

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

Rui Zhang

ruizhang@ncccl.org.cn

Beijing Hospital <https://orcid.org/0000-0003-4660-2042>

Duo Wang

National Center for Clinical Laboratories, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing Hospital, Beijing, China. <https://orcid.org/0009-0000-5802-9226>

Jiaxin Zhao

Beijing Hospital

Qingwang Chen

State Key Laboratory of Genetic Engineering, Human Phenome Institute, School of Life Sciences and Shanghai Cancer Center, Fudan University <https://orcid.org/0000-0002-7191-5801>

Yanxi Han

National Center for Clinical Laboratories, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing Hospital

Yaqing Liu

State Key Laboratory of Genetic Engineering, School of Life Sciences and Human Phenome Institute, Fudan University, Shanghai, China. <https://orcid.org/0000-0001-9356-9164>

Yuanfeng Zhang

Beijing Hospital <https://orcid.org/0000-0003-2308-019X>

Cong Liu

National Center for Clinical Laboratories, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing Hospital/National Center of Gerontology

Wanwan Hou

Fudan University

Ying Yu

 <https://orcid.org/0000-0002-4084-908X>

Leming Shi

State Key Laboratory of Genetic Engineering, Human Phenome Institute and School of Life Sciences, Shanghai Cancer Center, Fudan University, Shanghai, China <https://orcid.org/0000-0002-2981-4150>

Yuanting Zheng

State Key Laboratory of Genetics and Development of Complex Phenotypes, Human Phenome Institute and School of Life Sciences, Fudan University, Shanghai, China. <https://orcid.org/0000-0003-4480-8303>

Jinming Li

Beijing Hospital <https://orcid.org/0000-0002-1476-3397>

Article

Keywords: RNA sequencing, Quartet, Alternative Splicing, Best practices, Benchmarking

Posted Date: November 5th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-7716867/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on August 8th, 2026. See the published version at <https://doi.org/10.1038/s41467-026-76380-z>.

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

Duo Wang^{1,2,3,#}, Jiabin Zhao^{1,2,4,#}, Qingwang Chen^{5,#}, Yanxi Han^{1,2,4}, Yaqing Liu⁵, Yuanfeng Zhang^{1,2,4}, Cong Liu^{1,2,4}, Wanwan Hou⁵, Ying Yu⁵, Leming Shi^{5,6,*}, Yuanting Zheng^{5,*}, Jinming Li^{1,2,4,*}, Rui Zhang^{1,2,4,*}

¹ National Center for Clinical Laboratories, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing Hospital/National Center of Gerontology, P. R. China

² Beijing Engineering Research Center of Laboratory Medicine, Beijing Hospital, Beijing, P. R. China

³ College of Life Sciences, University of Chinese Academy of Sciences, Beijing, P. R. China

⁴ National Center for Clinical Laboratories, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, P. R. China

⁵ State Key Laboratory of Genetics and Development of Complex Phenotypes, School of Life Sciences, Human Phenome Institute and Shanghai Cancer Center, Fudan University, Shanghai, China

⁶ International Human Phenome Institutes (Shanghai), Shanghai, China

[#] These authors contributed equally: Duo Wang, Jiabin Zhao, Qingwang Chen

*Corresponding authors:

Leming Shi (lemingshi@fudan.edu.cn)

Yuanting Zheng (zhengyuanting@fudan.edu.cn)

Jinming Li (jmli@nccl.org.cn)

Rui Zhang (ruizhang@nccl.org.cn)

Abstract

Previous limited characterization of RNA-seq accuracy in alternative splicing (AS) due to methodological diversity and lack of reference standards has left unclear how to achieve optimal performance for informed applications—an issue increasingly critical with the rise of long-read sequencing. To address this, we conducted the first large-scale reference-based benchmarking of AS detection accuracy across 42 laboratories and 159 analysis pipelines leveraging the Quartet reference materials. We show that high-quality RNA-seq enables relatively accurate isoform detection. Best practices for experimental and bioinformatic designs were identified, and achieved mean Pearson correlations of 0.82 for isoform quantification and Matthews correlation coefficients of 0.69 for differential expression analysis, representing improvements of 0.21–0.31 and 0.46–0.61 across laboratories, respectively, compared to the poorest workflows. AS event-level accuracy remained limited, yet the best tools still outperformed the poorest by 0.02–0.25 in event quantification and 0.10–0.25 in differential splicing analysis. Beyond technical variables, low expression levels were the primary constraint on isoform and event detection accuracy, followed by their compositional complexity. Collectively, this study provides practical guidance for maximizing AS profiling accuracy with existing methodologies, contributing to effective RNA-seq application in splicing research.

Keywords: RNA sequencing, Quartet, Alternative Splicing, Best practices, Benchmarking

Alternative splicing (AS) is a post-transcriptional process by which different combinations of splice sites are selectively utilized to produce multiple mRNA isoforms from a single gene, contributing to transcriptomic and proteomic diversity¹. Aberrant AS is linked with various diseases, particularly cancer, neurodegenerative disorders, and autoimmune diseases²⁻⁴. Thus, detection of AS is essential for unraveling developmental mechanisms, elucidating disease pathogenesis, and identifying biomarkers for disease diagnosis and therapy. RNA sequencing (RNA-seq) enables transcriptome-wide profiling of RNA splicing, providing quantification of AS-derived isoform abundances, measurement of AS events via percent spliced-in (PSI), and the comparative analysis of their differential expression across physiological and pathological conditions⁵. This technology has greatly advanced the study of RNA splicing⁶⁻⁹. Despite its widespread use, the relatively short read lengths of RNA-seq make it challenging to fully capture AS regions, and the limited knowledge regarding its accuracy hinders clinical translation. This issue has grown increasingly pressing as long-read sequencing technologies have demonstrated superior performance in resolving full-length isoforms^{10, 11}, raising critical questions about whether short-read RNA-seq (srRNA-seq) remains reliable for AS profiling in clinical applications.

Previous benchmarking studies have substantially advanced our understanding of RNA-seq-based AS detection. Owing to the lack of ground truths, most focused on evaluating consistency across sequencing platforms or analysis tools. Early efforts by the MicroArray/Sequencing Quality Control (MAQC/SEQC) consortium and the Association of Biomolecular Resource Facilities (ABRF), employed RNA reference materials in multi-center evaluations and revealed that isoform-level detection is more challenging than gene-level analysis^{12, 13}. Subsequent studies assessed the consistency of analysis tools for isoform quantification, AS event detection, and differential expression or splicing analyses, revealing particularly poor overlap at the AS event level¹⁴⁻¹⁹. To approximate accuracy, synthetic RNA spike-ins were introduced; for example, the Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) revealed that srRNA-seq achieved high quantification accuracy for Spike-In RNA Variants (SIRV)²⁰. Moreover, other evaluations of analysis tools have relied primarily

on simulated data, incorporating different combinations of tools, and revealing differences in accuracy particularly at the event level, while the resulting performance rankings were not always consistent²¹⁻²⁶. Despite these advances, fundamental questions remain unresolved: how accurate RNA-seq truly is for isoform and AS event detection, and how to achieve the highest possible accuracy using existing methods.

Specifically, previous consistency-based analyses have been unable to clearly reflect the overall accuracy of RNA-seq in AS detection or to determine which experimental protocols and bioinformatics tools achieve the highest practical accuracy. Furthermore, benchmarking using spike-ins or simulated data, which are far less complex than real human transcriptomes, tends to overlook influencing factors and may lead to overestimation of performance^{21, 27, 28}. Additionally, benchmarking at the event level has been particularly hindered, largely due to differences among AS detection tools in algorithms, event-type definitions, and output formats. Consequently, most prior studies have focused on performance at the splicing gene level where events occur^{18, 23, 24}, substantially limiting the understanding of event-level detection accuracy. When overall accuracy remains unclear, the factors affecting RNA-seq accuracy in AS detection are also poorly characterized. Therefore, reference-based benchmarking of RNA-seq accuracy in isoform and AS event profiling, along with standardized benchmarking at the AS event level, is essential for achieving accurate AS detection.

To address these limitations, we conducted a comprehensive benchmarking study of RNA-seq for AS analysis using the Quartet RNA reference materials²⁹, spanning 42 real-world laboratories and 159 bioinformatics pipelines. The well-validated Quartet reference datasets, established in our companion study under the Quartet project using long-read sequencing, served as the ground truth, covering 10,218 isoforms and 6,032 AS events³⁰. Additional ground truths were provided through RT-qPCR validation of 121 isoforms and 59 events, as well as the incorporation of known mixing ratios in the reference sample design. Leveraging these large-scale, multi-type ground truths, we systematically evaluated the accuracy and consistency of splice junction identification, isoform and AS event quantification, and differential expression or splicing analyses. We further provided best-practice recommendations by investigating technical and

111 biological factors affecting isoform- and event-level performance across multiple
112 centers, platforms, and pipelines, and by developing a standardized benchmarking
113 framework to harmonize AS event formats and evaluate four widely used methods
114 (SUPPA2, MAJIQ, PSI-Sigma, and rMATS). Together, this study facilitates achieving
115 optimal performance of RNA-seq in AS profiling and provides important guidance for
116 its effective application in both research and clinical settings.

Results

Study design

This study generated multi-center RNA-seq datasets involving 42 laboratories and 159 analysis pipelines for comprehensive RNA-seq benchmarking using the Quartet RNA reference materials (Fig.1a). The sample preparation and data generation were detailed in our previous study²⁸. The Quartet samples consisted of M8, F7, D5, and D6²⁹, and two additional samples, T1 and T2, were generated by mixing M8 and D6 at ratios of 3:1 and 1:3, respectively²⁸. This sample design, with additional RT-qPCR validation of 121 isoforms and 59 AS events, incorporated three types of ground truth for isoform and event level evaluations, including the Quartet reference datasets, RT-qPCR validated reference datasets, and the built-in truth based on the known mixing ratios. The 42 laboratories involved diverse RNA-seq workflows, including two mRNA enrichment methods, eight library preparation kits, and two sequencing platforms, thereby reflecting real-world practice. Subsequent 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 (Fig.1b). The detailed experimental and analysis workflow are illustrated in Fig.S1–2.

Leveraging these extensive datasets, we first evaluated the accuracy and consistency across laboratories and analysis pipelines, focusing on splice junction identification, isoform and AS event quantification, as well as the identification of differentially expressed isoforms (DEIs) and differential splicing events (DSEs). Then, we investigated key factors influencing accuracy, considering technical variables in both experimental and bioinformatics processes as well as biological features of isoforms and AS events, and developed a standardized benchmarking tool to harmonize AS event formats for direct evaluation of event-level detection tools, ultimately providing best-practice recommendations for experimental protocols and bioinformatics pipelines to achieve optimal accuracy with current methods (Fig.1a).

Quality control for identifying low-quality RNA-seq data

We examined the signal-to-noise ratio (SNR) of quantification data at the gene, isoform,

and event levels, which reflected the ability to distinguish biological signals from technical noise (Fig.S3a). SNR values varied widely across laboratories (gene level: 0.1–36.1; isoform level: 0–32.1; AS event level: 0–20.6). Notably, the mean SNR for AS events was significantly lower than that for isoform-level, and both were lower than gene-level (20.8, 18.4, and 10.6 for gene, isoform, and event level, respectively). This indicates that achieving quantitative accuracy becomes increasingly challenging when transitioning to resolution of isoforms or individual splicing events within each isoform.

By incorporating the SNR metric with previously used basic RNA-seq quality metrics²⁸, data from 19 laboratories met all quality criteria, while the remaining 23 laboratories were flagged as low quality (Fig.S3b and Supplementary Data 1). The low-quality classifications were primarily attributed to poor library preparation stability, low sequencing quality, and cross-contamination.

srRNA-seq missed certain splice junctions captured by long-read sequencing

Accurate junction-level resolution is essential for reliable AS detection using RNA-seq. To evaluate the accuracy in junction detection, we constructed junction reference datasets using seven batches of high-quality long-read RNA-seq generated in our companion study³⁰. These reference datasets resolved annotated junctions for 6,130 genes, including 48,366 positive and 35,766 negative junctions. In addition, 69,013 high-confidence novel positive junctions were identified (Fig.S4).

All laboratories demonstrated generally high accuracy for annotated junctions, with a mean F1 score of 0.968 (0.928–0.981) (Fig.S5a). The false-negative (FN) rate was significantly higher than the false-positive (FP) rate, with mean values of 4.8% (1.6–13%) and 1.5% (0.5–7.1%), respectively (Fig.S5b). For novel junctions, the mean FN rate was as high as 45.1% (28.4–72.4%), highlighting the limited ability of short-read sequencing to detect novel isoforms or events. High-quality laboratories consistently exhibited lower FN rates compared to low-quality laboratories (Fig.S5c).

Both technical and biological factors influenced junction detection in RNA-seq. Laboratories clustered according to sequencing technology, mRNA enrichment method, and library strandedness, highlighting protocol effects (Fig.S6). Sequencing depth was

a key determinant, with higher depth strongly reducing FN rates for both annotated and novel junctions (Fig.S5d). Junction-specific features associated with FN events included lower read counts, uneven coverage, longer junction-correlated isoforms, more exons per isoform, and lower mappability (Fig.S5e and Fig.S7–13), indicating that short reads struggle to capture junctions in complex isoforms or poorly mappable regions, especially at low depth. FP were linked to slightly higher GC content, lower mappability, and shorter isoforms with fewer exons, suggesting that misalignments may occur in specific genomic contexts.

High-quality srRNA-seq exhibited relatively accurate isoform-level detection

The performance in isoform detection across real-world laboratories was evaluated from both isoform quantification and DEI detection perspectives. For isoform quantification, substantial inter-laboratory variation was observed, with PCA analysis indicating that inter-laboratory differences often obscured biological signals, whereas high-quality laboratories clustered more clearly by sample (Fig.2a–b and Fig.S14). Benchmarking against the Quartet and RT-qPCR references demonstrated high accuracy, with mean PCC of 0.71 (0.61–0.77) and 0.91 (0.86–0.93), respectively, for high-quality laboratories (Fig.2c–d and Fig.S15). Higher RMSE and lower PCC against the Quartet reference datasets highlighted the greater challenge of accurately quantifying a broader, more complex isoform set. The built-in truth was used to evaluate the laboratories' ability to recover the expected mixing ratios (Methods), serving as an indicator of both accuracy and reproducibility (Fig.S16), showing trends consistent with evaluations against both reference datasets (Fig.2c). Moreover, isoform quantification accuracy correlated positively with sequencing depth and negatively with junction detection errors, as higher FN junction rates were associated with poorer accuracy (Fig.2e–f).

For DEI identification, laboratories varied widely in the number of detected DEIs, largely driven by sequencing depth, which positively correlated with DEI detection (Fig.S17). Benchmarking against the Quartet and RT-qPCR reference datasets showed generally high precision of 0.81 (0.38–0.88) and 0.87 (0.76–1), respectively, but recall

remained modest at 0.38 (0.005–0.61) and 0.50 (0.10–0.71), yielding moderate Matthews correlation coefficients (MCC) of 0.51 (0.03–0.65) and 0.59 (0.28–0.69) (Fig.2g–h). Recall improved with higher depth, and MCC positively correlated with data quality (Fig.2i–j). Tracing all FN DEIs revealed that FNs were characterized by low coverage of reads and high inter-replicate coefficient of variation (CV) (Fig.2k), where low coverage caused inaccurate isoform quantification and poor reproducibility undermined statistical robustness of differential expression analysis.

Identifying technical and biological factors influencing isoform detection for best practices

To assess the technical and biological determinants of isoform detection performance, we examined 19 high-quality laboratories with different experimental protocols and 138 analysis pipelines to identify best practices (Fig.1 and Table 1). We then assessed various isoform features to evaluate their specific impact on RNA-seq methodologies, including the K-value (indicating isoform-exon structure complexity), isoform length, the number of isoforms per gene, exon length, the number of exons per isoform, GC content, and their expression levels.

Principal variance component analysis (PVCA) revealed that 11 experimental factors contributed significantly more to the variation than biological differences between Quartet and mixed samples (69.4% vs. 6.8%) (Fig.3a and Supplementary Data 2). mRNA enrichment method and strandedness accounted for 42.3% of the variation, with read length, library preparation kits, and sequencing depth also contributing. Cluster analysis of laboratories based on accuracy metrics against three ground truths showed that the highest accuracy was associated with Poly(A) selection and higher sequencing depth, while three laboratories (lab04, lab33, and lab36) using rRNA depletion also performed well due to exceptionally high sequencing depth (Fig.3b). No clear clustering was observed based on strandedness, library preparation kits, or read length.

The 138 analysis pipelines included 18 isoform quantification pipelines, with the accuracy of isoform quantification evaluated at the relative expression level ($\log_2FC$)

(Fig.3c and Fig.S18). Except for Subjunc-StringTie, all tools showed high agreement with RT-qPCR (mean PCC > 0.9), while the Quartet reference datasets revealed clearer distinctions among pipelines. featureCounts, Bowtie2-eXpress, and Bowtie2-Salmon struggled to distinguish isoforms due to featureCounts distributing reads from shared exons equally across isoforms and Bowtie2's untrimmed alignment discarding isoform-informative reads (Fig.3d), making them unsuitable for isoform-level analysis. Their apparently higher PCC and lower RMSE are primarily due to the dominance of MANE transcripts in the reference datasets and the use of relative expression calculations, which mask their limited ability to differentiate isoforms. Cuffdiff and STAR-eXpress demonstrated relatively superior accuracy, followed by RSEM, StringTie, and pseudo-alignment-based approaches (Salmon, Sailfish, Kallisto) (Fig.3c–e). Built-in truth confirmed these results, with StringTie achieving RMSE comparable to Cuffdiff, likely due to its higher reproducibility (Fig.3c and Fig.S19). Further evaluation of SNR and inter-replicate reproducibility revealed consistent trends (Fig.S19–20). Overall, Cuffdiff and STAR-eXpress emerged as the top-performing pipelines, achieving mean PCC of 0.82 (0.53–0.90) and 0.92 (0.72–0.95) against Quartet and RT-qPCR, respectively, with inter-laboratory improvements of 0.26 (0.21–0.31) and 0.15 (0.07–0.22) over the poorest pipelines.

For DEI identification, the 138 analysis pipelines, comprising 11 differential expression methods, were benchmarked against three ground truths. DESeq2 consistently achieved slightly higher MCC against the Quartet and RT-qPCR references, while edgeR, limma, and Cuffdiff showed similarly high MCC (Fig.3f–g). For edgeR, the quasi-likelihood method (edgeR-QL) was slightly less affected by quantification methods than edgeR-qCML and edgeR-GLM (Fig.S21). NOISeq-BIO reached comparably high MCC, particularly versus RT-qPCR, favoring recall over precision, whereas NOISeq-real was more sensitive to the quantification pipeline, favoring precision at the cost of recall (Fig.S22). Sleuth showed high precision but low recall, resulting in lower MCC, and Ballgown performed poorly overall (Fig.S22–23). Overall, each analysis step influenced DEI detection accuracy, with the top-performing pipeline reaching MCC of 0.69 (0.57–0.78) and 0.71 (0.67–0.76) for the Quartet and RT-qPCR

references, respectively, representing inter-laboratory improvements of 0.54 (0.46–0.61) and 0.49 (0.34–0.59) over the poorest pipeline. A comprehensive ranking of all pipelines is provided in [Supplementary Data 3](#).

Isoform-inherent features determine their detectability by RNA-seq, reflecting methodological limitations. Benchmarking against the Quartet reference datasets, most quantification pipelines performed poorly with high K-values, short isoforms, large isoform numbers per gene, long exon lengths, high GC content, and low expression levels, highlighting challenges in quantifying complex or lowly expressed isoforms ([Fig.3h](#) and [Fig.S24–30](#)). Notable exceptions included featureCounts, which showed reduced accuracy for highly expressed isoforms due to its approach of distributing reads from shared exons across isoforms, thereby affecting the quantification of other isoforms sharing these exons when detecting highly expressed isoforms. The effect of exon number varied by pipeline, with Cuffdiff and StringTie performing poorly for isoforms with more exons, while other tools showed the lowest performance at intermediate exon numbers. Using the built-in truth as a benchmark, these findings were largely consistent, except for featureCounts and pipelines employing Bowtie2 aligner ([Fig.S24–31](#)). Finally, ANOVA-based analysis revealed isoform expression levels as the primary factor affecting most pipelines, followed by isoform complexity such as K-value and isoform number ([Fig.3i](#)). Additionally, featureCounts, Bowtie2-eXpress, and Bowtie2-Salmon were significantly influenced by exon length, isoform length, and isoform number.

Limited performance of srRNA-seq in AS event detection with SUPPA2

The performance of RNA-seq in AS event detection across real-world laboratories was evaluated from both event quantification and DSE detection perspectives, using SUPPA2 to infer AS events from isoform expression data. Substantial variation was observed even among 19 high-quality laboratories, with PCA showing clustering by laboratory rather than sample groups ([Fig.4a](#)), largely reflecting differences in experimental protocols ([Fig.S32](#)). Accuracy was generally low and highly variable ([Fig.4b](#)), with high-quality laboratories showing mean PCC of only 0.21 (0.16–0.25)

and 0.47 (0.32–0.58) with the Quartet and RT-qPCR reference datasets, respectively (Fig.4c). Quantification accuracy improved with greater sequencing depth and was closely linked to junction detection, with higher FN junction rates corresponding to lower PCC (Fig.4d–e). Overall, RNA-seq exhibited limited AS event quantification performance, leading to poor inter-laboratory consistency: only 4% were detected across all laboratories (Fig.S33). Such reduced inter-laboratory consistency was associated with threshold effects at extreme PSI values, decreased reproducibility of PSI estimates, and low isoform expression levels (Fig.S34).

For DSE identification, benchmarking against the Quartet and RT-qPCR reference datasets revealed very low recall and precision for DSE detection, with mean MCC of 0.16 (0.05–0.24) and 0.12 (0–0.28), respectively (Fig.4f–g). The number of detected DSEs varied across laboratories, showing a strong dependence on sequencing depth (Fig. S35), with the corresponding recall of DSEs also positively correlated with depth (Fig.4h).

Identifying technical and biological factors influencing AS event detection for best practices with SUPPA2

To investigate factors influencing AS event detection and identify best practices, we evaluated 19 experimental protocols and 18 SUPPA2-based event analysis pipelines (Fig.1 and Table 1). We further assessed the impact of adjacent AS event number, AS-related isoform number, GC content, mappability, PSI level, and isoform expression on detection performance.

Among experimental factors, the mRNA enrichment method and strandedness were the primary sources of inter-laboratory variation in AS event quantification, accounting for 32% of the variance (Fig.5a), with library preparation kits, sequencing depth, and read length also contributing. Clustering high-quality laboratories based on RMSE showed that experimental protocols yielding the highest AS event quantification accuracy were consistent with those optimal for isoform expression. Specifically, the most accurate cluster included laboratories using Poly(A) selection with higher sequencing depth, except for two rRNA-depletion laboratories (lab33 and lab36) with

exceptionally high sequencing depth (Fig.5b).

In bioinformatics processes, the choice of isoform quantification method significantly affected SUPPA2's performance in detecting AS events. Benchmarking against the Quartet and RT-qPCR datasets revealed that SUPPA2, when combined with Salmon, RSEM, Cuffdiff, Sailfish, or Kallisto, achieved relatively higher dPSI quantification accuracy, though the maximum mean PCC did not exceed 0.36 and 0.54, respectively (Fig.5c). Combinations with featureCounts, Bowtie2-eXpress, and Bowtie2-Salmon performed the worst, with aberrant dPSI distributions near zero, reinforcing our conclusion of their limited isoform-level resolution (Fig.5d and Fig.S36). Normalized RMSE (nRMSE) was applied to mitigate the impact of aberrant distributions, showing consistent trends with PCC (Fig.5c). For DSE detection, pipelines with higher dPSI quantification accuracy also achieved greater detection accuracy, with maximum mean MCC values of 0.41 and 0.27 (Fig.5e). Their accuracy generally corresponded to the quality of AS event quantification data (SNR), except for StringTie, which had high SNR but lower accuracy (Fig.S37). Overall, despite generally low accuracy, the best-performing pipelines improved PCC by 0.18 (0.09–0.28) and 0.15 (0.02–0.29), and MCC by 0.31 (0.15–0.44) and 0.24 (0.03–0.45), compared with the poorest pipelines, based on the Quartet and RT-qPCR datasets.

Inherent features of AS events also constrained detection accuracy. Benchmarking with the Quartet reference datasets revealed that SUPPA2's quantification accuracy decreased for events with more adjacent AS events, fewer associated isoforms, higher GC content, lower mappability, intermediate PSI values, and lower expression levels of related isoforms (Fig.5f–g and Fig.S38–43). Notably, when SUPPA2 was combined with featureCounts, the influence of these factors followed different trends, consistent with observations at the isoform quantification level. The effects of these factors on reproducibility were similar to those on accuracy (Fig.S38–43). Finally, we quantified the magnitude of the impact on quantification accuracy and found that, except for pipelines combined with featureCounts, Bowtie2-eXpress, or Bowtie2-Salmon, the primary determinants were the expression levels of associated isoforms and the PSI levels (Fig.5f).

Evaluation of multiple AS event detection methods using a standardized benchmarking tool

Different AS event detection tools employ distinct algorithmic strategies, leading to divergent definitions and output formats for each event type. To address this, we developed a standardized benchmarking tool for direct evaluation across tools, including four tools using distinct principles: SUPPA2 based on isoform expression, MAJIQ quantifying local splicing variations (LSVs) from junctions, and PSI-Sigma and rMATS using junction read counts. A unified format for seven AS event types were defined (Fig.S44), including skipping exon (SE), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), alternative first exon (AF), alternative last exon (AL), mutually exclusive exon (MX), and retained intron (RI).

These tools showed limited overlap in detected AS events, especially for AF, AL, MX, and RI (Fig.6a and Fig.S45–48). The overlap in detected DSEs was even smaller, with most identified by only a single tool (Fig.6b and Fig.S49–54). In dPSI quantification, PSI-Sigma and MAJIQ exhibited the highest consistency, followed by rMATS, while SUPPA2 showed poorer consistency with the other tools (Fig.6c and Fig.S55–56). This is due to SUPPA2's unique approach, deriving AS events indirectly from isoform expression, unlike the junction-centered methods of the other tools. Consistency was highest for SE, A5SS, and A3SS, moderate for RI, and lowest for MX, AF, and AL.

Benchmarking against the Quartet and RT-qPCR reference datasets confirmed that MAJIQ and PSI-Sigma, which showed higher dPSI consistency, also exhibited greater agreement with both references datasets (Fig.6d). MAJIQ achieved mean PCC of 0.45 and 0.75 with the Quartet and RT-qPCR reference datasets, respectively, whereas PSI-Sigma, SUPPA2, and rMATS reached 0.43, 0.36, and 0.35 (Quartet reference) and 0.65, 0.61, and 0.56 (RT-qPCR). The PCC gap between best and poorest tools reached 0.10 (0.02–0.25) and 0.19 (0.002–0.35) across laboratories. For DSE detection, all tools performed poorly (Fig.6e). SUPPA2 balanced recall and precision, MAJIQ and PSI-Sigma favored precision over recall, while rMATS prioritized recall at the expense of

387 precision. Overall, SUPPA2 achieved the highest mean MCC of 0.41 (Quartet
388 reference), surpassing rMATS, PSI-Sigma, and MAJIQ (MCC = 0.25), representing an
389 inter-laboratory difference of 0.16 (0.10–0.25) ([Fig.S57](#)). Evaluation against RT-qPCR
390 showed comparable MCC across tools, with SUPPA2 slightly better, likely due to the
391 limited number of experimentally validated true positives reducing statistical power.

Discussion

This study represents the most extensive cross-laboratory RNA-seq benchmarking for accurate AS profiling, leveraging the Quartet reference materials and large-scale, multi-type ground truths. Our objectives were to address key challenges in RNA-seq applications: (i) the accuracy level of using RNA-seq for AS detection, (ii) best practices for experimental protocols and bioinformatic pipelines to achieve the most accurate isoform- and AS event-level detection using existing methods. The final recommendations are summarized in Table 1.

This study provided reference-based evidence for the AS detection accuracy across real-world laboratories, experimental protocols, and bioinformatics pipelines. We utilized the Quartet reference datasets derived from long-read sequencing and the largest AS-level RT-qPCR validation datasets to date, both were recommended approaches for establishing standards²⁷, and superior to previously used computational simulations or synthetic RNA²⁰⁻²⁵. Furthermore, incorporating the built-in truth based on known mixing ratios further allowed evaluation across all detected features^{31, 32}. Performance rankings derived from the built-in truth were largely consistent with those from the Quartet and RT-qPCR datasets, supporting the generalizability of conclusions for features included in these reference datasets to all transcriptome features. Notably, the Quartet reference datasets provided the clearest discrimination in accuracy metrics.

High-quality RNA-seq generally achieves high isoform-level accuracy, with optimal workflows reaching mean PCC of 0.81 and 0.92 for isoform quantification, and mean MCC of 0.69 and 0.71 for DEI detection, based on the Quartet and RT-qPCR references, even close to previously reported gene-level accuracy²⁸. Importantly, our evaluation focused primarily on the annotated transcriptome, and for novel isoforms, the high FN rate of up to 45.1% for novel junctions indicates limited detection capability. In contrast, AS event detection remains poor, with the best SUPPA2-based workflows achieving mean PCC of 0.36 and 0.50 for event quantification, and mean MCC of 0.41 and 0.21 for DSE detection, respectively, even lower than previous reports using simulated or limited RT-qPCR datasets^{18, 24, 33}. Overall, RNA-seq is reliable for annotated isoform detection but remains limited for AS event analysis.

Our benchmarking of RNA-seq provides best-practice recommendations for experimental and bioinformatics design in future studies (Table 1). Unlike most previous evaluations that focused on isolated experimental methods or individual analysis steps^{14-18, 21-25}, our approach encompassed entire experimental and computational workflows across multiple centers. This comprehensive perspective enables full consideration of interactions between different steps and provides direct, practical guidance for isoform- and AS event- level profiling.

For isoform detection, the Poly(A) selection method and higher sequencing depth were strongly associated with greater quantification accuracy, consistent with previous reports that Poly(A) selection provides superior gene quantification in high-quality RNA samples^{28, 34, 35}. While the benefits of increased depth are well recognized, the appropriate depth depends on transcriptome complexity, sequencing mode, library type, and RNA quality. Liu et al. indicated that 100M reads are sufficient for AS detection, whereas even 300–400M reads may not saturate differential expression analysis³⁶; our results corroborate this, as laboratories exceeding 400M reads still exhibited performance gains. In bioinformatics design, STAR and HISAT2 were superior aligners, offering high mapping efficiency²⁸, and comparable accuracy when combined with other analysis tools. Cuffdiff and eXpress performed better for isoform quantification. Our findings are consistent with those of Dapas et al., who reported that eXpress exhibited higher agreement with RT-qPCR¹⁶, but differ from Zhang et al., who found that pseudo-alignment tools performed best using simulated data²². This discrepancy likely reflects the simplicity of simulated data, which lack variants, sequencing errors, intronic signals, and non-uniform coverage, thereby reducing the challenge of pseudo-alignment, as also noted by Sarantopoulou et al.²⁶. Other studies have recommended RSEM for its reproducibility^{14, 25}, consistent with our results. For DEI identification, DESeq2, edgeR, limma, Cuffdiff, and NOIseq-BIO generally performed well. Previous evaluations of differential expression tools have mainly focused on the gene level^{23, 28, 37-39}, we extended to isoform-level analysis and revealed substantial overlap with previously recommended tools. Overall, identifying optimal pipelines helps achieve the highest accuracy, raising mean PCC for isoform quantification by 0.26 and 0.15, and

mean MCC for DEI detection by 0.54 and 0.49 over the poorest pipelines, based on the Quartet and RT-qPCR datasets. Importantly, irrespective of experimental or bioinformatics choices, certain factors consistently constrain the performance of RNA-seq methodologies, including low expression levels and high isoform complexity, representing a common challenge that may inform software development and optimization efforts.

For AS event detection, experimental protocols that performed well at the isoform level also showed relatively higher accuracy in event quantification. Regarding bioinformatics design, we demonstrated that SUPPA2 achieved higher accuracy when combined with upstream quantifiers such as Salmon, Sailfish, Kallisto, Cuffdiff, or RSEM, an aspect not systematically assessed before. The optimal pipeline improved mean PCC for event quantification by 0.18 and 0.15, and mean MCC for DSE detection by 0.31 and 0.24 over the poorest-performing pipeline. Nonetheless, the highest PCC with RT-qPCR was only 0.55 (0.42–0.66), markedly lower than the previously reported 0.68³³, likely due to the larger number and broader dPSI range of events in our RT-qPCR validation. Moreover, AS detection is also constrained by methodological limitations, with reduced accuracy for events associated with low isoform expression, intermediate PSI values, or high complexity.

To establish best practices for different AS event detection tools, we developed a standardized benchmarking tool that harmonizes event formats, enabling evaluation across four widely used tools. Although the earlier DICAST served as a benchmarking tool for AS event analysis¹⁵, it included tools that have been shown to perform poorly²⁴, did not cover the most widely used tools currently in practice, and was incapable of quantitative evaluation. We incorporated SUPPA2, rMATS, MAJIQ, and PSI-Sigma, representing several distinct AS event algorithmic strategies of isoform expression-based, LSV-based, and junction-based approaches⁴⁰. The four tools exhibited remarkably little overlap, largely due to differences in the event databases inferred from reference transcriptomes and in the algorithms used to define and quantify each event type. Despite overall weak performance, MAJIQ and PSI-Sigma performed better in event quantification, whereas SUPPA2 performed better in DSE detection with a more

balanced recall–precision trade-off, consistent with previous evaluations based on simulated data at both the splicing gene or event levels²⁴. Complex event types such as MX, RI, and AF/AL remained particularly challenging due to spanning multiple exons, long introns, or terminal exons that are prone to capture bias and uneven coverage. Importantly, our standardized benchmarking tool can be extended to additional tools, promoting standardization of event-level analyses.

Our evaluations based on the Quartet reference datasets also provide a perspective on the comparison between srRNA-seq and long-read sequencing. Together with recent benchmarking efforts for long-read RNA-seq^{20, 30, 41, 42}, our study provides a comprehensive insight of current AS detection technologies, guiding users in choosing between short- and long-read sequencing for splicing studies. For isoform-level analysis, we confirmed high concordance of srRNA-seq with long-read technologies, particularly for well-annotated transcriptomes. The advantages of greater sequencing depth and lower cost further highlight its practical utility. Nevertheless, further algorithmic improvements are required, especially to enhance the detection of low-expression and structurally complex isoforms²⁰. At the AS event level, we demonstrated that srRNA-seq exhibits overall limited performance. Given the current scarcity of long-read-specific analysis tools for event-level analysis, the development and optimization srRNA-seq remain essential. Potential directions include establishing comprehensive AS annotation databases, standardizing outputs across different algorithms, combining multiple tools to improve DSE detection, and targeted optimization for complex event types.

Overall, this study elucidates RNA-seq accuracy in AS detection, identifies key technical determinants, and provides best practice recommendations for experimental and bioinformatic designs (Table 1). Its insights will support the effective application of RNA-seq and advance research on alternative splicing, ultimately contributing to a deeper understanding of transcriptomic complexity and its functional implications.

Online Methods

RNA reference samples preparation

Quartet RNA reference materials were obtained from four Epstein–Barr virus–immortalized B-lymphoblastoid cell lines derived from the Quartet family, which have been certified by China’s State Administration for Market Regulation²⁹. The details of RNA reference sample preparation were provided in our previous study²⁸. In brief, four Quartet samples (M8, F7, D5, and D6) were used, and two additional samples (T1 and T2) were generated as defined mixtures of M8 and D6 at ratios of 3:1 and 1:3, respectively. Each sample was provided in triplicate to 42 laboratories for sequencing.

RNA-seq data

Details of RNA-seq data generation are available in our previous study²⁸. Initially, 45 laboratories were enrolled, of which 42 provided complete sequencing data and were included in this study. Each laboratory applied its own routine library preparation and sequencing protocol, thereby reflecting real-world RNA-seq practices. Laboratory identifiers were retained as lab01–lab45 for consistency with previous studies.

RT-qPCR validation

A total of 121 isoforms and 59 AS events included in the Quartet reference datasets were selected for validation. SYBR Green dye-based RT-qPCR was used to validate the quantification of isoforms and AS events. Specific primers were designed to target isoform-unique sequences or sequences corresponding to isoforms that either include or exclude the AS event. For cases where dye-based RT-qPCR results were inconsistent with the Quartet reference datasets, additional validation was performed using the TaqMan probe method. The TaqMan probes were designed on isoform-unique sequences or sequences specific to AS event-related isoforms. C1ORF43 was selected as the reference for the PCR method. Primers and probes were synthesized by Sangon Biotech, and the sequences are shown in [Supplementary Data 4](#).

The RT-qPCR experiment mainly consisted of reverse transcription and PCR steps. Five µg of each RNA sample was reverse transcribed using the PrimeScript™ RT reagent Kit (RR037A, Takara) in a 50 µl reaction. This reaction mixture was incubated at 37°C for 15 min, then for 5 s at 85°C and finally for termination at 4°C. cDNA was

used as a template for qPCR. The dye-based qPCR reactions were performed in a final volume of 20 μ L using Premix Ex TaqTM II (Tli RNaseH Plus) (RR820A, Takara) containing 1 μ L of cDNA, and 0.8 μ L of each forward and reverse primer. The probe-based qPCR reactions were carried out using Premix Ex TaqTM (RR390A, Takara) containing 2 μ L of cDNA, 0.4 μ L of each forward and reverse primer, and 0.8 μ L of TaqMan probes in a 20 μ L final volume reaction. The qPCR was performed on an Applied Biosystems 7500 Real-Time PCR System using the following cycling conditions: 30 s at 95°C followed by 45 cycles of 5 s at 95°C and 34 s at 56°C. Three replicates per sample per gene were conducted for eliminating random variations.

Normalized isoform expression was quantified using the delta Ct method, with the C1ORF43 gene serving as the endogenous control. Fold changes (FC) for isoform expression were calculated using the delta delta Ct method for the three Quartet sample comparisons (M8/D6, F7/D6, and D5/D6). PSI for AS event was calculated as the ratio of the expression level of isoforms containing a specific AS event to the total expression of all isoforms, and dPSI was calculated as the difference in PSI values between the three sample pairs. An isoform is classified as a DEI when the student's t-test $p < 0.05$ and $FC \geq 2$ or ≤ 0.5 . An event is classified as a DSE when the student's t-test $p < 0.05$ and $dPSI \geq 0.05$ or ≤ -0.05 .

Construction of the junction reference datasets

Seven batches of high-quality long-read sequencing data from our companion study³⁰ were used to build the junction reference datasets through three steps: identifying high-confidence junctions, selecting candidate gene sets with sufficient coverage, and generating the final datasets. Detailed data processing and building steps are available in the Supplementary Methods.

The ground truth inherent in RNA reference samples

The sample design incorporated three types of ground truth: the Quartet reference datasets, the RT-qPCR reference datasets, and the built-in truth based on the mixing ratios of the T1 and T2 samples.

Quartet reference datasets. The details for constructing the Quartet reference datasets are available in our companion study³⁰. These datasets provide isoform-level and AS

event-level relative quantification results, as well as differential analysis results, for the M8/D6, F7/D6, and D5/D6 sample pairs. The isoform reference dataset comprises 10,218 results, reporting relative expression as log₂FC, among which 1,077 isoforms were classified as DEIs, enabling evaluation of RNA-seq accuracy in isoform quantification and DEI identification. The AS event reference datasets contain 5,574 results, reporting relative quantification as dPSI, among which 78 events were defined as DSEs, providing a basis for evaluating the accuracy of AS event quantification and DSE detection.

RT-qPCR reference datasets. The selected isoforms and AS events were validated across the M8/D6, F7/D6, and D5/D6 sample pairs, yielding a total of 362 and 177 relative quantification results for isoforms (log₂FC) and AS events (dPSI), respectively. Among them, 78 were identified as DEIs and 41 as DSEs. The RT-qPCR reference datasets are shown in [Supplementary Data 5](#).

Built-in truth of mixing ratios in samples T1 and T2. The built-in truth can be defined at two levels. First, the relative expression of isoforms in the mixed sample pairs of T1/D6 and T2/D6 is expected to follow a defined functional relationship with that of M8/D6, which can be used to evaluate the performance of isoform quantification. The expected relative expression in comparisons of T1/D6 or T2/D6 could be calculated from the relative expression observed in the comparison of M8/D6 using Eq. (1). The detailed computational procedure is described in our previous study²⁸.

$$\begin{aligned} \log_2 E_{\frac{T1}{D6}} &= \log_2 \left(k1 + k2 \times 2^{\log_2 E_{\frac{M8}{D6}}} \right) \\ \log_2 E_{\frac{T2}{D6}} &= \log_2 \left(k2 + k1 \times 2^{\log_2 E_{\frac{M8}{D6}}} \right) \end{aligned} \quad (1)$$

where $E_{\frac{T1}{D6}}$ and $E_{\frac{T2}{D6}}$ represents the expected relative isoform expression in comparisons of T1/D6 or T2/D6, and $E_{\frac{M8}{D6}}$ represents the relative isoform expression in the comparison of M8/D6. Here, k1 and k2 represent the mixing coefficients. After correction for the total mRNA proportions in the samples, the coefficients were estimated to be 0.245 and 0.745 for the T1/D6 comparison, and 0.740 and 0.240 for the T2/D6 comparison.

Construction of the Quartet 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)³⁰. 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.

Relative expression calculation

Relative expression quantification was performed on a feature-by-feature basis within each laboratory. Specifically, relative expressions were calculated based on log₂FPKM values. For each gene, the mean of expression profiles of replicates of reference sample(s) (for example, D6) was first calculated and then were subtracted from the log₂FPKM values of that gene in other samples.

Evaluation of RNA-seq performance in isoform and AS event detection

The evaluation of RNA-seq performance was conducted at three levels. First, data quality was assessed using the SNR. Second, metrics reflecting the accuracy of quantification and differential analysis results for isoforms and AS events were calculated based on the three types of ground truth. Finally, reproducibility metrics were included to assess the consistency of quantification across replicate samples.

SNR. The PCA-based SNR was used to assess the data quality of the Quartet and mixed samples at the isoform expression and AS event PSI level, and the calculation method of PCA-based SNR is as shown in the previous study²⁹. Isoform-level analysis was based on FPKM values, including only those isoforms with FPKM > 0 in at least one sample. For AS event-level analysis, PSI values were used, and only events with PSI values between 0 and 1 in all samples were included in the analysis. An SNR below 10 was considered indicative of low-quality data. In addition to calculating the SNR across all 18 samples, we also computed SNR₁₇, representing the SNR calculated from any 17 out of the 18 samples (12 Quartet and 6 mixed samples). If the removal of a particular

sample results in a significant increase in SNR17 compared to SNR18 (with a difference greater than 6), the quality of the RNA-seq data of the excluded sample is considered low, and the sample is treated as an outlier. Such random occurrences of problematic library preparation or sequencing in individual samples are referred to as random failures.

PCC. PCC was used to assess the consistency between RNA-seq-derived relative isoform expression ($\log_2FC$) and AS event relative quantification (dPSI) with the reference datasets.

RMSE and nRMSE. RMSE was used to quantify the absolute deviation between RNA-seq-derived relative quantification of isoforms and AS events and the reference datasets. In addition, RMSE was also used to assess the consistency between the observed relative isoform expression in the sample pairs of T1/D6 and T2/D6 and the expected relative expression calculated from the sample pair of M8/D6 using Eq. (1). Normalized RMSE (nRMSE) was obtained by adjusting RMSE with the range of observed values using Eq. (2), enabling direct comparison across different analysis pipelines.

$$nRMSE = \frac{\sqrt{\frac{1}{l} \sum_{i=1}^l (\theta_i - \hat{\theta}_i)^2}}{\hat{\theta}_{max} - \hat{\theta}_{min}} \quad (2)$$

where θ_i and $\hat{\theta}_i$ denote the expected and observed quantification results, respectively, and $\hat{\theta}_{max}$ and $\hat{\theta}_{min}$ denote the maximum and minimum observed quantification values.

Recall, precision, and F1 score. To evaluate the performance of RNA-seq in detecting junctions, DEIs, and DSEs, recall, precision, and F1 score were calculated based on the comparison with the reference datasets. Recall was defined as the proportion of true positives (TP) correctly identified among all reference positives (TP + FN), while precision was defined as the proportion of TP among all detected positives (TP + FP). The F1 score, representing the harmonic mean of recall and precision, was calculated using Eq. (3).

$$F1 \text{ score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (3)$$

MCC. The MCC was used to evaluate the accuracy of DEI or DSE identification for a given pair of samples based on the Quartet and RT-qPCR reference datasets. Then MCC was calculated using Eq. (4).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

CV and SD. The reproducibility of isoform expression (FPKM) and AS event quantification (PSI) across three replicates was assessed using coefficient of variation (CV) and standard deviation (SD), respectively. Values were summed across all isoforms and AS events to provide an overall measure of reproducibility.

Uniform RNA-seq data analysis pipeline

The GRCh38 reference genome (https://ftp.ensembl.org/pub/release-109/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz) and transcriptome (https://ftp.ensembl.org/pub/release-109/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz) were obtained from Ensembl. Gene annotation from Ensembl release-109 (https://ftp.ensembl.org/pub/release-109/gtf/homo_sapiens/Homo_sapiens.GRCh38.109.gtf.gz) was used.

To evaluate the performance of RNA-seq experimental processes, RNA-seq data from all 42 laboratories were processed using a uniform data analysis pipeline, thereby eliminating variations introduced from the bioinformatics processes. Preliminary processing of raw reads was performed using fastp (v0.23.2) to remove adapter sequences⁴³. Cleaned reads were aligned to the GRCh38 genome using STAR (v2.7.10b)⁴⁴. The STAR --twopassMode Basic option was applied to improve junction detection, and splice junctions were extracted from BAM files using the sjFromSAMcollapseUandM_inclOverlaps.awk script provided by STAR. Isoform quantification was conducted with StringTie (v3.0.0) using the -e option to enable expression estimation guided by the annotation⁴⁵. Log₂ transformation of isoform expression values was carried out on Fragments Per Kilobase of transcript per Million mapped reads (FPKM), after adding a pseudocount of 0.01 to avoid infinite values. Differential isoform expression analysis was performed using edgeR (v4.4.2) in the

classical mode⁴⁶. Isoforms with log₂FC greater than ≥ 1 or ≤ -1 and FDR < 0.05 were classified as DEIs. For AS event analysis, StringTie-derived TPM data were used as input to SUPPA2 (v2.4)³³. AS events with dPSI ≥ 0.05 or ≤ -0.05 and *p-value* < 0.05 were considered DSEs.

Combined RNA-seq data analysis pipelines

To specifically assess the impact of bioinformatics designs, we incorporated four alignment tools, eight isoform quantification tools, ten DEI detection tools, and four AS event detection tools to combine 138 isoform-level and 21 AS event-level analysis pipelines. These pipelines were applied to high-quality RNA-seq data. Unless otherwise specified, all tools were run with default parameters. The details of RNA-seq analysis tools, versions, and the command line are listed in [Supplementary Data 6](#).

Alignment tools. Four alignment tools, STAR (v2.7.10b)⁴⁴, HISAT2 (v2.2.1)⁴⁷, and Subjunc (v2.0.8)⁴⁸, and Bowtie2 (v2.5.4)⁴⁹ were included. STAR, HISAT2, and Subjunc were applied to align reads to the reference genome, while Bowtie2 was used to align reads to the reference transcriptome. In addition, STAR was also employed for transcriptome alignment using the parameters `--quantMode TranscriptomeSAM` and `--quantTranscriptomeBan Singleend`. For Cuffdiff analysis, STAR was run with the additional parameters `--outSAMstrandField intronMotif` and `--outFilterIntronMotifs RemoveNoncanonical`, while HISAT2 was run with the `--dta-cufflinks` option. Subjunc alignments were not used for Cuffdiff analysis, as it does not provide parameters compatible with Cuffdiff.

Isoform quantification tools. Eight commonly used isoform quantification tools consist of genome-alignment-based tools like featureCounts (v2.0.8)⁵⁰, StringTie (v3.0.0)⁵¹, and Cuffdiff (v2.2.1)⁵², transcriptome-alignment-based tools like RSEM (v1.3.3)⁵³, eXpress (v1.5.1)⁵⁴, and Salmon (v1.10.1)⁵⁵ in alignment-based mode, as well as pseudo-alignment-based tools like Kallisto (v0.51.1)⁵⁶, Sailfish (v0.10.0)⁵⁷, and Salmon (v1.10.1)⁵⁵ in mapping-based mode. FeatureCounts was run with the additional parameters `-g transcript_id`, `-O`, and `-M`. RSEM performs both alignment and quantification directly by specifying the `--hisat2-hca`, `--star`, or `--bowtie2` options, rather than requiring a separate alignment step. Salmon was run using optional parameters: --

gcBias, --seqBias, and --posBias to correct GC biases, sequence-specific biases, and 5' or 3' positional bias.

Differential isoform expression analysis tools. Seven differential expression analysis tools, including Cuffdiff (v2.2.1)⁵², Ballgown (v2.38.0)⁵⁸, edgeR (v4.4.2)⁵⁹, limma (v3.62.2)⁶⁰, DESeq2 (v1.46.0)⁶¹, NOIseq (v2.50.0)⁶², and Sleuth (v0.30.1)⁶³, along with different modes, were combined to generate a total of 11 analysis approaches. Among these tools, edgeR and DESeq2 are based on negative binomial models, Cuffdiff employs a beta-negative binomial model, and NOISeq uses a non-parametric approach. Ballgown and limma are based on linear models, while Sleuth applies an error linear model. These tools collectively represent the major statistical algorithms currently used for differential expression analysis. Ballgown is compatible with StringTie-derived quantification, whereas Sleuth is compatible with quantification from Salmon, Kallisto, and Sailfish. For differential expression analysis tools, we also applied the available analysis modes. Specifically, for edgeR, three versions were considered: v1, based on the quantile-adjusted conditional maximum likelihood (qCML) method; v2, using generalized linear models (GLM) for parameter estimation; and v3/v4, which adopt the quasi-likelihood (QL) generalization of the negative binomial (NB) model. These complementary strategies were included to comprehensively evaluate the impact of different statistical tools on the detection of differentially expressed isoforms. For limma, we used the voom and trend modes, in which the mean–variance relationship for count data was modeled either with precision weights or with an empirical Bayes prior trend. For NOISeq, we applied both the real and BIO modes, with the latter optimized for biological replicates.

AS event detection tools. Four AS event detection tools were employed, including SUPPA2 (v2.4), rMATS (v4.3.0)⁶⁴, MAJIQ (v2.5.6)⁶⁵, and PSI-Sigma (v2.3)⁶⁶. SUPPA2 was run using transcript-level quantification data (TPM) generated from 18 isoform quantification pipelines. For the standardized benchmarking of the four AS event detection tools, SUPPA2 was supplied with TPM input generated by the STAR–Salmon, as this combination had been demonstrated to yield the highest accuracy. rMATS was run with cleaned reads as input, and performed alignment using its built-in

STAR module prior to AS event detection. Both MAJIQ and PSI-Sigma were executed based on STAR-aligned data. Following the rMATS manual, DSEs were defined using $|\text{IncLevelDifference}| \geq 0.05$ and $\text{FDR} \leq 0.01$. For PSI-Sigma, DSEs were identified with $|\Delta\text{PSI} (\%)| \geq 10$ and $\text{FDR (BH)} < 0.01$. MAJIQ employed a probability changing threshold greater than 0.95, whereas SUPPA2 used $|\text{dPSI}| \geq 0.05$ and $p\text{-value} \leq 0.05$.

Principal variance component analysis

Principal variance component analysis (PVCA) was performed to estimate the sources of variation in isoform (FPKM) or AS event (PSI) expression profiles across laboratories using the pvca package (v1.40.0)⁶⁷. We examined a total of 11 experimental factors, including RNA input, mRNA enrichment method, strandedness, library kit, read length, library concentration, sequencing platform, flowcell, lane, the number of reads in the exonic region, and insert size. Sample grouping also served as a factor in PVCA, allowing for a comparison between biological differences among samples and variations introduced by various factors.

Annotation of features for junction, isoform and AS event

Features were annotated for each junction, isoform, and AS event, including mappability, GC content, the number of isoforms per gene, isoform expression, isoform length, the number of exons per isoform, exon length, the K-value, the number of adjacent AS events, and PSI level. Detailed definitions of these features are provided in the Supplementary Methods.

Standardized benchmarking tool

The outputs of four alternative splicing detection tools, SUPPA2, rMATS, MAJIQ, and PSI-Sigma, were converted into a unified format. Definitions of unified format for the seven AS event types are provided in the Supplementary Methods. Using standardized formats, the four tools were evaluated for both qualitative results—including overlaps in detected AS events and DSEs—and quantitative consistency in dPSI measurements.

Statistical analysis

All statistical analyses were performed using R statistical packages (v4.5.1) and python (v3.10.10). PCA was conducted with the univariate scaling, using the prcomp (v3.6.2) function. Principal variance component analysis (PVCA) was performed by pvca

package (v1.40.0)⁶⁸ to quantifies the proportion of variance explained by each influencing factor.

Data availability

The high-quality raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2021) in National Genomics Data Center (Nucleic Acids Res 2022), China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences under accession code of GSA-Human: HRA005937 [<https://ngdc.cncb.ac.cn/gsa-human/browse/HRA005937>]⁶⁹. The constructed junction reference datasets, and expression profiles at isoform and event level, can be accessed at figshare (<https://figshare.com/account/home#/projects/264799>).

Code availability

Custom scripts for standardized benchmarking of AS event detection tools were provided in <https://github.com/wangduo-ux/Standardized-benchmarking-of-alternative-splicing-detection.git>.

Reference

1. Wright, C.J., Smith, C.W.J. & Jiggins, C.D. Alternative splicing as a source of phenotypic diversity. *Nat Rev Genet* **23**, 697-710 (2022).
2. Tao, Y., Zhang, Q., Wang, H., Yang, X. & Mu, H. Alternative splicing and related RNA binding proteins in human health and disease. *Signal Transduct Target Ther* **9**, 26 (2024).
3. Kahles, A. et al. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* **34**, 211-224.e216 (2018).
4. Nikom, D. & Zheng, S. Alternative splicing in neurodegenerative disease and the promise of RNA therapies. *Nat Rev Neurosci* **24**, 457-473 (2023).
5. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13 (2016).
6. Moya-Molina, M. et al. Deciphering the impact of PROM1 alternative splicing on human photoreceptor development and maturation. *Cell Death Dis* **15**, 721 (2024).
7. Chen, S. et al. CLK1/SRSF5 pathway induces aberrant exon skipping of METTL14 and Cyclin L2 and promotes growth and metastasis of pancreatic cancer. *J Hematol Oncol* **14**, 60 (2021).
8. Leclair, N.K. et al. RNA splicing as a biomarker and phenotypic driver of meningioma DNA-methylation groups. *Neuro Oncol* **26**, 2222-2236 (2024).
9. Sciarillo, R. et al. The role of alternative splicing in cancer: From oncogenesis to drug resistance. *Drug Resist Updat* **53**, 100728 (2020).
10. Glinos, D.A. et al. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**, 353-359 (2022).
11. Monzó, C., Liu, T. & Conesa, A. Transcriptomics in the era of long-read sequencing. *Nat Rev Genet* **26**, 681-701 (2025).
12. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* **32**, 903-914 (2014).
13. Li, S. et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol* **32**, 915-925 (2014).
14. Sahraeian, S.M.E. et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat Commun* **8**, 59 (2017).
15. Fenn, A. et al. Alternative splicing analysis benchmark with DICAST. *NAR Genom Bioinform* **5**, lqad044 (2023).
16. Dapas, M., Kandpal, M., Bi, Y. & Davuluri, R.V. Comparative evaluation of isoform-level gene expression estimation algorithms for RNA-seq and exon-array platforms. *Brief Bioinform* **18**, 260-269 (2017).
17. Kanitz, A. et al. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol* **16**, 150 (2015).
18. Mehmood, A. et al. Systematic evaluation of differential splicing tools for RNA-seq studies. *Brief Bioinform* **21**, 2052-2065 (2020).
19. Muller, I.B. et al. Computational comparison of common event-based differential splicing tools: practical considerations for laboratory researchers. *BMC Bioinformatics* **22**, 347 (2021).
20. 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).
21. Srivastava, A. et al. Alignment and mapping methodology influence transcript abundance estimation. *Genome Biol* **21**, 239 (2020).

835 22. Zhang, C., Zhang, B., Lin, L.L. & Zhao, S. Evaluation and comparison of computational tools
836 for RNA-seq isoform quantification. *BMC Genomics* **18**, 583 (2017).

837 23. Merino, G.A., Conesa, A. & Fernández, E.A. A benchmarking of workflows for detecting
838 differential splicing and differential expression at isoform level in human RNA-seq studies.
839 *Brief Bioinform* **20**, 471-481 (2019).

840 24. Jiang, M., Zhang, S., Yin, H., Zhuo, Z. & Meng, G. A comprehensive benchmarking of
841 differential splicing tools for RNA-seq analysis at the event level. *Brief Bioinform* **24** (2023).

842 25. Teng, M. et al. A benchmark for RNA-seq quantification pipelines. *Genome Biol* **17**, 74 (2016).

843 26. Sarantopoulou, D. et al. Comparative evaluation of full-length isoform quantification from
844 RNA-Seq. *BMC Bioinformatics* **22**, 266 (2021).

845 27. Mangul, S. et al. Systematic benchmarking of omics computational tools. *Nat Commun* **10**, 1393
846 (2019).

847 28. Wang, D. et al. A real-world multi-center RNA-seq benchmarking study using the Quartet and
848 MAQC reference materials. *Nat Commun* **15**, 6167 (2024).

849 29. Yu, Y. et al. Quartet RNA reference materials improve the quality of transcriptomic data through
850 ratio-based profiling. *Nat Biotechnol* **42**, 1118-1132 (2024).

851 30. Chen, Q. et al. A ratio-based framework using Quartet reference materials for integrating long-
852 and short-read RNA-seq. *bioRxiv*, 2025.2009.2015.676287 (2025).

853 31. Holik, A.Z. et al. RNA-seq mixology: designing realistic control experiments to compare
854 protocols and analysis methods. *Nucleic Acids Res* **45**, e30 (2017).

855 32. Dong, X. et al. Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures.
856 *Nat Methods* **20**, 1810-1821 (2023).

857 33. Trincado, J.L. et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis
858 across multiple conditions. *Genome Biol* **19**, 40 (2018).

859 34. Kumar, A. et al. The impact of RNA sequence library construction protocols on transcriptomic
860 profiling of leukemia. *BMC Genomics* **18**, 629 (2017).

861 35. Schuierer, S. et al. A comprehensive assessment of RNA-seq protocols for degraded and low-
862 quantity samples. *BMC Genomics* **18**, 442 (2017).

863 36. Liu, Y. et al. Evaluating the impact of sequencing depth on transcriptome profiling in human
864 adipose. *PLoS One* **8**, e66883 (2013).

865 37. Liu, X. et al. A comparison of transcriptome analysis methods with reference genome. *BMC*
866 *Genomics* **23**, 232 (2022).

867 38. McDermaid, A., Monier, B., Zhao, J., Liu, B. & Ma, Q. Interpretation of differential gene
868 expression results of RNA-seq data: review and integration. *Brief Bioinform* **20**, 2044-2054
869 (2019).

870 39. Rigai, G. et al. Synthetic data sets for the identification of key ingredients for RNA-seq
871 differential analysis. *Briefings in Bioinformatics* **19**, 65-76 (2016).

872 40. Aquino, J. et al. A novel splicing graph allows a direct comparison between exon-based and
873 splice junction-based approaches to alternative splicing detection. *Briefings in Bioinformatics*
874 **26** (2025).

875 41. Su, Y. et al. Comprehensive assessment of mRNA isoform detection methods for long-read
876 sequencing data. *Nat Commun* **15**, 3972 (2024).

877 42. Chen, Y. et al. A systematic benchmark of Nanopore long-read RNA sequencing for transcript-
878 level analysis in human cell lines. *Nature Methods* **22**, 801-812 (2025).

879 43. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
880 *Bioinformatics* **34**, i884-i890 (2018).

881 44. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

882 45. Pertea, M., Kim, D., Pertea, G.M., Leek, J.T. & Salzberg, S.L. Transcript-level expression
883 analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* **11**, 1650-
884 1667 (2016).

885 46. Chen, Y., Chen, L., Lun, Aaron T.L., Baldoni, Pedro L. & Smyth, Gordon K. edgeR v4:
886 powerful differential analysis of sequencing data with expanded functionality and improved
887 support for small counts and larger datasets. *Nucleic Acids Research* **53** (2025).

888 47. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment
889 and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915 (2019).

890 48. Liao, Y., Smyth, G.K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping
891 by seed- and-vote. *Nucleic Acids Res* **41**, e108 (2013).

892 49. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment
893 of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

894 50. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for
895 assigning sequencing reads to genomic features. *Bioinformatics (Oxford, England)* **30**, 923-930
896 (2014).

897 51. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq
898 reads. *Nat Biotechnol* **33**, 290-295 (2015).

899 52. Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq.
900 *Nat Biotechnol* **31**, 46-53 (2013).

901 53. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or wit
902 hout a reference genome. *BMC bioinformatics* **12**, 323 (2011).

903 54. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing
904 experiments. *Nat Methods* **10**, 71-73 (2013).

905 55. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast and bias-
906 aware quantification of transcript expression. *Nature methods* **14**, 417-419 (2017).

907 56. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
908 quantification. *Nat Biotechnol* **34**, 525-527 (2016).

909 57. Patro, R., Mount, S.M. & Kingsford, C. Sailfish enables alignment-free isoform quantification
910 from RNA-seq reads using lightweight algorithms. *Nat Biotechnol* **32**, 462-464 (2014).

911 58. Fu J, F.A., Collado-Torres L, Jaffe AE, Leek JT (Bioconductor Bioconductor 2025).

912 59. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential
913 expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**, 139-
914 140 (2010).

915 60. Ritchie, M.E. et al. limma powers differential expression analyses for RNA-sequencing and m
916 icroarray studies. *Nucleic Acids Res* **43**, e47 (2015).

917 61. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for
918 RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).

919 62. Tarazona, S. et al. Data quality aware analysis of differential expression in RNA-seq with
920 NOISeq R/Bioc package. *Nucleic Acids Res* **43**, e140 (2015).

921 63. Pimentel, H., Bray, N.L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq
922 incorporating quantification uncertainty. *Nat Methods* **14**, 687-690 (2017).

923 64. Wang, Y. et al. rMATS-turbo: an efficient and flexible computational tool for alternative splicing
924 analysis of large-scale RNA-seq data. *Nat Protoc* **19**, 1083-1104 (2024).
925 65. Vaquero-Garcia, J. et al. RNA splicing analysis using heterogeneous and large RNA-seq datasets.
926 *Nat Commun* **14**, 1230 (2023).
927 66. Lin, K.T. & Krainer, A.R. PSI-Sigma: a comprehensive splicing-detection method for short-
928 read and long-read RNA-seq analysis. *Bioinformatics* **35**, 5048-5054 (2019).
929 67. Bushel, P. pvca: Principal Variance Component Analysis (PVCA). doi:10.18129/B9.bioc.pvca,
930 R package version 1.42.0, <https://bioconductor.org/packages/pvca>. (2023).
931 68. Li, J., Bushel, P.R., Chu, T.-M. & Wolfinger, R.D. in Batch Effects and Noise in Microarray
932 Experiments 141-154 (2009).
933 69. Wang, D. et al. (Genome Sequence Archive (GSA), Genome Sequence Archive (GSA); 2023).
934

Acknowledgements

This study was supported in part by the Beijing Natural Science Foundation (L254025), National Key R&D Program of China (2023YFC3402500/2023YFC3402503), and Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0533100/2024ZD0533102). We thank the 42 laboratories for performing RNA-seq and for returning the raw data on time.

Author Contributions

R.Z., J.L., Y.T.Z., and L.S. conceived the study, provided overall supervision, and revised the manuscript. Y.T.Z., and L.S. provided the RNA reference materials and long-read sequencing data. D.W. performed the experiments, conducted data analysis, generated the figures, and wrote the original manuscript. J.Z. conducted data analysis and performed the RT-qPCR experiments. Q.C. generated the Quartet reference datasets and helpful discussions. Y.F.Z. managed the RNA-seq data. Y.H., Y.L., C.L., H.W., and Y.Y. participated in the study design, provided constructive discussions and invaluable advice for this study. All authors reviewed and approved the manuscripts.

Competing Interests

The authors declare no competing interests.

Figure legends

Fig.1 | Overview of study design.

a, Certified Quartet RNA reference materials from immortalized cell lines of a Chinese Quartet with father (F7), mother (M8), and monozygotic twin daughters (D5 and D6) were used. Samples M8 and D6 were mixed at 3:1 and 1:3 ratios to create samples T1 and T2, respectively. Each sample was prepared with three replicates. The reference samples were tested in 42 laboratories with distinct protocols and analyzed with 159 bioinformatics pipelines. We systematically evaluated RNA-seq accuracy in isoform quantification, AS event quantification, and differential expression and splicing analyses. To identify best practices, we investigated influencing factors in experimental and bioinformatics processes as well as intrinsic isoform and event features. In addition, we developed a standardized benchmarking tool to harmonize event formats and facilitate the evaluation of four AS event detection methods. **b**, Overview of sequencing platforms, library preparation methods, and bioinformatics pipelines (Fig.S2). * indicates that differential expression analysis tools are applied only to their compatible input: Cuffdiff for Cuffdiff results, Ballgown for StringTie results, and Sleuth for data from Salmon, Kallisto, and Sailfish. Circles and triangles in different colors represent different performance metrics and types of ground truth, respectively. RMSE, root mean square error; PCC, Pearson correlation coefficient; RMR, recovery of mixing ratios; MCC, Matthews correlation coefficient. Created in BioRender. Zhang, R. (2025) <https://BioRender.com/3rpzq7f>.

Fig.2 | Evaluation of isoform detection performance of RNA-seq.

a–b, Principal component analysis (PCA) plots on isoform-level expression profiles from (a) all 42 laboratories and (b) the 19 high-quality laboratories. **c**, Assessment of relative isoform expression quantification using root mean square error (RMSE) and Pearson correlation coefficient (PCC) based on three types of ground truth. Mixing ratio-based ground truth was employed to evaluate the consistency between measured and theoretical isoform expression levels in mixed samples T1 and T2, quantified using RMSE. **d**, Comparison of RMSE values 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. ** indicates a two-sided p -value < 0.01 . **e–f**, Correlation of RMSE with (e) sequencing depth in exonic regions and (f) false-negative rate of junction detection across laboratories. **g–h**, Recall, precision and Matthews correlation coefficient (MCC) of differentially expressed isoforms (DEIs) detection across laboratories based on the (f) Quartet and (g) RT-qPCR reference datasets. **i**, Correlation between recall based on the Quartet reference datasets and sequencing depth in exonic regions across laboratories. **j**, Correlation between MCC and SNR across laboratories. **k**, Average read counts and inter-replicate coefficient of variation (CV) for isoforms identified as false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN) across all laboratories. R represents the Pearson correlation coefficient.

Fig.3 | Technical and biological factors affecting isoform detection performance.

a, Principal variance component analysis quantifies the proportion of variance explained by each experimental factor for the Quartet and mixed samples. RNA-seq data with high-quality flags were processed using a unified STAR + StringTie pipeline to exclusively investigate the sources of variation from the experimental processes. **b**, Clustering of 19 high-quality laboratories based on the root-mean-square error (RMSE) values. The horizontal clustering represents the combinations of three ground truths and 18 isoform quantification pipelines. One laboratory with abnormal RSEM run was excluded. **c**, Comparison of 18 isoform quantification pipelines based on RMSE values calculated against three types of ground truth. **d**, The distribution of coefficients of variation (CV) of expression levels among isoforms belonging to the same gene, as estimated by all quantification pipelines. Data are shown for the M8 sample from lab01 as an example. **e**, Clustering of the 18 quantification pipelines based on RMSE values. **f**, Evaluation of the 11 differentially expressed isoforms (DEIs) detection methods based on the Matthews correlation coefficient (MCC). The colors of the points represent different quantification pipelines, consistent with panel c. **g**, Evaluation of each DEI detection tool with respect to different isoform quantification pipelines. **h**, Evaluation

of quantification pipelines with respect to multiple isoform features, including the isoform length, the number of isoforms per gene, the mean length of exons per isoform, the number of exons per isoform, GC content, K-value, and expression level (FPKM). **i**, Evaluation of the extent to which isoform features affect the performance of 18 isoform quantification pipelines. RMSE values based on the Quartet reference datasets (left panel) and built-in truth (right panel) from 19 high-quality datasets were grouped by each feature and analyzed using one-way ANOVA to derive *p-value* and effect sizes (η^2). Data in boxplots are presented as median values (center lines) and the upper and lower quartiles (box limits). The tools are abbreviated as follows: STA (STAR), HIS (HISAT2), Bow (Bowtie2), Cuff (Cuffdiff), Stri (StringTie), feat (featureCounts), RSEM (RS), eXpress (eXpr), Sal (Salmon), Sailfish (Sail), and Kall (Kallisto).

Fig.4 | Evaluation of AS event detection performance of RNA-seq.

a, Principal component analysis (PCA) plots on alternative splicing (AS) event-level quantification data (PSI) from the 19 high-quality laboratories. **b**, Evaluation of AS event relative quantification (delta PSI) using root mean square error (RMSE) and Pearson correlation coefficient (PCC) based on the Quartet and RT-qPCR reference datasets. **c**, Comparison of PCC values between high-quality and low-quality laboratories. Significance testing among groups was conducted using Mann-Whitney U test. * and ** indicates a two-sided *p-value* < 0.05 and *p-value* < 0.01. **d–e**, Correlation of RMSE with (d) sequencing depth in exonic regions and (e) false-negative rate of junction detection across laboratories. **f–g**, Recall, precision and Matthews correlation coefficient (MCC) of differential splicing events (DSEs) detection across laboratories based on the (f) Quartet and (g) RT-qPCR reference datasets. **h**, Correlation between recall and sequencing depth in exonic regions across laboratories. Data in boxplots are presented as median values (center lines) and the upper and lower quartiles (box limits). *R* represents the PCC.

Fig.5 | Technical and biological factors affecting AS event detection performance.

a, Principal variance component analysis quantifies the proportion of variance in

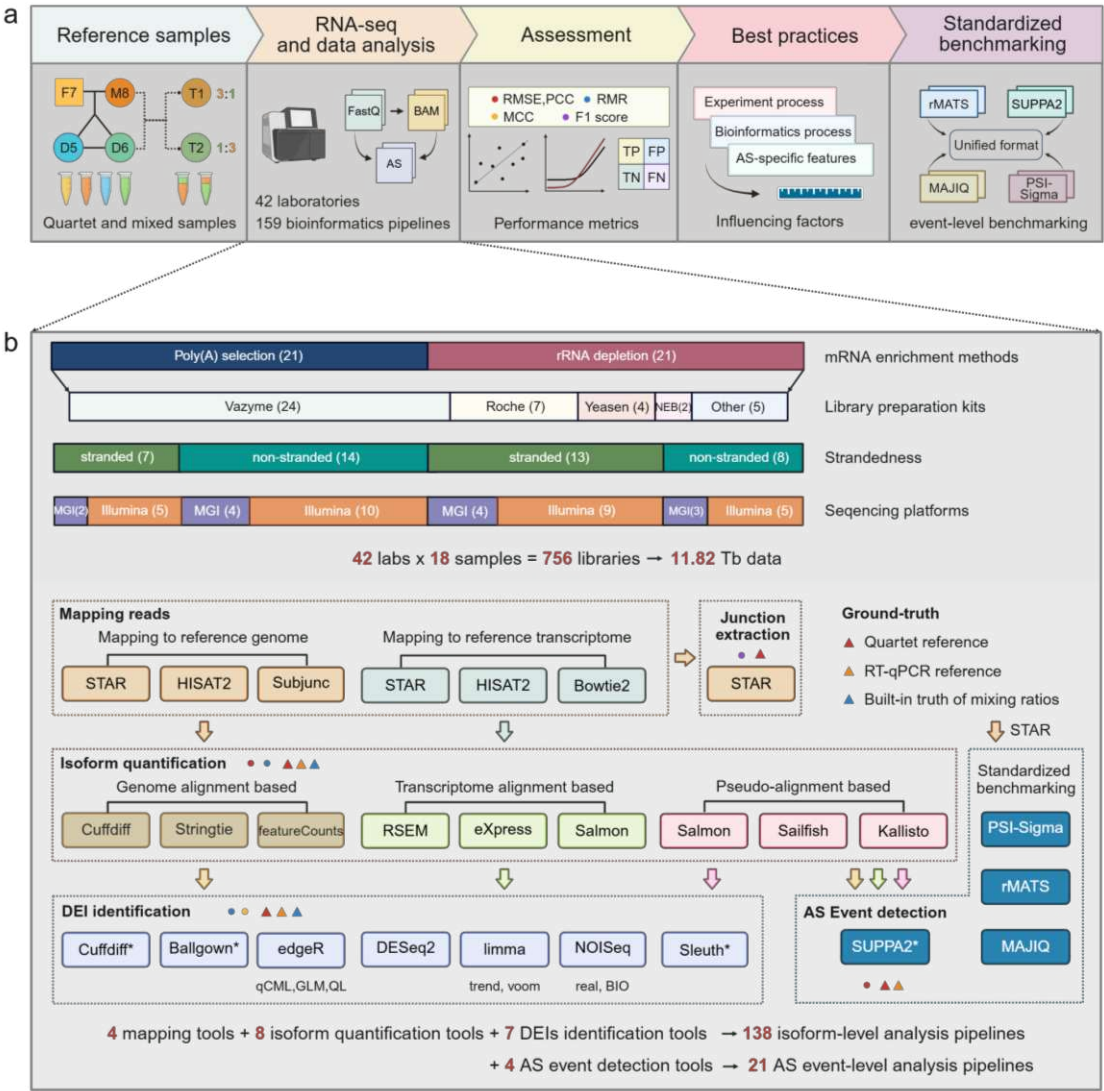
percent spliced-in (PSI) data explained by each experimental factor for the Quartet and mixed samples. RNA-seq data with high-quality flags were processed using a unified STAR + StringTie + SUPPA2 pipeline to exclusively investigate the sources of variation from the experimental processes. **b**, Clustering of 19 high-quality laboratories based on the root-mean-square error (RMSE) values. Clustering of 19 high-quality laboratories based on the root-mean-square error (RMSE) values. The horizontal clustering represents the combinations of two reference datasets and 18 SUPPA2-based analysis pipelines. One laboratory with abnormal RSEM run was excluded. **c**, Comparison of delta PSI (dPSI) quantification accuracy of SUPPA2 combined with 18 isoform quantification pipelines using normalized RMSE (nRMSE) and Pearson correlation coefficient (PCC). **d**, Distribution of dPSI values from SUPPA2 combined with 18 isoform quantification pipelines. Data are shown for lab01 as an example. **e**, Differential splicing events (DSEs) detection accuracy of SUPPA2 across 18 isoform quantification pipelines based on the Quartet and RT-qPCR reference datasets. **f**, Evaluation of the extent to which AS event features affect the performance of 18 pipelines. RMSE values from 19 high-quality datasets were grouped by each feature and analyzed using one-way ANOVA to derive *p-value* and effect sizes (η^2). **g**, Evaluation of SUPPA2 combined with 18 isoform quantification pipelines with respect to multiple AS event features, including number of adjacent AS events, AS-related isoform number, GC content, mappability, and PSI level, and AS-related isoform expression (FPKM). Data in boxplots are presented as median values (center lines) and the upper and lower quartiles (box limits). The tools are abbreviated as follows: STA (STAR), HIS (HISAT2), Bow (Bowtie2), Cuff (Cuffdiff), Stri (StringTie), feat (featureCounts), RSEM (RS), eXpress (eXpr), Sal (Salmon), Sailfish (Sail), Kall (Kallisto), and SUP (SUPPA2).

Fig.6 | Standardized benchmarking of four AS event detection tools

a–b, The UpSet plots show the intersection of (a) alternative splicing (AS) events and (b) differentially splicing events (DSEs) from four AS event detection tools, using data from lab1 as an example. AS events with percent spliced-in (PSI) values between 0.05

1074 and 0.95 were included. **c.** Evaluation of the consistency of delta PSI (dPSI)
1075 quantification across the four AS event detection tools for each event type. The data in
1076 the boxplots represent results from 19 high-quality laboratories across three sample
1077 pairs, presented as median values (center lines) with upper and lower quartiles (box
1078 limits). **d.** The consistency of dPSI quantification results between the four AS event
1079 detection tools and the Quartet and RT-qPCR reference datasets. **e.** The recall and
1080 precision of DSE detection against the Quartet and RT-qPCR reference datasets by four
1081 AS event detection tools.

1082 **Figures**



1083

1084 **Fig.1 | Overview of study design.**

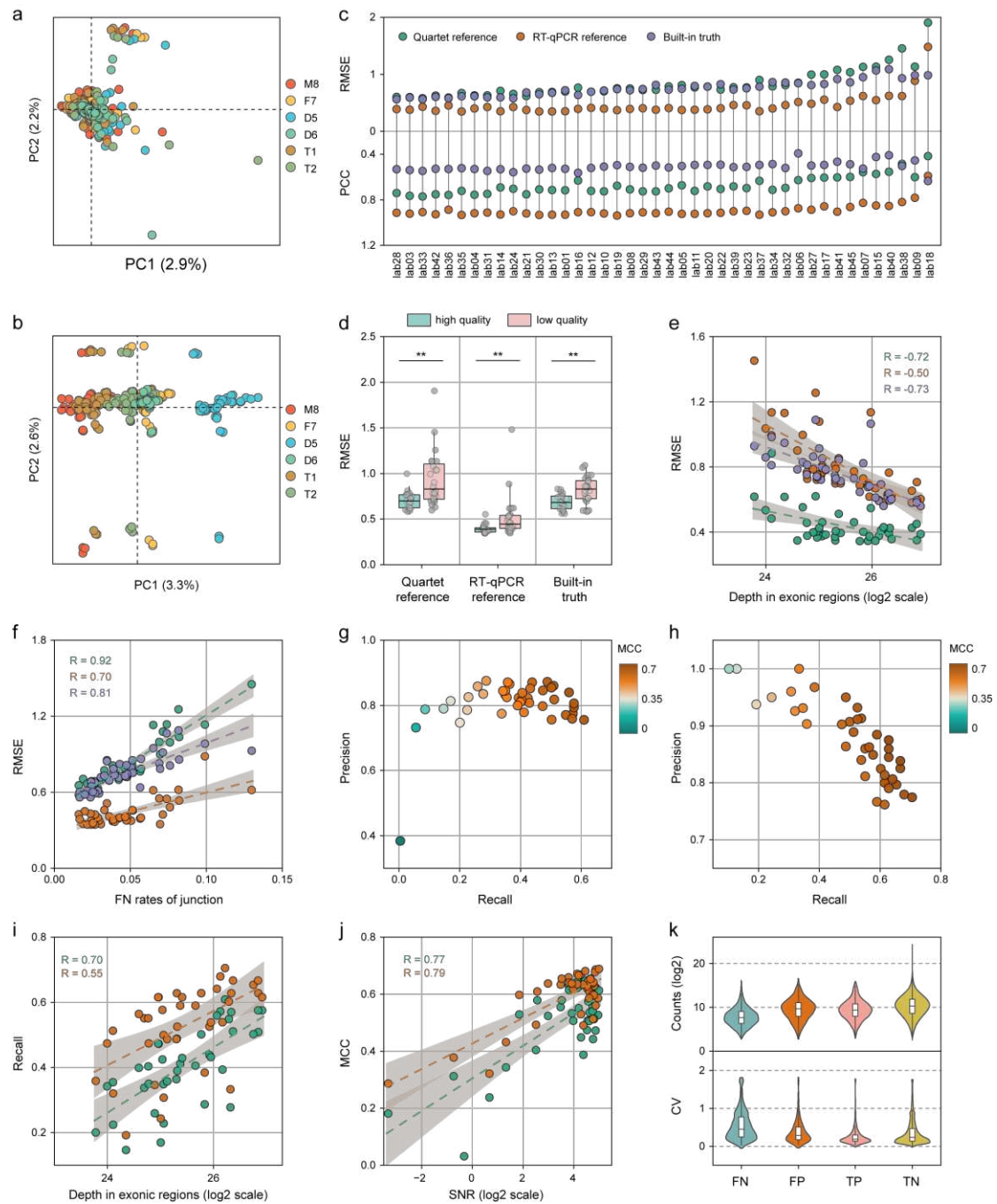


Fig.2 | Evaluation of isoform detection performance of RNA-seq.

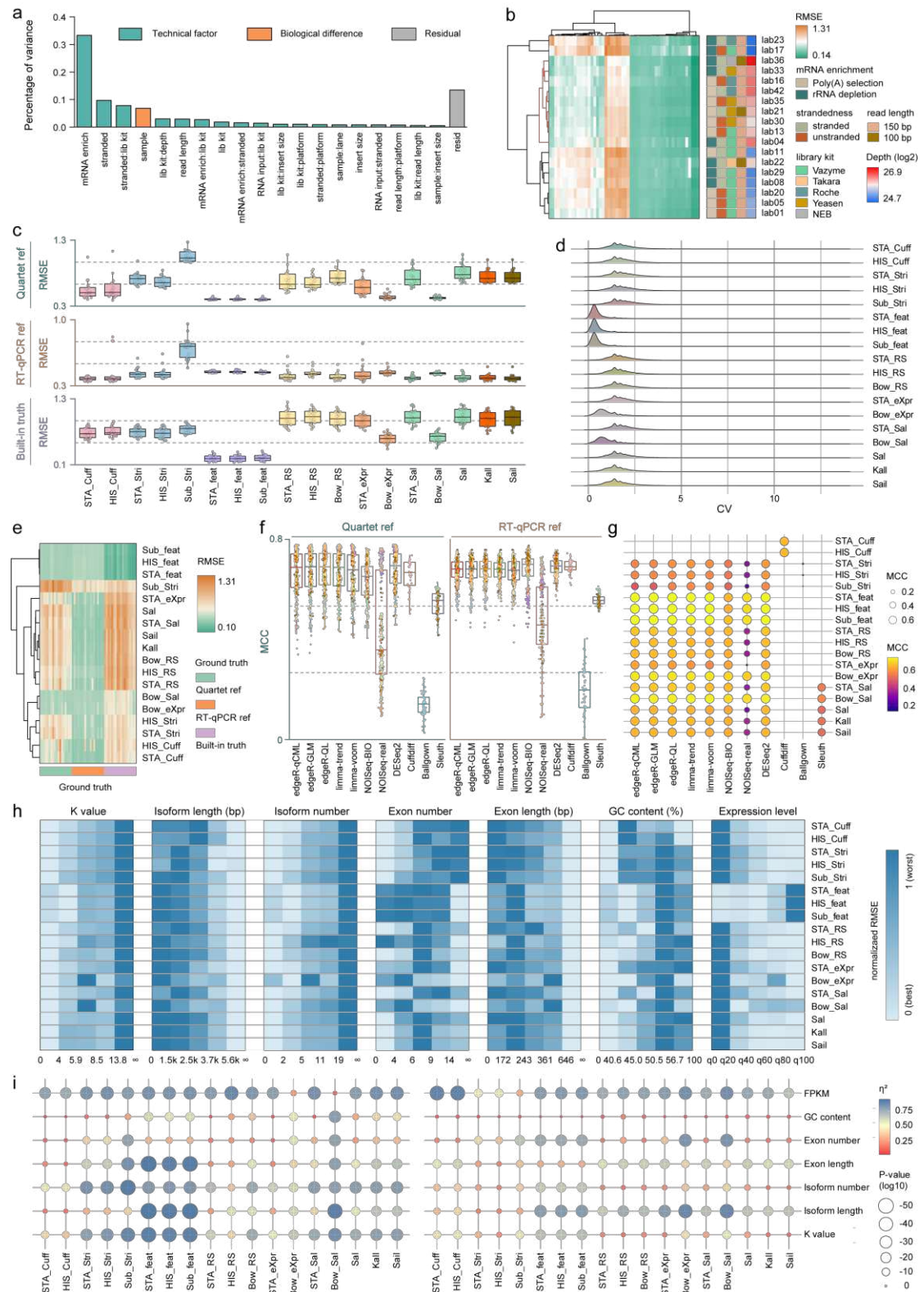


Fig.3 | Technical and biological factors affecting isoform detection performance.

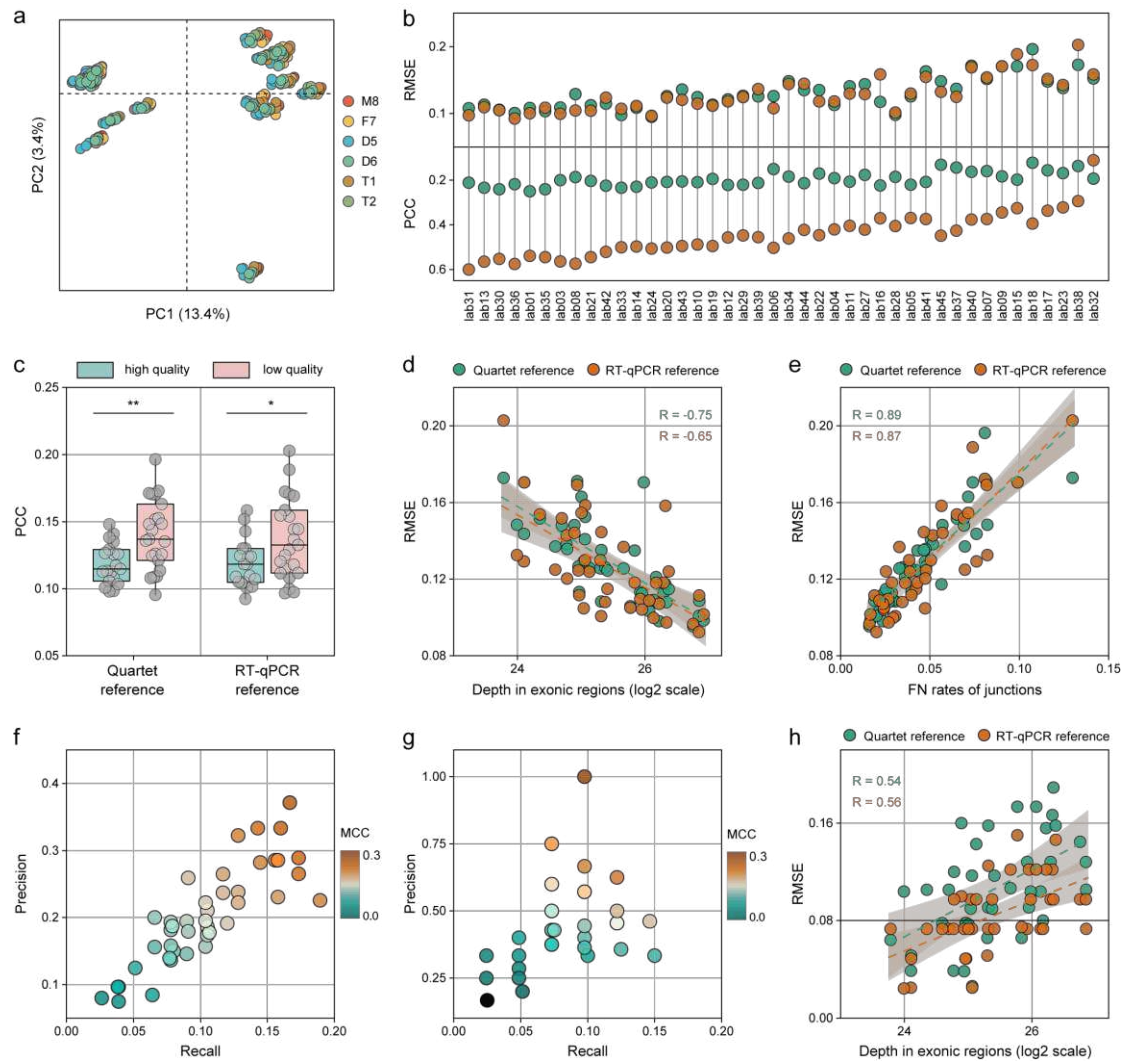


Fig.4 | Evaluation of AS event detection performance of RNA-seq.

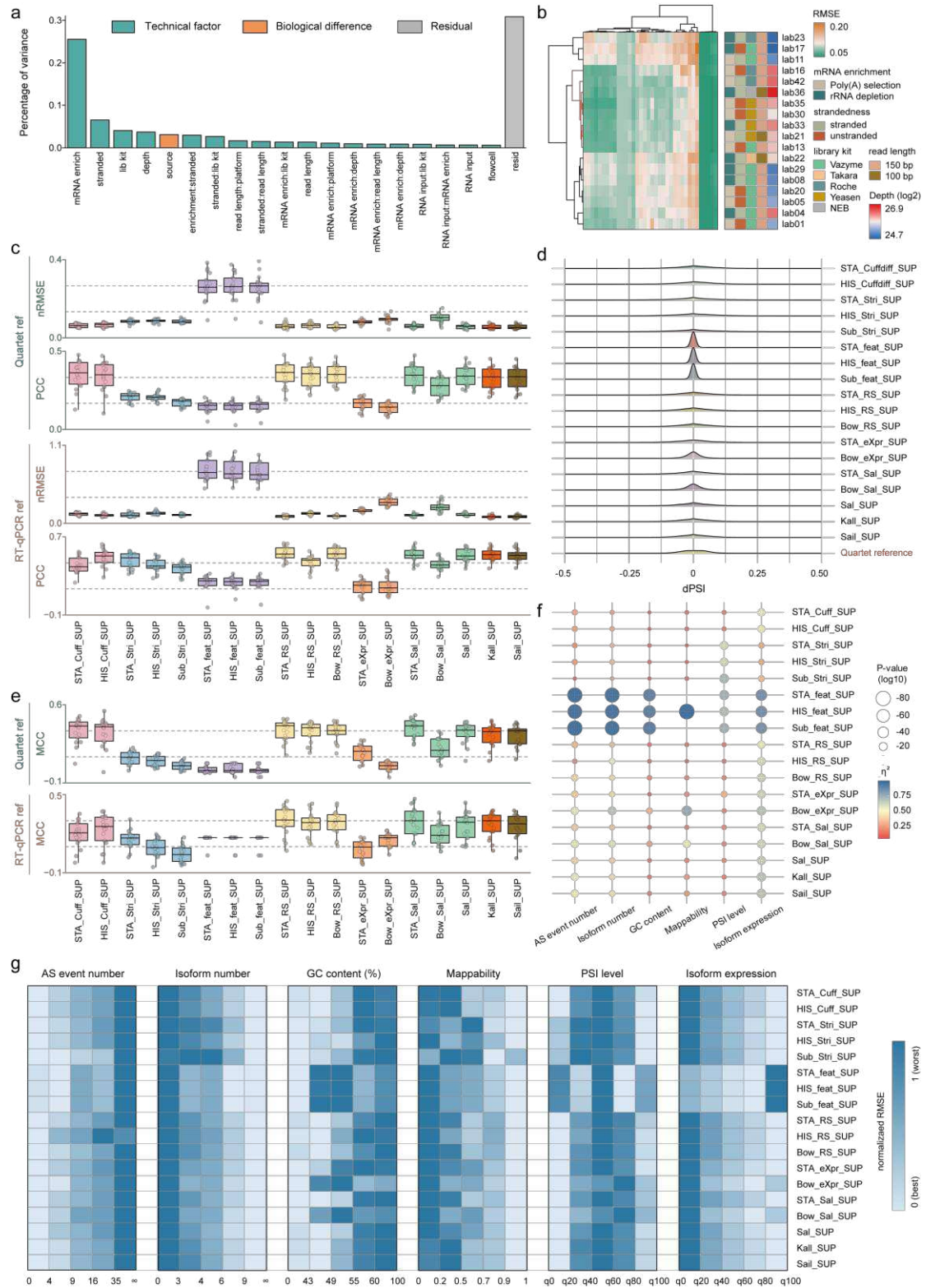


Fig.5 | Technical and biological factors affecting AS event detection performance.

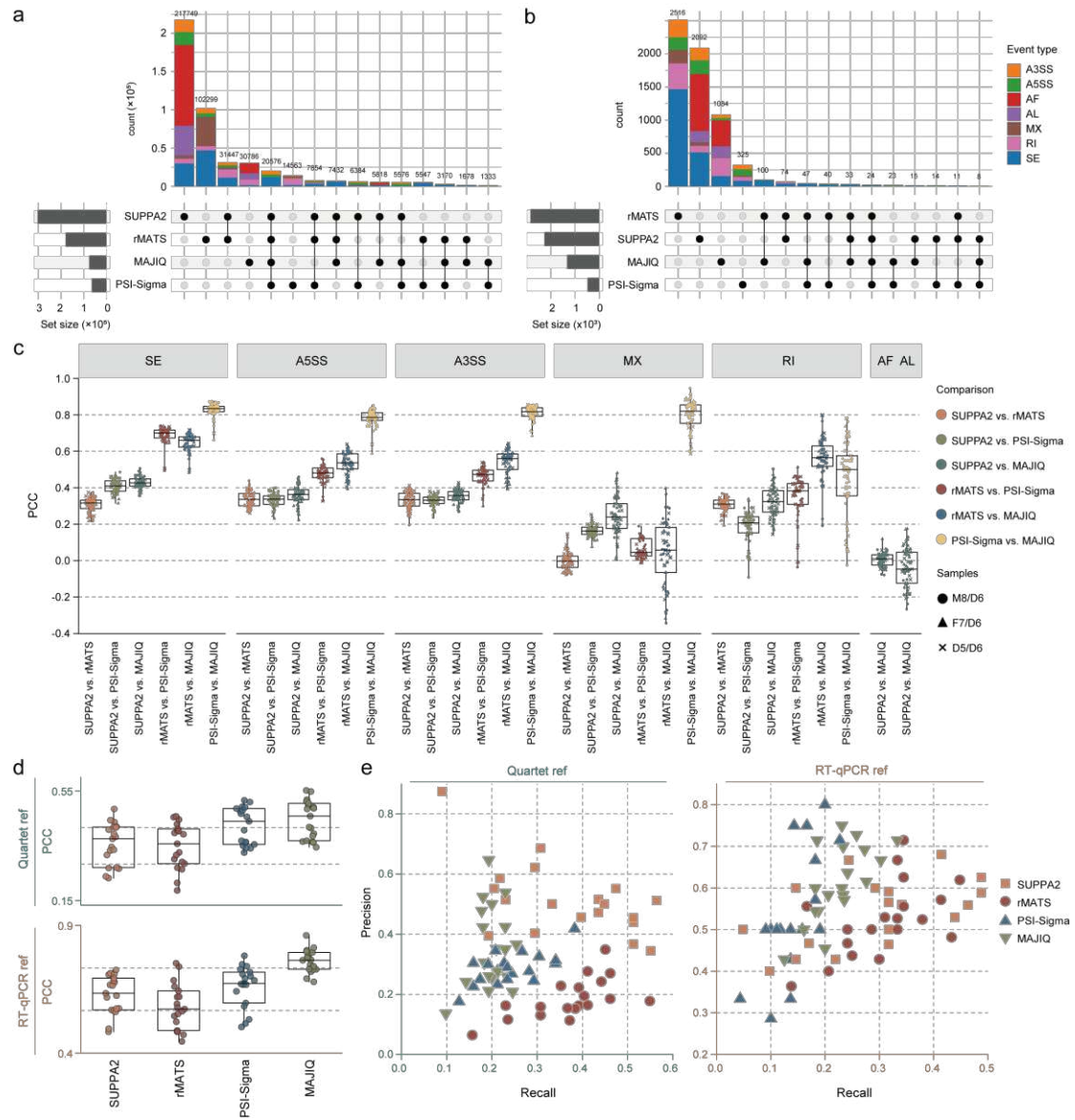


Fig.6 | Standardized benchmarking of four AS event detection tools.

Tables

Table 1 Best-practice recommendations for RNA application.

RNA-seq processes	Recommendations	Performance gap* (optimal vs. poorest)
Quality control	Rigorous quality control	
	- To reduce the impact of low-quality experimental operations - Recommended quality metrics: SNR, duplicate rate, 5'–3' bias, number of exonic reads, proportion of exonic reads, cross-contamination, and mapping rate	SNR gap (high- vs. low-quality): 25 vs. 4.8 and 14.8 vs. 7.1 at isoform and AS event level, respectively
Isoform-level analysis	Best practices for experiment design	PCC gap: 0.15 (0.02–0.18) ^Q , 0.05 (0.02 – 0.21) ^R across quantification pipelines
	Best practices for bioinformatics design	PCC gap: 0.26 (0.21–0.31) ^Q , 0.15 (0.07 – 0.22) ^R across laboratories
	- Isoform quantification: STAR/HISAT2+Cuffdiff; STAR+eXpress - DEI detection: DESeq2, edgeR, limma, and NOISeq-BIO in combination with RSEM, Salmon, Sailfish, Kallisto, or eXpress; Cuffdiff	MCC gap: 0.54 (0.46–0.61) ^Q , 0.49 (0.34 – 0.59) ^R across laboratories
	Influencing factors associated with isoform features	
	- Factors positively associated with accuracy: Isoform length, isoform expression level - Factors negatively associated with accuracy: K-values, isoform number per gene, exon length, GC content - Other: Exon number per isoform: Cuffdiff and StringTie–high associated with lower accuracy; others–intermediate associated with lower accuracy	NA

RNA-seq processes	Recommendations	Performance gap* (optimal vs. poorest)
AS event- level analysis	Best practices for experiment design	PCC gap: 0.16 (0.04–0.25) ^Q ,
	The Poly(A) selection method and higher sequencing depth	0.19 (0.06 – 0.27) ^R across quantification pipelines
	Best practices for bioinformatics design using SUPPA2	PCC gap: 0.18 (0.09–0.28) ^Q ,
	• AS event quantification and DSE detection: SUPPA2 in combination with Salmon, RSEM, Cuffdiff, Sailfish, or Kallisto	0.15 (0.02–0.29) ^R across laboratories
		MCC gap: 0.31 (0.15–0.44) ^Q , 0.24 (0.03–0.45) ^R across laboratories
	Influencing factors associated with isoform features	
	• Factors positively associated with accuracy: AS event-related isoform number, mappability, AS event- related isoform expression level	NA
	• Factors negatively associated with accuracy: GC content, the number of adjacent AS events	
	Best practices for AS event detection tools	PCC: MAJIQ, PSI-Sigma,
	• Individual AS event detection tools generally show limited performance	SUPPA2, and rMATS (0.45, 0.42, 0.36, and 0.35) ^Q and
	• MAJIQ and PSI-Sigma show higher accuracy for event quantification	(0.75, 0.65, 0.61, and 0.56) ^R
	• SUPPA2 show higher accuracy for DSE identification	MCC: SUPPA2, rMATS, MAJIQ, and PSI-Sigma (0.41, 0.25, 0.25, and 0.25) ^Q and
	• Standardization of AS event detection tools is required	(0.26, 0.26, 0.25, and 0.19) ^R

* featureCounts- and Bowtie2-based pipelines, deemed unsuitable for AS analysis, were excluded from optimal–poorest pipeline comparisons. NA, not available.

^Q and ^R represent the evaluation against the Quartet and RT-qPCR reference datasets.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryData1.xlsx](#)
- [SupplementaryData3.xlsx](#)
- [SupplementaryData2.xlsx](#)
- [SupplementaryData4.xlsx](#)
- [SupplementaryData6.xlsx](#)
- [SupplementaryData5.xlsx](#)
- [Supplementaryinformation.pdf](#)