Briefings in Bioinformatics, 2026, 27, bbag267
https://doi.org/10.1093/bib/bbag267
Published: 27 May 2026
Review

OXFORD

The impact of transcriptome assembly algorithms on downstream quantification in RNA-seq data analysis

Zeming Tan[1], Jiahao Li[1], Enfeng Qi[2,*], Ting Yu[1,*]

[1]Research Center for Mathematics and Interdisciplinary Sciences, Frontiers Science Center for Nonlinear Expectations (Ministry of Education), Shandong University, 72 Binhai Road, Jimo, Qingdao, 266237, China
[2]School of Mathematics and Statistics, Guangxi Normal University, No. 15 Yucai Road, Qixing District, Guilin, Guangxi Zhuang Autonomous Region, 541004, China
*Corresponding authors. Enfeng Qi, School of Mathematics and Statistics, Guangxi Normal University, No.15 Yucai Road, Qixing District, Guilin, Guangxi Zhuang Autonomous Region, 541004, China. Email: ef521@126.com; Ting Yu, Research Center for Mathematics and Interdisciplinary Sciences, Frontiers Science Center for Nonlinear Expectations (Ministry of Education), Shandong University, 72 Binhai Road, Jimo, Qingdao, 266237, China. Email: yutingsdu@163.com

Abstract

Transcriptome assembly and quantification are crucial steps in the differential expression analysis of RNA-seq data. As transcriptome assembly precedes quantification, its results inevitably influence the outcomes of quantification. This study investigates the impact of transcriptome assembly algorithms on quantification outcomes in next-generation RNA-seq data analysis. From the perspective of quantification results, we evaluate the performance of transcriptome assembly algorithms. We assess the assembly quality and stability of three commonly used transcriptome assemblers—StringTie2, Scallop, and Cufflinks—on both simulated and real datasets. Our evaluation provides references for downstream analyses and identifies the most effective and stable pipeline, which is specifically the pipeline combining HISAT2 (for transcriptome alignment) and StringTie2 (for assembly). Furthermore, we compare simulated data generated by RNA-seq data simulation tools with real RNA-seq data and reveal that simulated data fails to fully capture the complexity of real data. Through this analysis, we identify transcript features associated with poor assembly and quantification performance, specifically highlighting two extreme cases: long, low-expression transcripts that are often overlooked and short transcripts that are prone to quantification errors. These findings offer valuable insights into future software development directions.

Keywords RNA-seq, DEG analysis, transcriptome assembly, transcriptome quantification

Introduction

Next-generation sequencing (NGS) [1], particularly RNA-seq developed in 2008, has revolutionized transcriptome profiling, enabling genome-wide quantification of gene expression and isoform discovery [2, 3]. While third-generation sequencing (TGS) offers advantages in read length, through which researchers can directly sequence of RNA molecules or get full-length RNA-seq data [4–6], NGS platforms remain dominant due to their higher throughput, lower error rates (about 1 to 2 orders of magnitude lower than TGS [7, 8]), and mature analytical workflows [9, 10]. Standard RNA-seq analysis usually involves [11]:

(i) Alignment of reads to a reference genome (e.g. using HISAT2 [12] or STAR [13]),

(ii) Transcriptome assembly (genome-guided by tools like StringTie2 [14], Scallop [15], or Cufflinks [16], or *de novo* assemblers like Trinity [17] where references are unavailable), and

(iii) Transcript quantification (e.g. via Salmon [18] or RSEM [19]) for downstream differential expression analysis (DEG).

Assembly is a critical precursor to quantification. Current assembly evaluation relies primarily on metrics like precision, recall, and F1-score, assessing reconstruction accuracy in isolation [20]. Crucially, the propagation of assembly errors into downstream quantification fidelity remains poorly characterized.

As quantification is the cornerstone of DEG analysis, understanding how assembly algorithms bias expression estimates is essential for biological interpretation. Srivastava *et al.* evaluated the transcriptome alignment algorithm using quantitative results and demonstrated that the quality of the assembly algorithm outputs indeed affects the reliability of the DEG analysis [21].

One key challenge lies in the difficulty of comparing quantification results. We noticed that those results are high-dimensional sparse vectors, so we chose Spearman correlation, Pearson's correlation, the root mean square error, and mean absolute error as an evaluating indicator.

We address this gap by introducing a downstream-centric evaluation framework, as illustrated in Fig. 1. We posit that the

Received: November 4, 2025. **Revised:** March 15, 2026. **Accepted:** May 4, 2026
© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

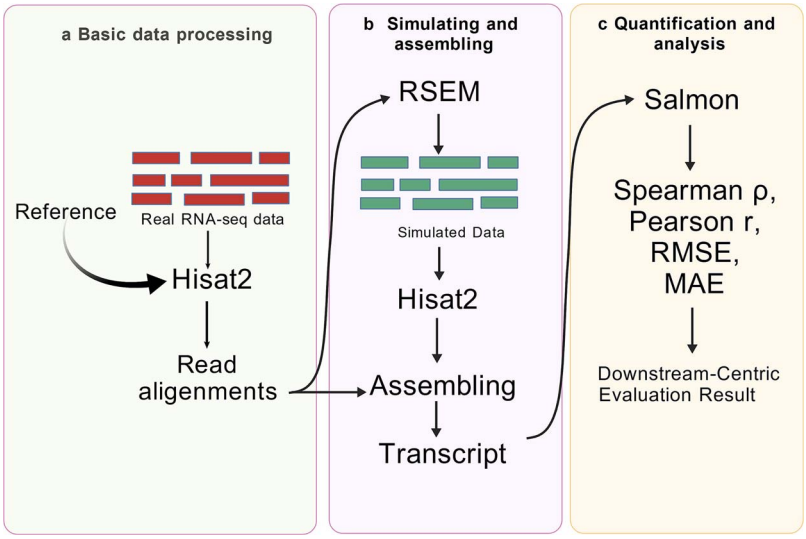

Figure 1 Pipeline. The overall workflow of the RNA-seq data analysis framework used in this study. Simulated RNA-seq data were generated using RSEM, while real RNA-seq datasets were aligned to the reference genome using HISAT2. Transcriptome assembly was performed using StringTie2, Scallop, and Cufflinks, followed by transcript quantification using Salmon.

performance of assembly algorithms should be judged by its impact on the accuracy and robustness of subsequent expression quantification. Specifically, we: (i) quantify how assembly outputs from StringTie2, Scallop, and Cufflinks impact transcript-level quantification accuracy (versus ground truth) using simulated data; (ii) assess pipeline performance stability using diverse real human RNA-seq datasets, contrasting it against simulated results, which may solve the problem that simulated data cannot reflect the complexity of real data [21]; (iii) identify transcript structural features associated with assembly-induced quantification errors; (iv) propose optimization strategies for both assembly algorithms and benchmarking methodologies.

This work establishes a novel paradigm for evaluating RNA-seq assemblers, moving beyond conventional reconstruction metrics to assess their functional impact on key biological inferences. The identified error-prone transcript features provide clear targets for algorithmic improvement. Our findings challenge the sufficiency of simulated data for tool validation and offer practical guidance for selecting robust analysis pipelines.

Results

Results on simulated data

Spearman correlation analysis on simulated data

Ten groups of simulated RNA-seq data for humans, five for mice, and three for *Arabidopsis* were generated. The three designated pipelines (involving StringTie2, Scallop, and Cufflinks for assembly) were executed on these simulated datasets. Initially, we calculated the average number of distinct transcripts identified by each pipeline across species, as presented in Table 1. Subsequently, we computed the Spearman correlation between the quantification results from each pipeline and the true TPM values. The average correlations between pipelines and ground truth across species are shown in Table 2, while the pairwise Spearman correlations between different pipelines are detailed in Table 3.

As reflected in Table 1, we noticed that StringTie2 always reconstructs the most transcripts, but Scallop always reconstructed the least on different species. So, the transcript assembly performance of various tools exhibits consistent species-specific trends, further confirming the substantial differences in assembly outcomes among tools.

As evidenced by Table 2, across different species, both StringTie2 and Cufflinks pipelines demonstrated substantially superior performance compared to Scallop in terms of quantification accuracy. For human data, StringTie2 and Cufflinks also slightly outperformed Salmon quasi-mapping, although the margin was narrow. When comparing the datasets in Table 1 and Table 2, we found that Scallop consistently assembled significantly fewer transcripts across both scenarios. This observation is likely attributable to Scallop's reliance solely on the alignment-generated BAM file, omitting reference genome annotation. In contrast, StringTie2 and Cufflinks utilize reference annotations, while quasi-mapping leverages the reference transcriptome.

Table 3 indicates exceptionally high Spearman correlation (>0.99) between the TPM values generated by StringTie2, Cufflinks, and quasi-mapping across all species. StringTie2 and Cufflinks always have the highest correlations on three different species and Scallop always has lower correlations with other two software. Given that the results for all three species revealed analogous trends regarding pipeline performance, human data was selected for subsequent detailed analysis to conserve computational resources and data volume.

Complementary evaluation metrics on simulated data

The inclusion of complementary evaluation metrics demonstrates that the conclusions are robust to the choice of performance measures. The results of different evaluation metrics are shown in Table 4.

According to Table 4, while Spearman's rank correlation indicates comparable ranking consistency between StringTie2 and Cufflinks, additional metrics reveal substantial differences among pipelines. Pearson correlation further confirms strong linear agreement for StringTie2 and Cufflinks.

Error-based metrics provide clearer separation between methods. StringTie2 achieves the lowest root mean square error (RMSE)

Table 1 Average number of transcripts assembled by different pipelines.

	Ground truth	StringTie2	Scallop	Cufflinks	Quasi-mapping
Humans	99 691 ± 13	62 640 ± 21	23 048 ± 19	60 709 ± 22	61 867 ± 20
Mice	66 490 ± 16	51 603 ± 23	22 999 ± 15	50 901 ± 20	–
Arabidopsis	30 797 ± 6	29 875 ± 9	21 435 ± 13	29 780 ± 10	–

Table 2 Spearman correlation between pipeline results and ground truth TPM values.

	StringTie2	Scallop	Cufflinks	Quasi-mapping
Humans	0.9187 ± 0.0002	0.8619 ± 0.0013	0.9180 ± 0.0002	0.9144 ± 0.0004
Mice	0.9477 ± 0.0004	0.8702 ± 0.0007	0.9471 ± 0.0003	–
Arabidopsis	0.9889 ± 0.0002	0.9495 ± 0.0002	0.9889 ± 0.0002	–

Table 3 Pairwise Spearman correlation between pipeline results.

	StringTie2-Scallop	StringTie2-Cufflinks	Scallop-Cufflinks	StringTie2-Quasi	Scallop-Quasi	Cufflinks-Quasi
Humans	0.9186 ± 0.0007	$0.9996 \pm 1.6 \times 10^{-5}$	0.9207 ± 0.0006	$0.9974 \pm 5.7 \times 10^{-5}$	0.9166 ± 0.0018	$0.9972 \pm 5.7 \times 10^{-5}$
Mice	0.9300 ± 0.0008	0.9996 ± 0.0000	0.9304 ± 0.0004	–	–	–
Arabidopsis	0.9573 ± 0.0003	$1.0000 \pm 2.4 \times 10^{-6}$	0.9575 ± 0.0004	–	–	–

Table 4 Quantification accuracy metrics.

Pipeline	Spearmanρ	Pearson r	RMSE	MAE
StringTie2	0.9187 ± 0.0011	0.9876 ± 0.0070	22.8394 ± 6.2721	1.2147 ± 0.1021
Scallop	0.8732 ± 0.0024	0.9065 ± 0.0579	96.8980 ± 43.3078	5.9299 ± 0.7146
Cufflinks	0.9175 ± 0.0011	0.9873 ± 0.0072	23.2267 ± 6.2591	1.2373 ± 0.1005
Salmon	0.9142 ± 0.0011	0.9526 ± 0.0240	70.7492 ± 11.2755	4.6940 ± 0.2915

(22.84 ± 6.27) and mean absolute error (MAE) (1.21 ± 0.10), closely followed by Cufflinks, while Salmon and Scallop exhibit substantially larger absolute errors. These results indicate that reliance on Spearman correlation alone would underestimate practical performance differences.

Importantly, all evaluation metrics consistently support the same performance ranking:

$$\text{StringTie2} \approx \text{Cufflinks} > \text{Salmon} > \text{Scallop}$$

demonstrating that the study's conclusions are robust across correlation-based and error-based evaluation criteria.

To address conventional evaluation criteria, we further assessed transcript reconstruction accuracy using gffcompare [22]-derived precision, recall, and F1-score metrics. The result is presented in Table 5.

As evidenced by Table 5, StringTie2 achieved near-perfect performance, followed by Cufflinks, while Scallop exhibited substantially lower recall and overall accuracy.

Importantly, the ranking obtained from structural evaluation is fully consistent with the quantification-based assessment, with StringTie2 outperforming Cufflinks and Scallop across both evaluation paradigms. This agreement demonstrates that the conclusions of our framework are robust across fundamentally different evaluation criteria.

Table 5 Conventional assembly evaluation metrics (precision, recall, and F1-score).

Methods	StringTie2	Scallop	Cufflinks
Precision	0.997	0.575	0.803
Recall	1	0.112	0.988
F1-score	0.9985	0.1875	0.885

However, the two evaluation strategies emphasize distinct aspects of performance. While gffcompare metrics primarily reflect transcript structure reconstruction accuracy, quantification-based metrics capture the fidelity of expression abundance estimation. The combined results indicate that accurate transcript reconstruction does not necessarily guarantee precise expression quantification, highlighting the necessity of incorporating downstream-centric evaluation metrics.

Analysis of anomalous transcripts in simulated data

Building upon the macroscopic Spearman correlation analysis, this part investigates pipeline performance at the individual transcript level.

Given the established high Spearman correlation implying a near-linear relationship between computed TPM values and ground truth

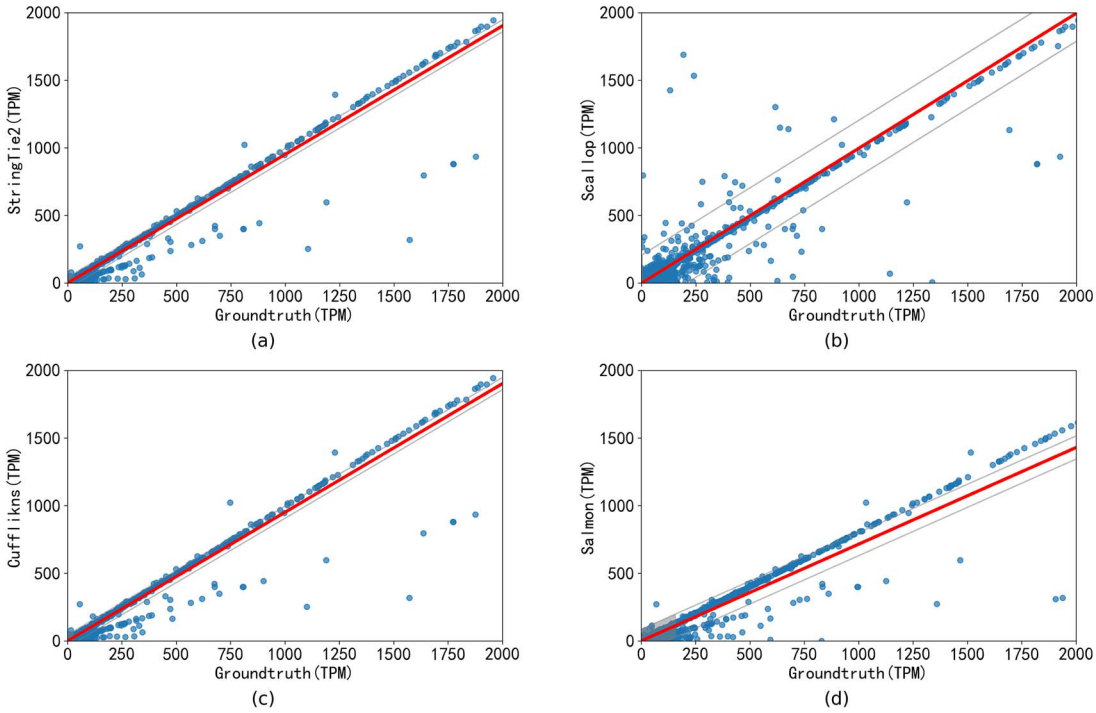

Figure 2 Linear correlation analysis between pipeline results and ground truth TPM values (TPM $\leq$ 2000), and indicates the 95% confidence intervals. Each point represents a transcript. Most transcripts fall within the confidence interval, while outliers represent anomalous transcripts. (a) StringTie2, (b) Scallop, (c) Cufflinks, (d) Salmon quasi-mapping.

Table 6 Count and mean ground truth TPM of anomalous transcripts.

Methods	Count	Mean ground truth TPM
StringTie2	179	1511.6
Scallop	62	1482.4
Cufflinks	179	1502.7
Quasi-mapping	181	2158.4

Table 7 Mean effective length of anomalous transcripts (bp).

Methods	StringTie2	Scallop	Cufflinks	Quasi-mapping
Mean effective length	1266.2475	1543.3256	1280.1916	1394.6132

TPM values, a linear correlation analysis was performed. Transcripts deviating outside the 95% confidence interval were identified as anomalous. Figure 2 illustrates the linear regression analyses (for TPM $\leq$ 2000) for the four pipelines against the ground truth. The red line represents the regression line, the grey band denotes the 95% confidence interval, and each point corresponds to a transcript. Most transcripts reside within the 95% confidence interval.

The quantity of anomalous transcripts and their overall contribution to quantification inaccuracy were statistically analyzed, as presented in Table 6. Although the absolute number of anomalous transcripts per pipeline appears relatively low, these transcripts exhibited significantly elevated average expression levels compared to the transcriptome average. Consequently, despite their small numbers, these highly expressed anomalous transcripts contributed substantially to the total quantification error. Notably, the quasi-mapping method exhibited the largest total TPM deviation, suggesting that the assembly step enhances quantification accuracy relative to direct quasi-mapping. This finding validates the rationale of employing quantification accuracy to evaluate assembly performance.

To identify shared characteristics of these anomalous transcripts, their mean effective lengths were calculated and compared to the overall transcriptome average (3514.4227 bp), as shown in Table 7.

The mean effective lengths of anomalous transcripts were significantly shorter than the overall transcriptome average across all pipelines. Pipelines with higher accuracy (StringTie2, Cufflinks) demonstrated even shorter mean lengths for their anomalous transcripts, indicating a systematic challenge in accurately quantifying shorter transcripts regardless of the specific method used.

Analysis of exon count statistics revealed that anomalous transcripts consistently possessed fewer exons than the transcriptome average (Table 8, Fig. 3). This finding reinforces the association between quantification anomalies and transcripts exhibiting shorter effective lengths and lower exon counts.

Further examination using the IsoPct metric (representing the percentage contribution of a transcript's expression to its parent gene's total expression, provided in the RSEM simulation output) assessed whether anomalies affected most transcripts within a gene (Table 8).

Anomalous transcripts exhibited very high mean IsoPct values (>90% for reference-dependent pipelines and >74% for Scallop), indicating that inaccurate transcripts are often expressed by the same parent gene.

Genomic annotation analysis of the parent genes associated with anomalous transcripts revealed that these genes themselves possessed

Table 8 Structural characteristics of anomalous transcripts and their associated genes.

Methods	Ground truth	StringTie2	Scallop	Cufflinks	Quasi-mapping
Mean exon count per transcript	13.29	6.92	6.65	6.83	7.31
Mean IsoPct value (%)	–	90.97	74.94	90.66	92.70
Mean total exon length of associated genes (bp)	464.33	285.26	367.93	282.82	367.93
Mean exon count per associated gene	9.69	6.88	6.75	6.84	7.22

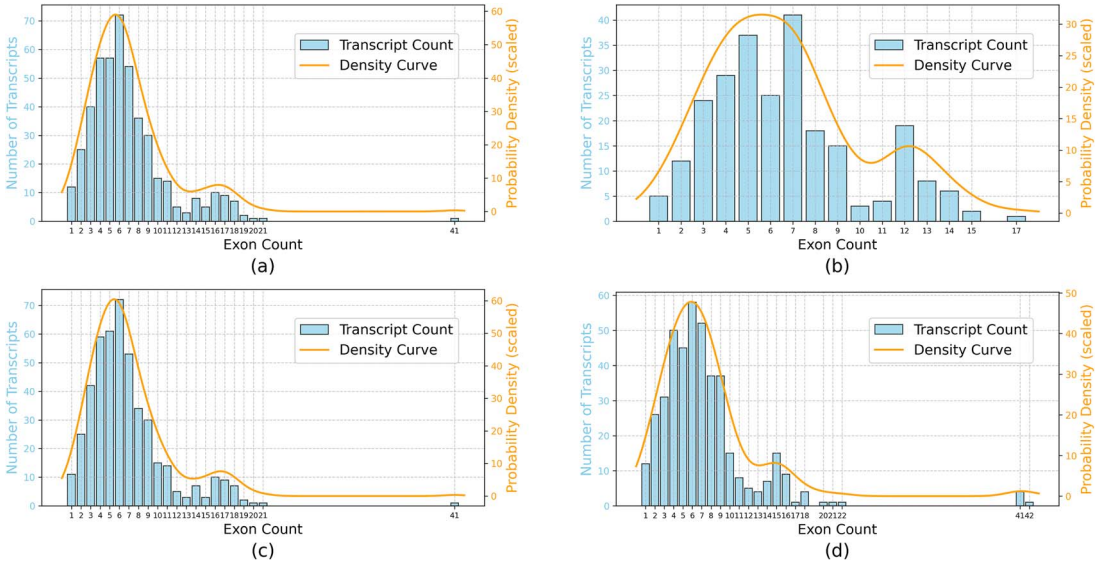

Figure 3 Bar charts and probability density curves of exon count distribution for anomalous transcripts. The distributions illustrate the exon count characteristics of anomalous transcripts identified in different pipelines. Anomalous transcripts consistently exhibit fewer exons compared to the transcriptome average. (a) StringTie2, (b) Scallop, (c) Cufflinks, (d) Salmon quasi-mapping.

significantly fewer and shorter exons compared to the transcriptome average (Table 8). This observation aligns with the transcript-level findings, suggesting that shorter transcripts, often arising from alternative splicing of genes with shorter exons, are particularly susceptible to quantification inaccuracies across the evaluated pipelines.

In summary, transcripts prone to quantification inaccuracies were predominantly characterized by high expression levels, significantly shorter effective lengths, fewer exons, and originated from parent genes possessing fewer and shorter exons.

Analysis of transcripts missing in expression quantification results

As evident from Table 1, a substantial fraction of transcripts present in the ground truth were absent in the quantification outputs (i.e. assigned TPM = 0) of various pipelines. The characteristics of these missing transcripts were analyzed (Table 9).

Missing transcripts exhibited low expression levels (mean TPM < 8 in Table 9) but were notably longer than the transcriptome average. Considering that the RNA-seq data simulated by RSEM had a read length of 100 bp (matching the original data source), longer transcripts are inherently more susceptible to mapping errors where reads may incorrectly map to shorter, homologous regions. Consequently, longer transcripts with low expression levels were more likely to be completely omitted during the assembly or quantification process.

Table 9 Characteristics of transcripts missing from quantification results.

Methods	StringTie2	Scallop	Cufflinks	Quasi-mapping
Mean ground truth TPM of missing transcripts	7.05	4.91	7.08	6.99
Mean effective length (bp)	4625.56	4320.68	4608.47	4631.32

Results on real data

To validate the conclusions drawn from simulated data, we tested 20 diverse human RNA-seq datasets sourced from NCBI (listed in Supplementary Information). Given the absence of ground truth expression values for real data, we employed pairwise Spearman correlation analysis among the expression quantification results of different pipelines. The average correlation coefficients were computed and compared with the corresponding values obtained from simulated human data (specifically, the human data averages from Table 3). This comparative analysis aimed to assess the stability and generalizability of the pipeline performances observed in simulation, and the above results are visualized in Fig. 4.

Analysis of the lower triangular matrix in Fig. 4 (average correlations) and comparison with simulated data (Table 3) revealed several key findings:

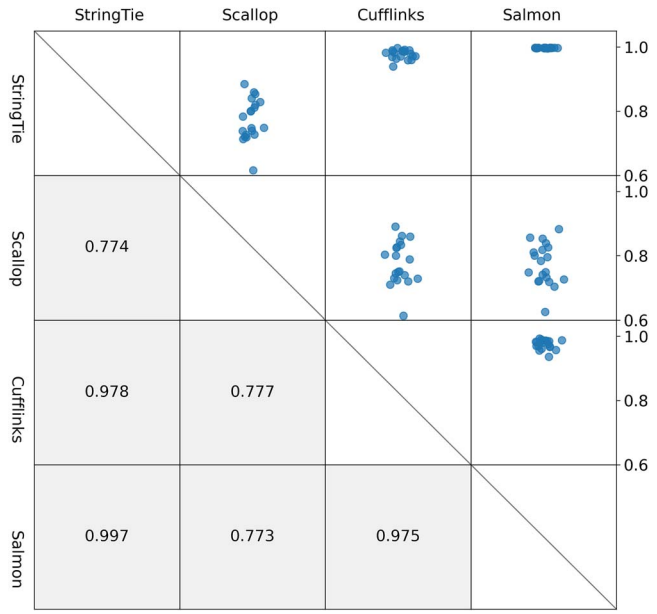

Figure 4 Expression quantification correlations between pipelines on real data. Lower triangular matrix: average Spearman correlation coefficients for each pipeline pair across 20 real RNA-seq datasets. Upper triangular matrix: scatter plots showing the distribution of correlation coefficients for each dataset. The results demonstrate differences in stability and agreement between pipelines when applied to real data.

Exceptional stability of StringTie2 and quasi-mapping

The correlation between StringTie2 and Quasi-mapping exhibited remarkable stability. The average correlation on real data (0.9974) was virtually indistinguishable from its counterpart on simulated data (0.9974).

Notable performance decline

Strikingly, the average pairwise correlations involving Scallop or Cufflinks demonstrated significant decreases when evaluated on real data compared to their simulated counterparts. The discrepancies were substantial: (i) StringTie2-Scallop: 0.7186 (Real) versus 0.9186 (Simulated), (ii) Scallop-Cufflinks: 0.7317 (Real) versus 0.9207 (Simulated), (iii) Cufflinks-Quasi: 0.9682 (Real) versus 0.9972 (Simulated), (iv) Scallop-Quasi: 0.7485 (Real) versus 0.9166 (Simulated).

Relative ranking persistence

Despite the numerical decreases, the relative ranking of pipeline correlations generally mirrored the simulated data findings. Pipelines that showed higher correlations in simulation (e.g. StringTie2-Cufflinks, StringTie2-Quasi) also maintained higher average correlations on real data compared to pairs involving Scallop.

Examination of the upper triangular matrix in Fig. 4 (scatter plots) provided insights into the stability of pairwise correlations:

StringTie2's superior robustness

Plotting against the highly stable Quasi-mapping results served as a reference benchmark. Among the assembly-based pipelines (StringTie2, Cufflinks, Scallop), StringTie2 demonstrated the most stable (tightest scatter, most horizontal alignment) correlation with Quasi-mapping, significantly outperforming both Cufflinks and Scallop in stability.

Discussion

Downstream-centric evaluation provides a complementary perspective

Transcriptome assembly plays a fundamental role in RNA-seq analysis pipelines, as it directly influences downstream transcript quantification and subsequent biological interpretation. Traditional benchmarking of assembly tools typically focuses on structural reconstruction accuracy using metrics such as precision, recall, and F1-score. While these metrics evaluate how accurately transcript structures are reconstructed, they do not directly measure the effect of assembly outputs on downstream quantitative analyses, which are often the primary goal of RNA-seq experiments.

In this study, we introduced a downstream-centric evaluation perspective in which transcriptome assembly algorithms are assessed based on the accuracy and stability of the expression quantification results they produce. By combining simulated datasets with known ground truth and diverse real RNA-seq datasets, our analysis demonstrates that the choice of assembly algorithm can substantially influence transcript abundance estimation even when the quantification tool remains fixed. This finding highlights the importance of evaluating assembly pipelines not only from a structural reconstruction perspective but also from their impact on downstream quantitative analyses.

Robustness of the results across multiple evaluation metrics

The multi-metric evaluation conducted in this study further demonstrates that the observed performance trends are robust across different evaluation criteria. In addition to Spearman's rank correlation, which is commonly used for sparse expression vectors, we incorporated Pearson correlation as well as error-based metrics such as RMSE and MAE. The consistent ranking of pipelines across these complementary metrics indicates that the conclusions are not dependent on a single evaluation measure and remain stable across different statistical perspectives.

Using multiple evaluation metrics also helps mitigate potential biases introduced by outliers or distributional properties of expression data. Rank-based measures capture consistency in transcript ranking, whereas error-based metrics directly quantify absolute deviations in abundance estimation. Together, these complementary perspectives provide a more comprehensive assessment of pipeline performance.

Transcript structural features associated with quantification errors

Our systematic analysis identified two distinct categories of transcripts that are particularly prone to assembly or quantification errors, each associated with specific structural characteristics.

The first category consists of long transcripts with low expression levels. These transcripts were frequently absent from quantification outputs (TPM = 0) across multiple pipelines despite being present in the ground truth. A plausible explanation lies in the inherent limitations of short-read RNA-seq data. When sequencing reads are relatively short (e.g. 100 bp), accurately mapping them across long genomic regions becomes difficult, especially when homologous segments exist elsewhere in the transcriptome. In such situations, reads originating from long transcripts may be incorrectly mapped

to shorter and more highly expressed transcripts, ultimately leading to the omission of the long transcripts during assembly or quantification. Emerging long-read sequencing technologies or hybrid assembly strategies that integrate short- and long-read data may offer promising solutions for improving the recovery of these transcripts.

The second category involves short transcripts derived from genes with relatively few and short exons. In contrast to the first category, these transcripts are often highly expressed but exhibit substantial quantification deviations from the ground truth. Notably, the anomalous transcripts identified in our analysis displayed very high IsoPct values, indicating that they account for a dominant proportion of the total expression of their parent genes. This suggests that the quantification inaccuracies often affect the majority of transcripts within the same gene. Genes associated with these transcripts also tend to have fewer and shorter exons, implying relatively simple genomic structures. These characteristics may introduce challenges for current assembly and quantification algorithms, particularly in resolving alternative splicing events or accurately assigning reads among closely related isoforms. Addressing these limitations may require algorithmic improvements in both transcript reconstruction and abundance estimation models to better capture the structural properties of transcripts originating from compact gene architectures.

Overlap between identified transcript features and structural characteristics of noncoding RNAs

Interestingly, several transcript features associated with substantial quantification deviations in our analysis resemble structural characteristics commonly reported for certain classes of noncoding RNAs (ncRNAs), particularly long noncoding RNAs (lncRNAs). Previous studies have shown that lncRNAs typically exhibit shorter transcript lengths, fewer exons, lower expression levels, and reduced isoform diversity compared with protein-coding genes [23–25].

These characteristics partially overlap with the structural properties observed in transcripts that displayed large quantification deviations in this study. For example, transcripts derived from genes with relatively simple exon architectures were more prone to quantification errors across multiple RNA-seq pipelines. In addition, several transcripts from a same gene showed correlated deviations, suggesting that limited isoform diversity may contribute to shared quantification biases. Although the coding potential of these transcripts was not explicitly examined, the structural similarities observed here suggest that transcripts with simplified exon structures or low expression levels may pose challenges for accurate transcript assembly and quantification.

Limitations of RNA-seq simulation frameworks

Another important observation concerns the discrepancy between simulated and real RNA-seq datasets. While simulated data generated by tools such as RSEM are valuable for benchmarking because they provide ground truth expression values, our results show that pipeline performance can differ substantially when evaluated on real datasets.

In particular, pipelines involving Scallop and Cufflinks exhibited noticeably reduced stability on real RNA-seq datasets compared with their simulated counterparts. This discrepancy indicates that existing RNA-seq simulation frameworks may not fully capture the biological complexity and technical variability present in real transcriptomes. Consequently, benchmarking studies that rely solely on simulated datasets may produce overly optimistic estimates of algorithm performance.

Implications for transcriptome analysis pipeline design

Overall, the results demonstrate that incorporating downstream quantification accuracy into the evaluation of transcriptome assembly pipelines provides a more biologically relevant perspective for assessing RNA-seq analysis workflows. By linking assembly outputs to their functional impact on expression estimation, the proposed framework offers additional insights beyond conventional structural evaluation metrics.

This downstream-aware perspective can help guide the selection of RNA-seq analysis pipelines and may also inform the development of improved transcriptome assembly algorithms capable of handling structurally challenging transcript classes.

Conclusions

This study presents a downstream-centric framework for evaluating transcriptome assembly algorithms based on their impact on transcript expression quantification. By systematically analyzing both simulated and real RNA-seq datasets, we demonstrate that assembly choices can significantly influence quantification results even when the quantification method remains unchanged.

Comparative evaluation of StringTie2, Scallop, and Cufflinks shows that the HISAT2 + StringTie2 pipeline consistently achieves the most stable and accurate quantification performance across multiple evaluation metrics. In addition, our analysis reveals two transcript categories that are particularly prone to quantification errors: long transcripts with low expression levels and short transcripts derived from genes with relatively simple exon structures.

Future work

Building upon the findings and limitations identified in this study, several promising avenues for future research emerge:

Broader benchmarking across diverse data types and transcript structures

Future studies should extend the evaluation framework to encompass a broader range of transcriptomic data types and datasets with increasingly complex transcript structures. In addition to conventional bulk RNA-seq datasets, emerging technologies such as single-cell RNA sequencing (scRNA-seq), long-read RNA sequencing from platforms like Pacific Biosciences and Oxford Nanopore Technologies, as well as spatially resolved transcriptomics, provide valuable opportunities to evaluate pipeline performance under diverse experimental conditions. Furthermore, transcriptomes enriched with diverse noncoding RNA species, particularly lncRNAs and other regulatory transcripts, may introduce additional challenges for transcript assembly and quantification due to their complex and variable structural properties. Benchmarking pipelines across such datasets would provide deeper insights into the robustness, scalability, and generalizability of the proposed downstream-oriented evaluation framework.

Incorporation of additional robust evaluation metrics

Although this study employs multiple complementary metrics—including Spearman correlation, Pearson correlation, RMSE, and

MAE—to evaluate quantification accuracy, future work may further explore robust statistical measures that are less sensitive to extreme values or outliers in expression data. Examples include median-based error metrics or additional rank-based similarity measures, which may provide further insights into the stability of transcript quantification across pipelines.

Development of advanced algorithms

Leverage insights into error-prone transcript features to inform the development of next-generation assembly and quantification algorithms, potentially integrating traditional bioinformatic methods with machine learning (e.g. deep learning models tailored for splice junction prediction or transcript abundance estimation focusing on problematic classes). The goal is to achieve higher accuracy, particularly for vulnerable transcript types (short, low-abundance long, isoforms from genes with short/few exons), and improved computational stability.

Improved simulation frameworks

As observed in this study, existing simulation tools do not fully capture the complexity of real RNA-seq datasets. Developing improved simulation frameworks capable of modeling realistic biases, expression variability, and transcript structural diversity would provide more reliable benchmarks for evaluating transcriptome analysis pipelines.

Refinement of the downstream-centric evaluation standard

Finally, the downstream-oriented evaluation strategy proposed in this work could be further extended to additional bioinformatics tasks, including *de novo* transcriptome assembly, alternative splicing analysis, and differential expression pipelines. Establishing standardized downstream-aware benchmarking practices may contribute to a more comprehensive evaluation of transcriptome analysis tools.

Methods

Selection of data and software

Data selection

RNA sequencing (RNA-seq) data has evolved significantly over the past decade, ranging from Sanger sequencing to second-generation sequencing (NGS) and third-generation sequencing (TGS) technologies. While TGS platforms (e.g. Oxford Nanopore Technologies, PacBio) enable direct RNA molecule sequencing or full-length RNA-seq reads, offering advantages in read length and completeness, NGS platforms (e.g. Illumina) remain dominant for RNA-seq analysis. NGS data boasts higher throughput, more mature analytical pipelines, and significantly lower error rates (1 to 2 orders of magnitude) compared to TGS data. Consequently, this study utilized short-read NGS data obtained from the NCBI Sequence Read Archive (SRA) as real datasets. Simulated data were also generated based on these real datasets to establish ground truth expression values.

Software selection

As this study focuses on reference genome-guided transcriptome assembly and quantification, the following software tools were selected for key processing steps:

Alignment: HISAT2 [12] was chosen for mapping RNA-seq reads to the reference genome, prioritizing alignment accuracy and computational efficiency over alternatives like TopHat2 [26] and STAR [13].

Assembly: three widely used reference-guided transcriptome assemblers were selected: StringTie2 [14] v2.2.1, Scallop [15] v0.10.5, and Cufflinks [17] v2.2.1. This selection enables comparative evaluation of their impact on downstream quantification.

Quantification: Salmon [18] v1.10.1 was chosen as the quantification tool. Its lightweight and accurate "quasi-mapping" mode (which bypasses traditional alignment/assembly but requires a reference transcriptome) was also included for comparative analysis following the approach of Srivastava *et al.* [21]

Data simulation: RSEM [19] v1.3.3 was used to simulate RNA-seq data based on real dataset expression profiles to establish ground truth values. Its common usage and integration with HISAT2 [12] alignment influenced this choice.

Conventional assembly evaluation metrics: to assess transcript reconstruction accuracy using conventional structural metrics, gffcompare [22] v0.12.6 was employed to compare assembled transcripts against reference annotation. This tool computes commonly used evaluation metrics including precision, recall, and F1-score, enabling quantitative assessment of the structural correctness of assembled transcript models.

Determination of evaluation metrics

After selecting the software, we used the Transcripts Per Million (TPM) metric output by Salmon [18] as the normalized expression value for evaluation. Since TPM values across transcripts form a high-dimensional vector, Spearman's rank correlation coefficient (ρ) was initially chosen as the primary metric for comparing quantification results. Spearman's ρ assesses monotonic relationships and is robust to non-normal distributions, making it suitable for expression data. It is defined as follows:

$$\rho = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}$$

where x_i and y_i represent the TPM values of transcript i from two different quantification results or a quantification result and the ground truth.

To complement the rank-based comparison, Pearson's correlation coefficient **(r)** was also calculated to measure the strength of the linear relationship between expression estimates. Pearson's correlation is defined as follows:

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}.$$

While Spearman evaluates rank consistency, Pearson reflects the degree of linear agreement between quantitative estimates.

In addition, two commonly used error-based metrics were included to quantify the magnitude of differences between estimated and reference expression values. The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$$

where n represents the number of transcripts being compared.

The MAE was also calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | x_i - y_i |$$

RMSE penalizes large deviations more strongly due to the squared term, while MAE provides a more interpretable measure of the average absolute difference between estimated and reference TPM values.

To mitigate the impact of sparse vectors containing many zero TPM values on the correlation calculation, transcripts assigned a TPM of zero in either of the two datasets being compared were excluded before computing ρ. This focuses the analysis on the commonly quantified transcripts.

For simulated data, the ground truth TPM file generated by RSEM provides a reference. Spearman's ρ, Pearson's r, RMSE, and MAE were calculated between the quantification results of each pipeline (e.g. HISAT2 + StringTie2 + Salmon) and these ground truth values, enabling direct assessment of quantification accuracy. Pairwise comparisons between pipelines were also performed.

For real data, where ground truth expression levels are unavailable, pairwise correlations were computed between TPM values generated by different pipelines. These correlations were then compared with the pairwise correlations observed on the simulated dataset to assess the consistency and robustness of the pipelines under realistic conditions.

Conventional assembly evaluation metrics

In addition to the quantification-based evaluation framework, we also assessed transcript reconstruction accuracy using conventional assembly evaluation metrics. Specifically, precision, recall, and F1-score were calculated using the gffcompare tool by comparing assembled transcripts against the reference annotation.

Precision measures the proportion of predicted transcripts that correctly match reference transcripts, while recall reflects the proportion of reference transcripts that are successfully reconstructed by the assembly tool. The F1-score represents the harmonic means of precision and recall, providing a balanced measure of reconstruction accuracy. These metrics are defined as:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

Unlike quantification-based metrics that evaluate the accuracy of transcript abundance estimation, these structural metrics primarily assess how accurately transcript models are reconstructed. Therefore, they provide a complementary perspective on pipeline performance, allowing us to compare conventional transcript reconstruction accuracy with downstream-centric quantification fidelity.

Experimental data processing workflow

Generation of simulated data

Reference genome download: human (GRCh38.p14), mouse (GRCm39), and *Arabidopsis thaliana* (TAIR10) reference genome FASTA files and corresponding GTF annotation files were downloaded from NCBI.

Real RNA-seq data download: representative human (SRR7807484), mouse (SRR327047), and *Arabidopsis* (SRR31593541) RNA-seq datasets were downloaded from NCBI SRA.

Reference preparation: the HISAT2 index was built for each reference genome using HISAT2-build.

Expression estimation: RSEM was used to estimate transcript expression levels from the real RNA-seq data:

Reference prepared with rsem-prepare-reference—HISAT2-hca.

Expression calculated using rsem-calculate-expression—paired-end.

Read simulation: simulated paired-end RNA-seq reads (100 bp reads matching the input data characteristics) were generated using rsem-simulate-reads based on the estimated expression profiles. Multiple replicates were generated per species (10 for human, 5 for mouse, 3 for *Arabidopsis*).

Acquisition of real data

Twenty human RNA-seq datasets listed by Srivastava *et al.* [21] were selected (listed in Supplementary information). SRA files were downloaded from NCBI and converted into paired-end FASTQ files using fastq-dump from the SRA Toolkit.

Transcriptome alignment (HISAT2)

A HISAT2 index was built for the human reference genome (GRCh38.p14).

Each real RNA-seq dataset (FASTQ pairs) was aligned to the reference genome using HISAT2 –dta -x -1 -2 -S.

SAM files were converted to sorted BAM files using samtools sort.

Sorted BAM files were indexed using samtools index.

Transcriptome assembly

Assemblers were run using the HISAT2-generated BAM file (sorted, indexed).

StringTie2: StringTie2 -e -G -o (-e indicates expression estimation mode using reference annotations).

Scallop: Scallop -i -o (does not require reference GTF).

Cufflinks: Cufflinks -G -o.

Annotation comparison and filtering (post-assembly for all):

Assembled GTF files were compared to the reference annotation using gffcompare (gffcompare -r) to identify perfectly matching transcript IDs (class code '=').

Assembled transcripts were extracted to FASTA format using gffread (gffread -w -g).

Transcript quantification based on assembly (Salmon)

A Salmon index was built for each assembler's transcript FASTA file: salmon index -t -k 31 -i.

The corresponding real RNA-seq reads (FASTQ files) were quantified against this index: salmon quant -i -l A -1 -2 -o.

The transcript ID and TPM columns were extracted from the resulting quant.sf files.

Salmon quasi-mapping quantification (reference transcriptome)

The human reference transcriptome FASTA file (based on GRCh38.p14) was downloaded.

A Salmon index was built: salmon index -t -k 31 -i.

The real RNA-seq reads were directly quantified against this index: salmon quant -i -l A -1 -2 -o.

The transcript ID and TPM columns were extracted from quant.sf.

Conventional assembly evaluation metrics

gffcompare -r -o eval.

Data processing and correlation analysis

Data filtering: for each pipeline result, transcripts with TPM = 0 were removed. Only transcripts identified as perfectly matching the reference by gffcompare (for assembly-based pipelines) and present in both datasets being compared were retained for correlation analysis. TPM values for these common transcripts formed the expression vectors for comparison.

Simulated data analysis

Ground truth TPM vectors were extracted (excluding TPM = 0 transcripts).

Spearman's ρ, Pearson's r, RMSE, and MAE were calculated between ground truth TPMs and each pipeline's TPMs.

Spearman's ρ was calculated pairwise between pipeline TPM vectors.

Visualization and outlier analysis (e.g. transcripts outside linear regression confidence intervals) were performed to investigate features associated with quantification errors.

Real data analysis

Spearman's ρ was calculated pairwise between TPM vectors generated by the different pipelines.

Average pairwise correlations across the 20 real datasets were computed and compared to the average pairwise correlations observed on the human simulated data.

Data and code availability

The description of the simulated data set and the accession codes for all the real data set used in this research are recorded in Supplemental Materials. The source code for the analysis in this research is available at https://github.com/ZemingTan/Impact-of-Transcriptome-Assembly-Algorithms-on-Downstream-Quantification.

Key Points

- This study systematically analyzes the impact of transcriptome assembly results on downstream quantification in RNA-seq data analysis.
- A quantification-based evaluation framework is proposed to assess assembly algorithms from a downstream perspective rather than traditional reconstruction metrics.
- Comparative experiments on both simulated and real RNA-seq datasets reveal that the HISAT2 + StringTie2 pipeline achieves the most stable and accurate quantification results.
- Two transcript types—long, low-expression and short, highly expressed transcripts—are identified as major sources of quantification error.
- The study highlights the limitations of simulated RNA-seq data and underscores the necessity of real-data-based evaluation for robust algorithm benchmarking.

Acknowledgements

This work was supported by the National Natural Science Foundation of China with code 1247146, the General Program of Guangxi Natural Science Foundation with code 2023GXNSFAA026410, and the Fundamental Research Funds for the Central Universities. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

Z.T.: Investigation, Formal analysis, Data curation, Visualization, Writing – original draft. J.L.: Investigation, Formal analysis, Data curation, Visualization, Writing – original draft. E.Q.: Conceptualization, Methodology, Resources. T.Y.: Conceptualization, Methodology, Supervision, Project administration.

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Competing interests statement

The authors declare that they have no competing interests.

Funding

None declared.

References

1. Grada A, Weinbrecht K. Next-generation sequencing: methodology and application. *J Invest Dermatol* 2013;**133**:1–4. https://doi.org/10.1038/jid.2013.248
2. Lister R, O'Malley RC, Tonti-Filippini J *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 2008;**133**:523–36. https://doi.org/10.1016/j.cell.2008.03.029
3. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 2011;**12**:87–98. https://doi.org/10.1038/nrg2934
4. Garalde DR, Snell EA, Jachimowicz D *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* 2018;**15**:201–6. https://doi.org/10.1038/nmeth.4577
5. Ramsköld D, Luo S, Wang YC *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* 2012;**30**:777–82. https://doi.org/10.1038/nbt.2282
6. Oikonomopoulos S, Wang YC, Djambazian H *et al.* Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci Rep* 2016;**6**:31602. https://doi.org/10.1038/srep31602
7. Weirather JL, de CM, Wang Y *et al.* Comprehensive comparison of Pacific biosciences and Oxford Nanopore technologies and their applications to transcriptome analysis. *F1000Res* 2017;**6**:100. https://doi.org/10.12688/f1000research.10571.2
8. Quail MA, Smith M, Coupland P *et al.* A tale of three next generation sequencing platforms: comparison of ion torrent, Pacific biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012;**13**:341. https://doi.org/10.1186/1471-2164-13-341
9. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet* 2020;**21**:597–614. https://doi.org/10.1038/s41576-020-0236-x
10. Athanasopoulou K, Boti MA, Adamopoulos PG *et al.* Third-generation sequencing: the spearhead towards the radical transformation of modern genomics. *Life* 2022;**12**:30. https://doi.org/10.3390/life12010030

11. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet* 2019;**20**:631–56. https://doi.org/10.1038/s41576-019-0150-2
12. Kim D, Paggi JM, Park C *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;**37**:907–15. https://doi.org/10.1038/s41587-019-0201-4
13. Dobin A, Davis CA, Schlesinger F *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21. https://doi.org/10.1093/bioinformatics/bts635
14. Pertea M, Pertea GM, Antonescu CM *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;**33**:290–5. https://doi.org/10.1038/nbt.3122
15. Shao M, Kingsford C. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat Biotechnol* 2017;**35**:1167–9. https://doi.org/10.1038/nbt.4020
16. Trapnell C, Williams BA, Pertea G *et al.* Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nat Biotechnol* 2010;**28**:511–5. https://doi.org/10.1038/nbt.1621
17. Grabherr MG, Haas BJ, Yassour M *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;**29**:644–52. https://doi.org/10.1038/nbt.1883
18. Patro R, Duggal G, Love MI *et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;**14**:417–9. https://doi.org/10.1038/nmeth.4197
19. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011;**12**:323. https://doi.org/10.1186/1471-2105-12-323
20. Abdurakhmonov IY. *Bioinformatics in the Era of Post Genomics and Big Data*. n.p: BoD – Books on Demand, 2018.
21. Srivastava A, Malik L, Sarkar H *et al.* Alignment and mapping methodology influence transcript abundance estimation. *Genome Biol* 2020;**21**:239. https://doi.org/10.1186/s13059-020-02151-8
22. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare [version 1; peer review: 3 approved]. *F1000Research* 2020;**9**:304. https://doi.org/10.12688/f1000research.23297.1
23. Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. *Cell* 2013;**154**:26–46. https://doi.org/10.1016/j.cell.2013.06.020
24. Derrien T, Johnson R, Bussotti G *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 2012;**22**:1775–89. https://doi.org/10.1101/gr.132159.111
25. Sparber P, Filatova A, Khantemirova M *et al.* The role of long noncoding RNAs in the pathogenesis of hereditary diseases. *BMC Med Genet* 2019;**12**:42. https://doi.org/10.1186/s12920-019-0487-6
26. Kim D, Pertea G, Trapnell C *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013;**14**:R36. https://doi.org/10.1186/gb-2013-14-4-r36

© The Author(s) 2026. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.
Briefings in Bioinformatics, 2026, **27**, bbag267
https://doi.org/10.1093/bib/bbag267
Review