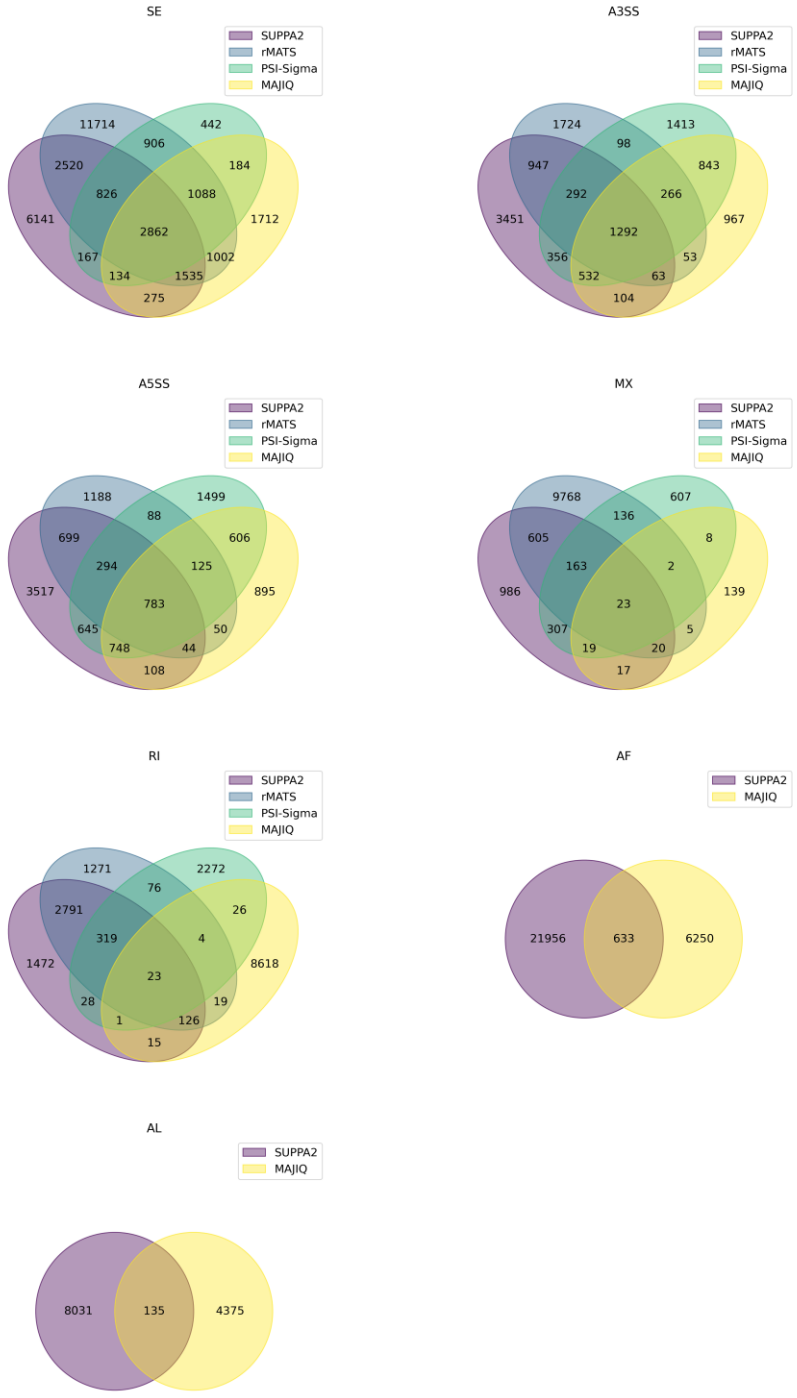
The standardized benchmarking tool generates the Upset plots indicating intersections of alternative splicing (AS) event across MAJIQ, PSI-Sigma, rMATS, and SUPPA2, including both annotated and novel events. AS events with percent spliced-in (PSI) values between 0.05 and 0.95 were included. Data from the M8 sample of lab1 are shown as an example.

Figure S46. The intersections of annotated AS events by four AS event detection tools.



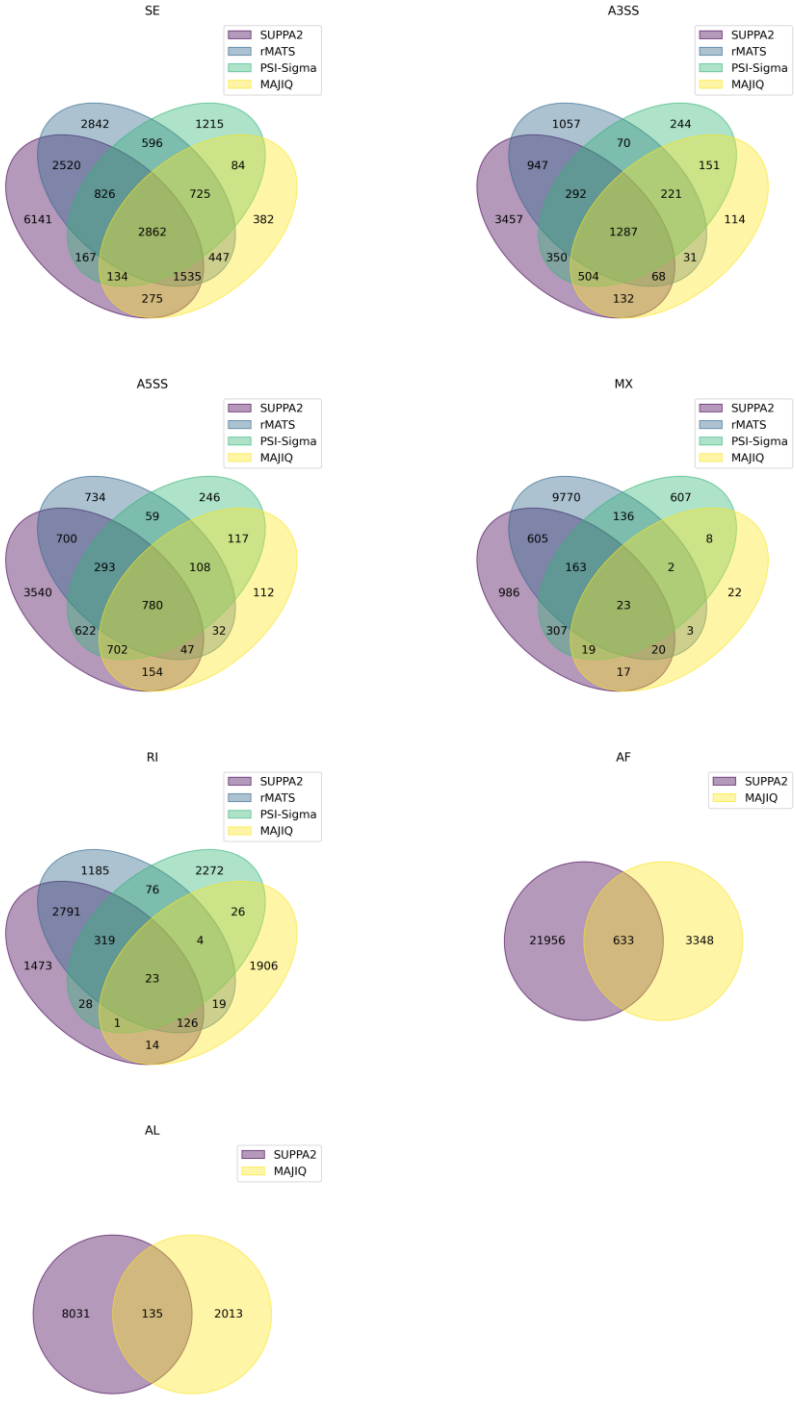
The standardized benchmarking tool generates the Upset plots indicating intersections of annotated alternative splicing (AS) event across MAJIQ, PSI-Sigma, rMATS, and SUPPA2. AS events with percent spliced-in (PSI) values between 0.05 and 0.95 were included. Data from the M8 sample of lab1 are shown as an example.

Figure S47. The intersections of annotated and novel AS events by four AS event detection tools.



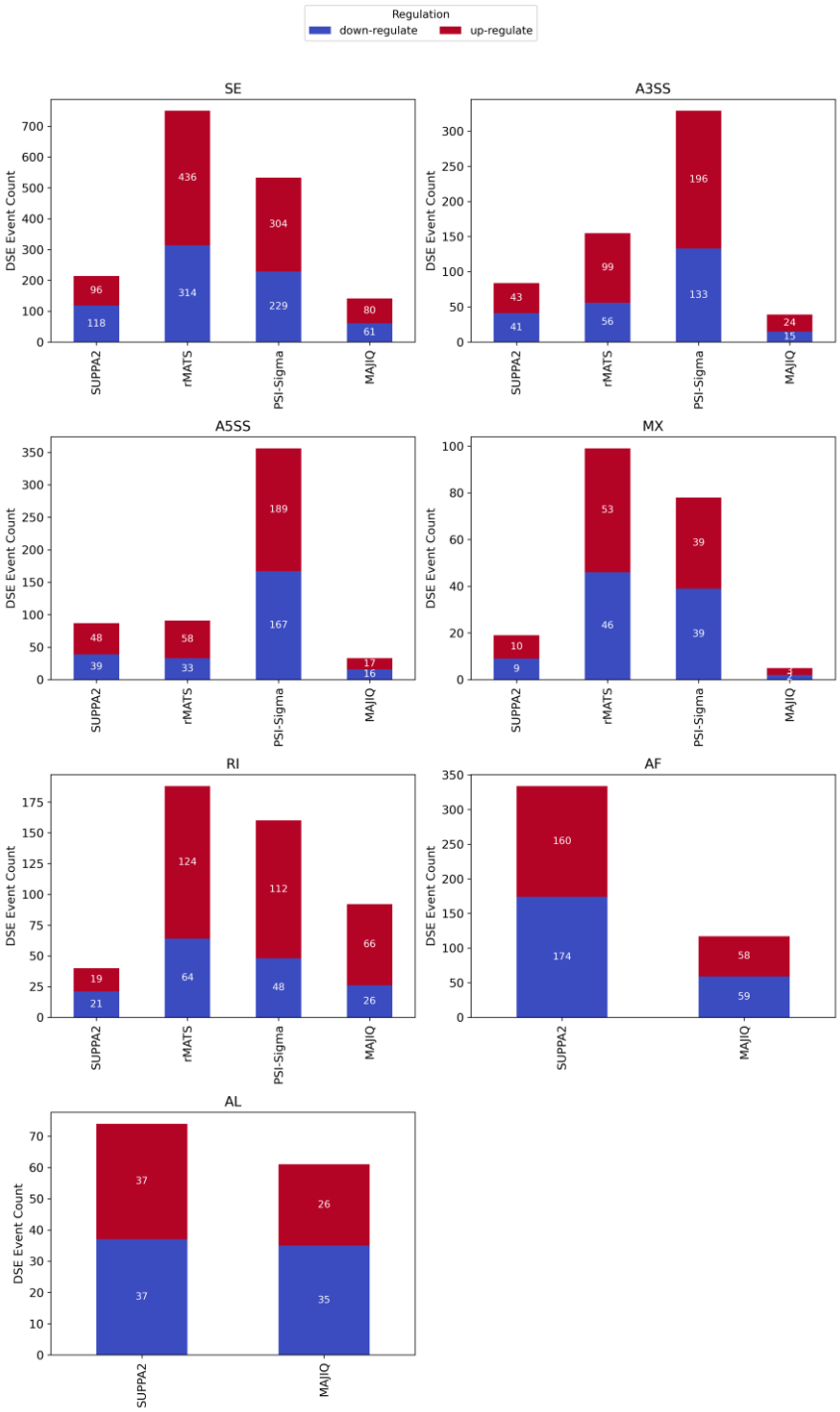
The standardized benchmarking tool generates the Venn plots indicating intersections of alternative splicing (AS) event across MAJIQ, PSI-Sigma, rMATS, and SUPPA2, including both annotated and novel events. AS events with percent spliced-in (PSI) values between 0.05 and 0.95 were included. Data from the M8 sample of lab1 are shown as an example.

Figure S48. The intersections of annotated AS events by four AS event detection tools.



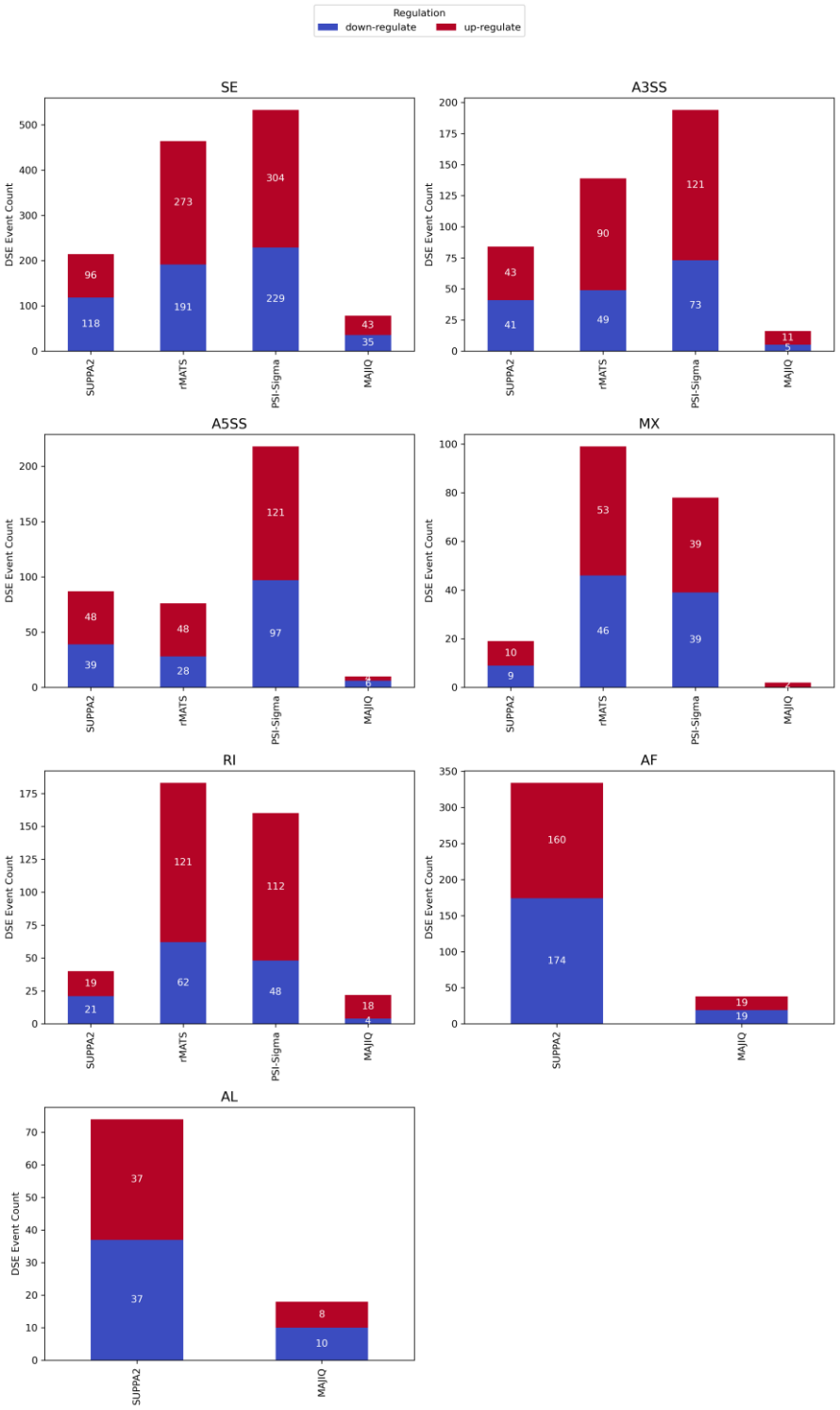
The standardized benchmarking tool generates the Venn plots indicating intersections of annotated alternative splicing (AS) event across MAJIQ, PSI-Sigma, rMATS, and SUPPA2. AS events with percent spliced-in (PSI) values between 0.05 and 0.95 were included. Data from the M8 sample of lab1 are shown as an example.

Figure S49. The number of annotated and novel DSEs by four AS event detection tools.



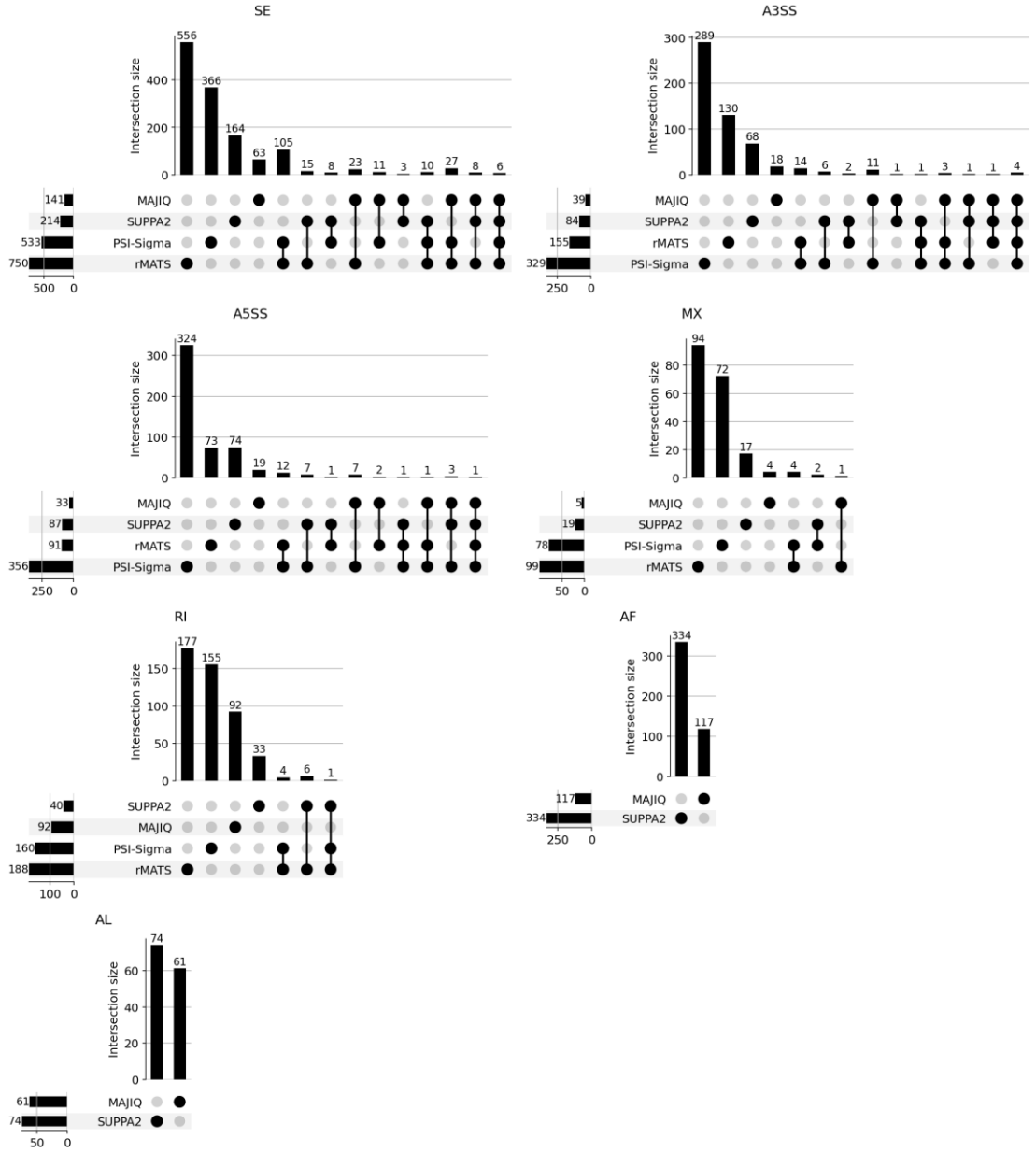
The standardized benchmarking tool generates the histogram indicating number of differential splicing events (DSEs) detected by MAJIQ, PSI-Sigma, rMATS, and SUPPA2, including both annotated and novel alternative splicing (AS) events. Data from the M8/D6 sample pair of lab1 are shown as an example.

Figure S50. The number of annotated DSEs by four AS event detection tools.



The standardized benchmarking tool generates the histogram indicating number of annotated differential splicing events (DSEs) detected by MAJIQ, PSI-Sigma, rMATS, and SUPPA2. Data from the M8/D6 sample pair of lab1 are shown as an example.

Figure S51. The intersections of annotated and novel DSEs by four AS event detection tools.

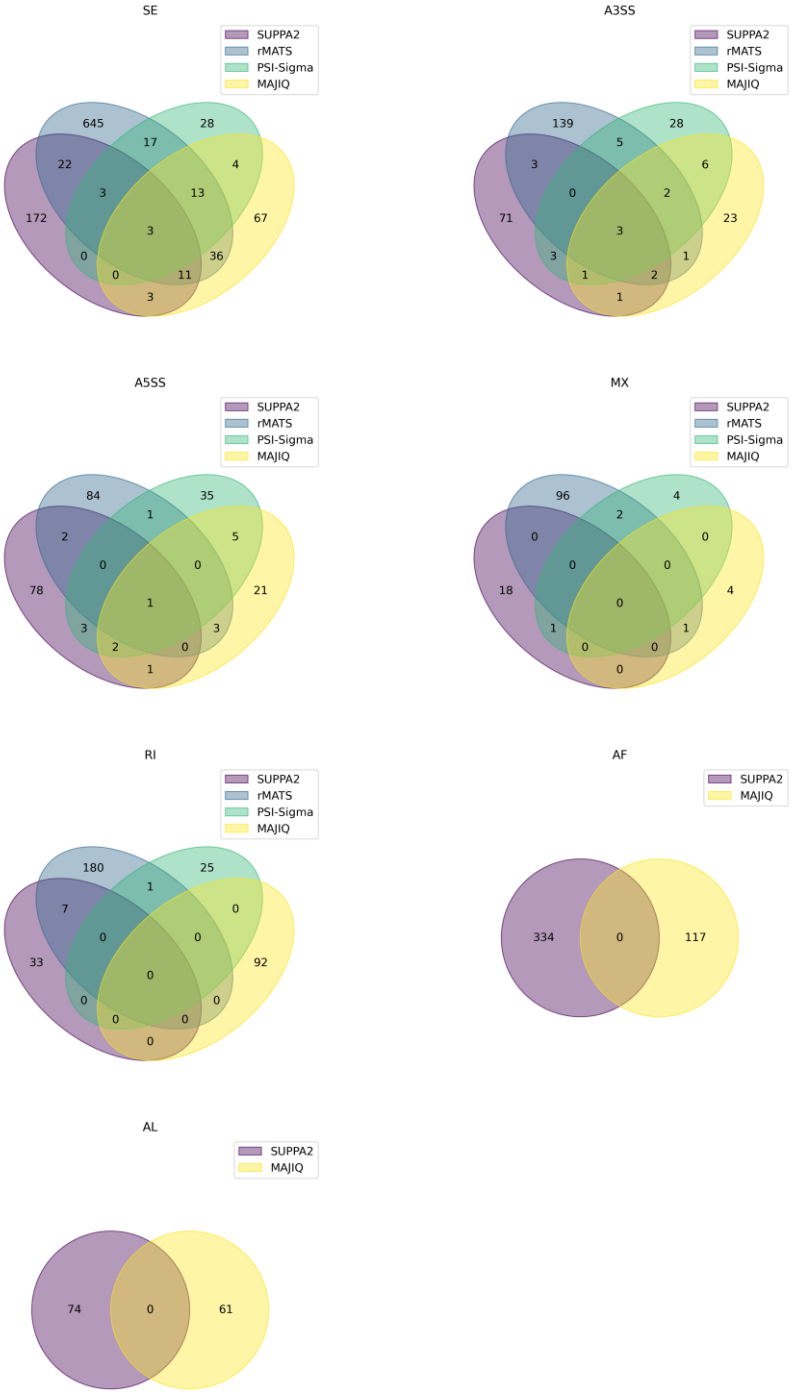


The standardized benchmarking tool generates the Upset plots indicating intersections of differential splicing events (DSEs) across MAJIQ, PSI-Sigma, rMATS, and SUPPA2, including both annotated and novel alternative splicing (AS) events. Data from the M8/D6 sample pair of lab1 are shown as an example.

Figure S52. The intersections of annotated DSEs by four AS event detection tools.

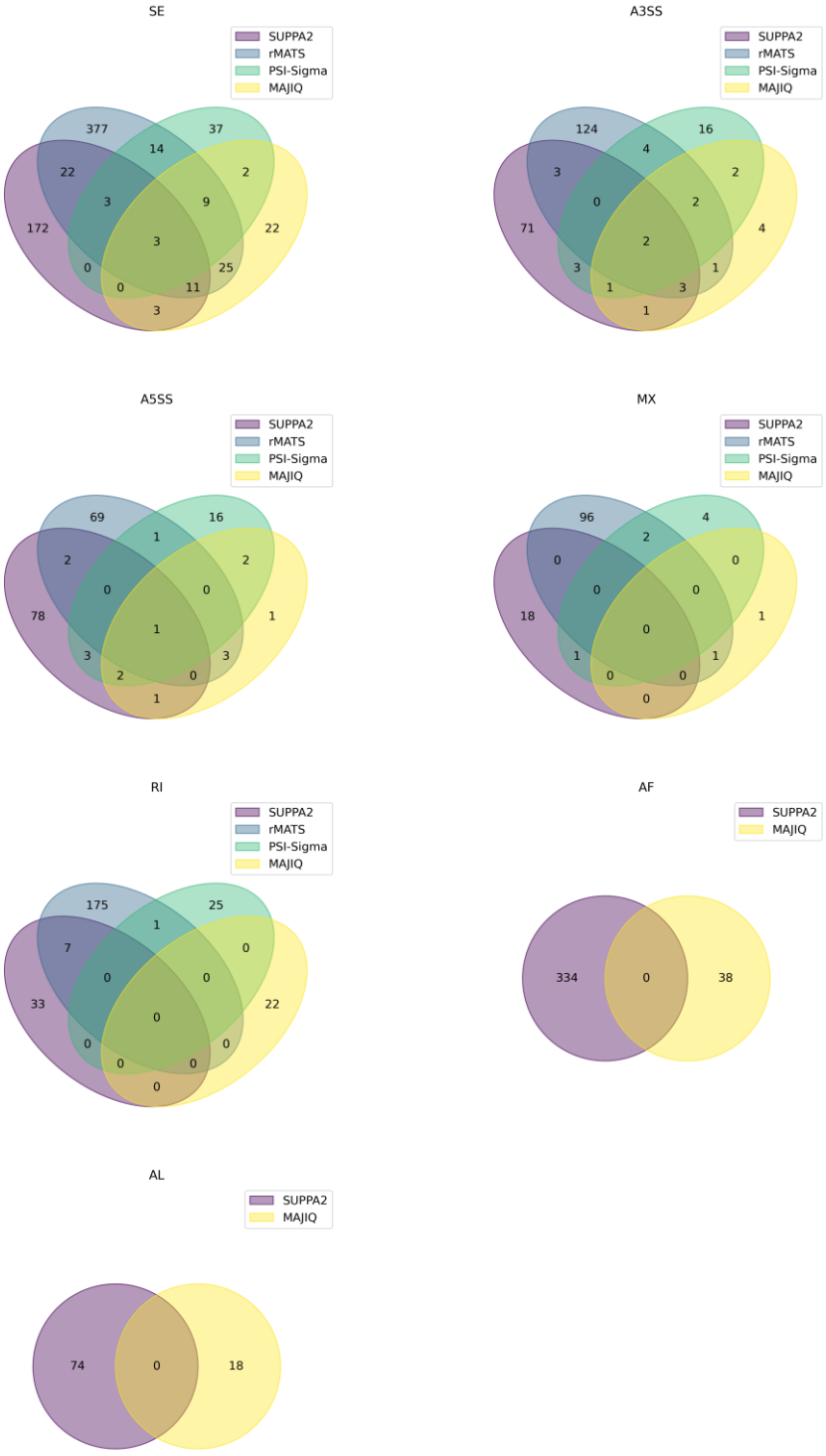
The standardized benchmarking tool generates the Upset plots indicating intersections of annotated differential splicing events (DSEs) across MAJIQ, PSI-Sigma, rMATS, and SUPPA2. Data from the M8/D6 sample pair of lab1 are shown as an example.

Figure S53. The intersections of annotated and novel DSEs by four AS event detection tools.



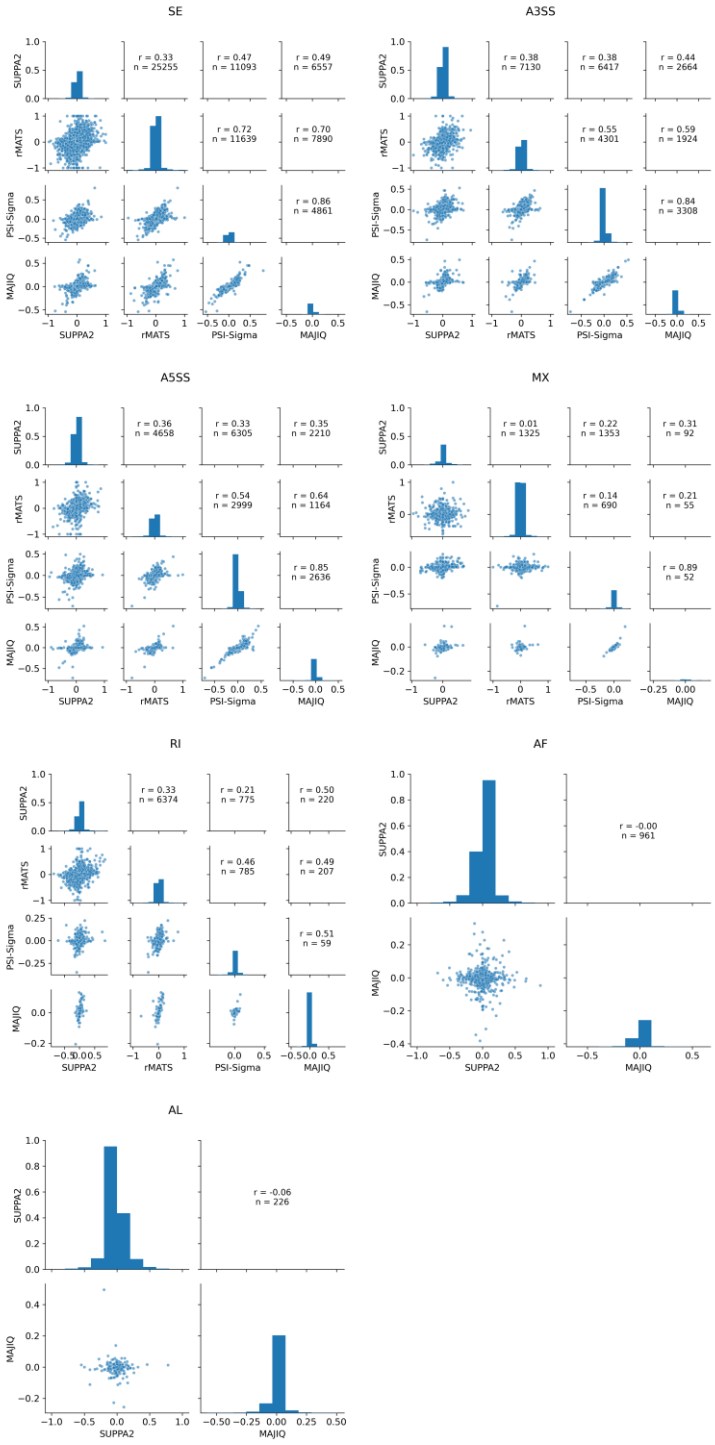
The standardized benchmarking tool generates the Venn plots indicating intersections of differential splicing events (DSEs) across MAJIQ, PSI-Sigma, rMATS, and SUPPA2, including both annotated and novel alternative splicing (AS) events. Data from the M8/D6 sample pair of lab1 are shown as an example.

Figure S54. The intersections of annotated DSEs by four AS event detection tools.



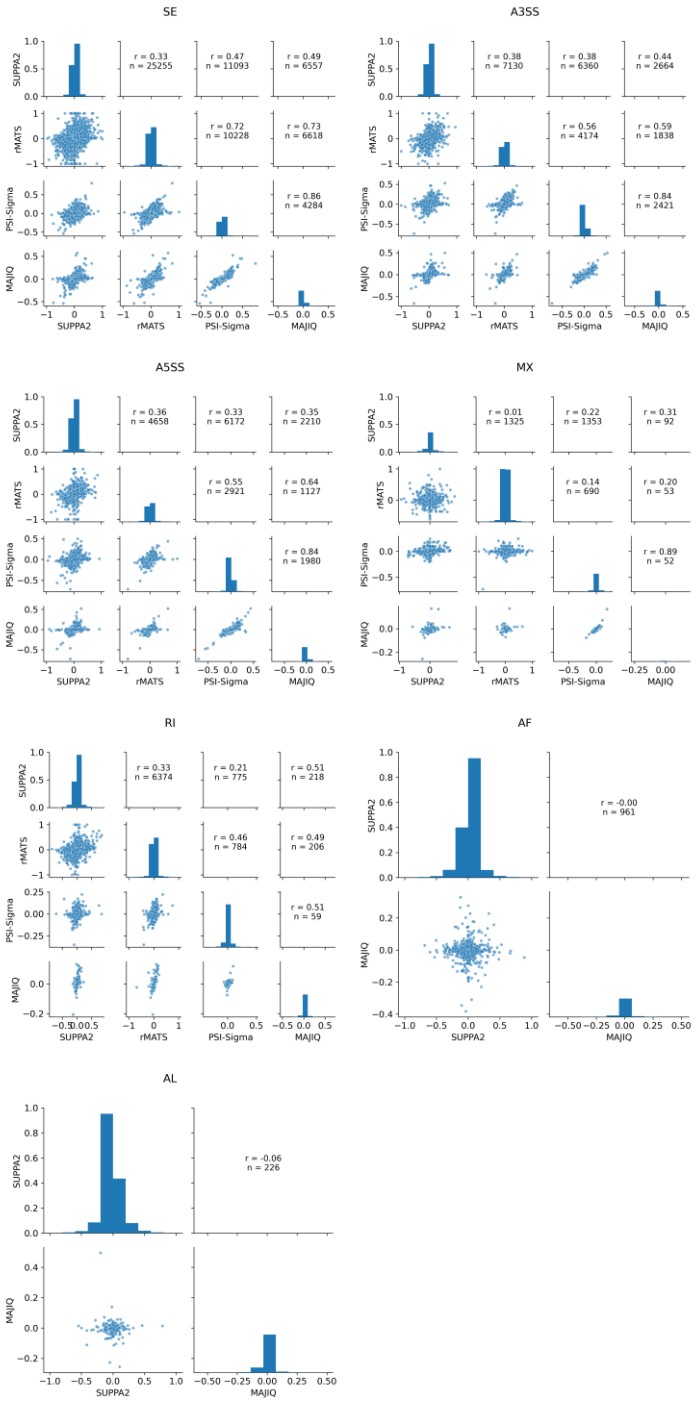
The standardized benchmarking tool generates the Venn plots indicating intersections of annotated differential splicing events (DSEs) across MAJIQ, PSI-Sigma, rMATS, and SUPPA2. Data from the M8/D6 sample pair of lab1 are shown as an example.

Figure S55. The dPSI consistency of annotated and novel AS events by four AS event detection tools.



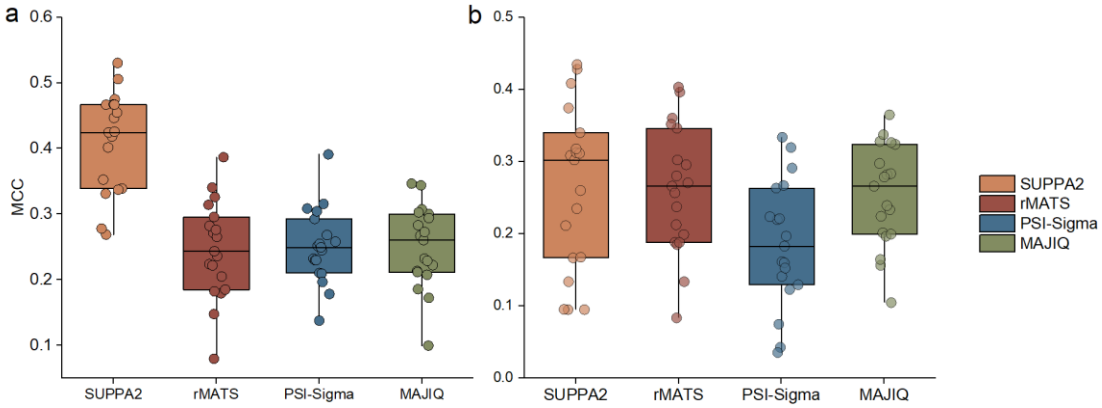
The standardized benchmarking tool generates the Scatter plots indicating the consistency of delta percent spliced-in (dPSI) results across MAJIQ, PSI-Sigma, rMATS, and SUPPA2, including both annotated and novel alternative splicing (AS) events. Data from the M8/D6 sample pair of lab1 are shown as an example.

Figure S56. The dPSI consistency of annotated AS events by four AS event detection tools.



The standardized benchmarking tool generates the Scatter plots indicating the consistency of delta percent spliced-in (dPSI) results for annotated alternative splicing (AS) events across MAJIQ, PSI-Sigma, rMATS, and SUPPA2. Data from the M8/D6 sample pair of lab1 are shown as an example.

Figure S57. The MCC of four AS event detection tools for DSE detection.



The Matthews correlation coefficient (MCC) of 19 high-quality laboratories in differential splicing event (DSE) detection against the (a) Quartet and (b) RT-qPCR reference datasets by four AS event detection tools. Data are presented as median values (center lines) with upper and lower quartiles (box boundaries).

Reference

1. Chen, Q. et al. A ratio-based framework using Quartet reference materials for integrating long- and short-read RNA-seq. *bioRxiv*, 2025.2009.2015.676287 (2025).
2. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
3. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
4. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184-2185 (2012).
5. Pockrandt, C., Alzamel, M., Iliopoulos, C.S. & Reinert, K. GenMap: ultra-fast computation of genome mappability. *Bioinformatics* **36**, 3687-3692 (2020).
6. Pardo-Palacios, F.J. et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods* **21**, 1349-1363 (2024).
7. Trincado, J.L. et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol* **19**, 40 (2018).
8. Wang, Y. et al. rMATS-turbo: an efficient and flexible computational tool for alternative splicing analysis of large-scale RNA-seq data. *Nat Protoc* **19**, 1083-1104 (2024).
9. Vaquero-Garcia, J. et al. RNA splicing analysis using heterogeneous and large RNA-seq datasets. *Nat Commun* **14**, 1230 (2023).
10. Lin, K.T. & Krainer, A.R. PSI-Sigma: a comprehensive splicing-detection method for short- read and long-read RNA-seq analysis. *Bioinformatics* **35**, 5048-5054 (2019).