Proceedings of ICBioMed 2025 Symposium: AI for Healthcare: Advanced Medical Data Analytics and Smart Rehabilitation
DOI: 10.54254/2753-8818/2025.AU28984

Statistical Strategies for Differential Expression Analysis of RNA Sequencing Data: Model Comparison and Benchmarking

Yihui Luo

New York University, New York, USA
yl11592@nyu.edu

Abstract. RNA sequencing (RNA-seq) has become the standard technique for genome-wide gene expression profiling. Accurate identification of differentially expressed (DE) genes is critical for understanding biological processes and disease mechanisms. A wide range of statistical models and normalization strategies have been developed, each with distinct assumptions, strengths, and limitations. This review provides a comprehensive overview of statistical approaches for differential expression analysis, including negative binomial models, linear modeling, Bayesian, and non-parametric methods. We discuss normalization techniques, compare methods in terms of sensitivity, specificity, computational efficiency, and robustness, and summarize benchmarking results from simulated and public datasets. Finally, we highlight current challenges, including small sample sizes, zero inflation, batch effects, and complex experimental designs, and discuss future directions involving multi-model integration, deep learning approaches, single-cell adaptations, and multi-omics integration.

Keywords: RNA sequencing, Differential expression analysis, Statistical strategy Yihui Luo

1. Introduction

Over the past decade, RNA sequencing (RNA-seq) has revolutionized transcriptomic studies, providing high-resolution, genome-wide quantification of gene expression. Compared to microarrays, RNA-seq offers several advantages, including higher sensitivity for low-abundance transcripts, the ability to detect novel isoforms, and a larger dynamic range of expression levels [1,2]. These advantages have established RNA-seq as the gold standard for studying gene expression in diverse biological contexts, ranging from fundamental cellular processes to disease-specific mechanisms.

Differential expression (DE) analysis is a central application of RNA-seq, aimed at identifying genes whose expression levels vary significantly across experimental conditions, treatments, or phenotypes. Accurate identification of DE genes provides crucial insights into molecular pathways, regulatory networks, and potential biomarkers, and underpins downstream functional studies and translational research [3,4]. Despite its widespread adoption, RNA-seq data present unique statistical challenges. Count-based measurements are discrete and often over-dispersed, exhibiting a variance

DOI: 10.54254/2753-8818/2025.AU28984

that exceeds the mean. In addition, RNA-seq experiments frequently generate high-dimensional datasets, with thousands of genes measured in relatively few biological replicates. Other complicating factors include batch effects, technical biases related to library preparation and sequencing platform, and variability in gene length or GC content [5,6]. These characteristics complicate statistical modeling and necessitate specialized methods for accurate DE detection.

A variety of statistical models have been developed to address these challenges. Negative binomial models (implemented in DESeq2 and edgeR) are widely used for bulk RNA-seq data, as they account for overdispersion and enable robust inference even with small sample sizes [7,8]. Linear modeling approaches, combined with variance-stabilizing transformations such as voom, allow the application of classical linear model frameworks to RNA-seq data, providing flexibility for complex experimental designs [9]. Bayesian and non-parametric methods offer alternative strategies for datasets with atypical distributions or extreme sparsity [10,11]. Given the diversity of available methods and their varying assumptions, selecting an appropriate DE analysis strategy is non-trivial. Differences in sensitivity, specificity, computational efficiency, and robustness can lead to substantially different results from the same dataset. Therefore, comprehensive benchmarking using both simulated and real RNA-seq datasets is essential to guide method selection and ensure reproducibility [12,13].

The present review aims to provide a detailed overview of statistical strategies for RNA-seq differential expression analysis. We describe the key characteristics of RNA-seq data, summarize normalization approaches, present commonly used statistical models, and compare methods in terms of performance metrics. Furthermore, we examine benchmarking studies and discuss current challenges and future directions, including the adaptation to single-cell RNA-seq, integration with multi-omics data, and potential applications of machine learning and deep learning approaches. This comprehensive perspective is intended to guide researchers in choosing appropriate analytical methods and to highlight areas where methodological improvements are still needed.

2. Background to RNA-seq data analysis

RNA-seq data are fundamentally count-based, representing the number of sequencing reads mapped to each gene. Characteristics include:

• Discrete, non-negative counts: Counts are integers and often follow over-dispersed distributions.

• High dimensionality: Thousands of genes measured with relatively small sample sizes.

• Heteroscedasticity: The variance depends on the mean expression levels.

• Technical biases: Sequencing depth, library preparation, GC content, and batch effects affect measurements [5].

Proper normalization and model selection are essential to ensure robust detection of DE genes.DE analysis typically involves three key steps:

• Normalization: Adjust for library size, sequencing depth, and technical biases.

• Statistical modeling: Estimate expression differences while accounting for variability.

• Multiple testing correction: Control false discovery rate (FDR) using methods such as Benjamini-Hochberg [14].

3. Statistical models and methods

Negative Binomial Distribution Model: The negative binomial (NB) model is widely used for bulk RNA-seq data, modeling overdispersed counts with mean and dispersion parameters. Methods such

DOI: 10.54254/2753-8818/2025.AU28984

as edgeR and DESeq2 use NB models with empirical Bayes shrinkage to improve dispersion estimates [7,8].

• Strengths: Handles overdispersion; suitable for small to medium sample sizes.

• Limitations: May be sensitive to extreme zero counts or large compositional biases.

Linear Modeling Approaches: Linear models, often combined with voom transformation, approximate count data to continuous log-scale expression values, enabling the use of classical linear modeling tools (limma) [9].

• Strengths: Efficient for large samples and multi-factor designs; flexible for complex experimental setups.

• Limitations: Assumes normality after transformation; may be less accurate for low-count genes.

Bayesian and Non-parametric Methods: Bayesian methods (e.g., baySeq): Estimate posterior probabilities of DE genes, suitable for small sample sizes but computationally intensive [10].

Non-parametric methods (e.g., NOISeq): Do not assume specific count distributions; robust to non-standard data but may have reduced sensitivity [11].

4. Normalization strategies

Normalization ensures comparability across samples by mitigating sequencing depth, library composition, and technical biases. Common approaches include:

• RPKM/FPKM/TPM: Corrects for gene length and sequencing depth, suitable for expression comparisons but less optimal for DE analysis [1,15].

• Upper Quartile (UQ) Normalization: Reduces the impact of highly expressed genes [5].

• TMM (Trimmed Mean of M-values): Adjusts for compositional biases; robust for bulk RNA-seq [8].

• DESeq Size Factor Normalization: Uses geometric means to estimate scaling factors; effective for small samples [16].

• Quantile Normalization: Ensures identical distributions across samples; may obscure true biological variation [17].

• Advanced Methods: RUV, SCnorm, cqn address batch effects, zero inflation, and GC content bias [6,18].

5. Model comparison

Sensitivity and specificity for different methods are shown in Table 1.

DOI: 10.54254/2753-8818/2025.AU28984

Table 1. Sensitivity and specificity of different methods

Method	Small samples	Medium samples	Large samples	Characteristics
DESeq2	High specificity, moderate sensitivity	Balanced	Stable	Shrinkage of dispersion
edgeR	High sensitivity	High sensitivity, good specificity	Medium	Empirical Bayes shrinkage
limma-voom	Moderate	High	Optimal	Linear modeling with voom weights
baySeq	Moderate	Moderate	Limited	Bayesian posterior probabilities
NOISeq	Moderate-low	Moderate	High	Non-parametric noise modeling

In terms of computational efficiency, limma-voom is suitable for large-scale datasets, edgeR for small to medium-sized datasets, DESeq2 is computationally intensive, and baySeq is computationally demanding due to its use of MCMC algorithms. Regarding robustness, DESeq2 and edgeR perform excellently with low counts and small sample sizes. limma-voom demonstrates robustness for large, multi-factor datasets. NOISeq adapts to non-standard data but exhibits sensitivity to high noise levels.

In summary, the applicability of different methods in differential expression analysis is outlined as follows:

• Small samples (n<5): DESeq2 is recommended, balancing sensitivity and false positive control.

• Medium samples (n=5-20): edgeR offers slightly superior sensitivity, while DESeq2 provides marginally better specificity.

• Large samples (n>20): limma-voom delivers the most comprehensive performance.

• Exploratory or atypical data: NOISeq and baySeq may serve as supplementary options.

6. Benchmarking

The performance evaluation of RNA-seq differential expression analysis methods, based on different models and normalization strategies, is of paramount importance for method selection and result reliability. Benchmarking typically involves testing against evaluation metrics, simulated data, and public benchmark datasets [4,12]. For a detailed comparison, see Table 2.

In the evaluation of RNA-seq methods, commonly used metrics include:

• True Positive Rate (TPR / Sensitivity): Measures the ability to identify genuinely differentially expressed genes.

• False Positive Rate (FPR): Measures the proportion of non-differentially expressed genes incorrectly identified as differentially expressed.

• Accuracy and Precision: Precision = TP / (TP + FP), used to evaluate result reliability.

• False Discovery Rate (FDR) control: Employing Benjamini-Hochberg correction to ensure overall controllable significance levels [14].

• Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC): Comprehensive evaluation of sensitivity and specificity, suitable for method performance comparison.

• Computational Efficiency (Time and Memory): For large-scale projects such as GTEx or TCGA, computational efficiency is a critical factor in practical applications.

DOI: 10.54254/2753-8818/2025.AU28984

Common datasets include:

• SEQC/MAQC-III: Multi-laboratory RNA-seq data for evaluating technical reproducibility and method robustness [19].

• GTEx (Genotype-Tissue Expression): Cross-tissue RNA-seq data for evaluating differential expression method performance across large-scale samples [20].

• ENCODE RNA-seq: High-quality datasets simulating diverse experimental design conditions [21].

Table 2. Benchmark findings

Method	Sensitivity	Specificity	FDR control	Computational efficiency	Notes
DESeq2	High	High	Stable	Medium	Small-medium samples
edgeR	High	Moderate-High	Stable	High	Slightly higher sensitivity
Limma-voom	Moderate-High	High	Stable	High	Large datasets
baySeq	Moderate	High	Controllable	Low	Small samples, heavy computation
NOISeq	Moderate-Low	Moderate-High	N/A	Medium	Robust for non-standard data

7. Challenges and future directions

Despite the remarkable progress in RNA-seq technologies and statistical methods for differential expression analysis, several challenges remain that hinder the full realization of its potential. A fundamental issue is the combination of small sample sizes and high-dimensional data, a common scenario in many biological experiments. The limited number of replicates relative to the number of genes reduces statistical power, particularly for lowly expressed genes, and increases the risk of false positives and false negatives, even when advanced models such as negative binomial or Bayesian approaches are employed [4,13]. In addition, the intrinsic sparsity and zero inflation of RNA-seq data, especially in single-cell RNA-seq, further complicate accurate detection of differentially expressed genes. Traditional statistical models often fail to adequately account for the large proportion of zero counts, leading to biased estimates of dispersion and fold changes [22]. Although specialized methods such as ZINB-WaVE or SCnorm have been proposed to address zero inflation, their computational complexity and sensitivity to parameter tuning present practical limitations for large-scale datasets.

Batch effects and technical biases pose additional challenges. Variability arising from library preparation, sequencing platforms, and laboratory-specific conditions can overshadow genuine biological differences. Methods such as RUV and ComBat can mitigate batch effects to some extent, yet complete removal of these confounders remains difficult, and residual technical noise may still affect downstream inference [6,23]. Moreover, contemporary RNA-seq experiments often involve complex designs with multiple conditions, time points, or tissue types. Accurately modeling such multi-factorial experiments requires linear mixed models or generalized linear models, which may still be limited by small sample sizes and high variability [24]. Another critical consideration is that statistical models rely on specific assumptions about data distribution. Deviations from these

DOI: 10.54254/2753-8818/2025.AU28984

assumptions, such as extreme zero inflation, compositional biases, or highly skewed expression patterns, can compromise the accuracy of differential expression detection.

Looking forward, several directions are emerging to overcome these challenges. Ensemble approaches that integrate multiple statistical methods may leverage complementary strengths, improving sensitivity, specificity, and robustness across diverse datasets [4]. Deep learning and generative modeling techniques, including variational autoencoders and generative adversarial networks, offer potential to capture complex, non-linear relationships and to generate realistic synthetic datasets for benchmarking or small-sample augmentation [25]. The rapid expansion of single-cell RNA-seq further motivates the development of models that explicitly account for sparsity, dropout events, and cross-batch variation. Integrating multi-omics data, including transcriptomics, epigenomics, and proteomics, holds promise for enhancing biological interpretability and identifying regulatory mechanisms that cannot be inferred from gene expression alone [26]. Finally, the development of standardized, automated, and reproducible pipelines is essential for ensuring consistent analysis across studies, enabling reliable cross-study comparisons, and fostering reproducibility in the field.

In summary, while RNA-seq differential expression analysis has advanced substantially, challenges related to small sample sizes, zero inflation, batch effects, complex experimental designs, and model assumptions persist. Continued methodological innovations, particularly those integrating multiple statistical frameworks, machine learning techniques, single-cell adaptations, and multi-omics data, are expected to improve the accuracy, robustness, and biological relevance of RNA-seq studies in the coming years.

References

[1] Mortazavi A, Williams B A, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-seq [J]. Nature Methods, Nature Publishing Group, 2008, 5(7): 621–628.
[2] Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics [J]. Nature Reviews Genetics, Nature Publishing Group, 2009, 10(1): 57–63.
[3] Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis [J]. Genome Biology, 2016, 17(1): 13.
[4] Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data [J]. BMC Bioinformatics, 2013, 14(1): 91.
[5] Bullard J H, Purdom E, Hansen K D, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments [J]. BMC Bioinformatics, 2010, 11(1): 94.
[6] Risso D, Ngai J, Speed T P, et al. Normalization of RNA-seq data using factor analysis of control genes or samples [J]. Nature Biotechnology, Nature Publishing Group, 2014, 32(9): 896–902.
[7] Love M I, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 [J]. Genome Biology, 2014, 15(12): 550.
[8] Robinson M D, McCarthy D J, Smyth G K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data [J]. Bioinformatics, Oxford Academic, 2010, 26(1): 139–140.
[9] Law C W, Chen Y, Shi W, et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts [J]. Genome Biology, 2014, 15(2): R29.
[10] Hardcastle T J, Kelly K A. baySeq: empirical bayesian methods for identifying differential expression in sequence count data [J]. BMC Bioinformatics, 2010, 11(1): 422.
[11] Tarazona S, García F, Ferrer A, et al. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases [J]. Embnet.journal, 2011, 17(B): 18–19.
[12] Rapaport F, Khanin R, Liang Y, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data [J]. Genome Biology, 2013, 14(9): 3158.
[13] Schurch N J, Schofield P, Gierliński M, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? [J]. RNA (new York, N.Y.), 2016, 22(6): 839–851.

DOI: 10.54254/2753-8818/2025.AU28984

[14] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing [J]. Journal of the Royal Statistical Society Series B (Statistical Methodology), Oxford Academic, 1995, 57(1): 289–300.
[15] Li B, Ruotti V, Stewart R M, et al. RNA-seq gene expression estimation with read mapping uncertainty [J]. Bioinformatics (oxford, England), 2010, 26(4): 493–500.
[16] Anders S, Huber W. Differential expression analysis for sequence count data [J]. Genome Biology, 2010, 11(10): R106.
[17] Bolstad B M, Irizarry R A, Astrand M, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias [J]. Bioinformatics (oxford, England), 2003, 19(2): 185–193.
[18] Bacher R, Chu L-F, Leng N, et al. SCnorm: Robust normalization of single-cell RNA-seq data [J]. Nature Methods, 2017, 14(6): 584–586.
[19] Su Z, Łabaj P P, Li S, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium [J]. Nature Biotechnology, Nature Publishing Group, 2014, 32(9): 903–914.
[20] The GTEx Consortium, Ardlie K G, Deluca D S, et al. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans [J]. Science, American Association for the Advancement of Science, 2015, 348(6235): 648–660.
[21] Dunham I, Kundaje A, Aldred S F, et al. An integrated encyclopedia of DNA elements in the human genome [J]. Nature, Nature Publishing Group, 2012, 489(7414): 57–74.
[22] Vallejos C A, Marioni J C, Richardson S. BASiCS: bayesian analysis of single-cell sequencing data [J]. PLOS Computational Biology, Public Library of Science, 2015, 11(6): e1004333.
[23] Leek J T, Scharpf R B, Bravo H C, et al. Tackling the widespread and critical impact of batch effects in high-throughput data [J]. Nature Reviews Genetics, Nature Publishing Group, 2010, 11(10): 733–739.
[24] Hoffman G E, Schadt E E. variancePartition: interpreting drivers of variation in complex gene expression studies [J]. BMC Bioinformatics, 2016, 17(1): 483.
[25] Eraslan G, Simon L M, Mircea M, et al. Single-cell RNA-seq denoising using a deep count autoencoder [J]. Nature Communications, Nature Publishing Group, 2019, 10(1): 390.
[26] Hoadley K A, Yau C, Hinoue T, et al. Cell-of-origin patterns dominate the molecular classification of 10, 000 tumors from 33 types of cancer [J]. Cell, 2018, 173(2): 291-304.e6.