6

Statistical Methods for RNA Sequencing Data Analysis

Dongmei Li

Clinical and Translational Science Institute, University of Rochester School of Medicine and Dentistry, Rochester, NY, USA

Author for correspondence: Dongmei Li, Clinical and Translational Science Institute, University of Rochester School of Medicine and Dentistry, 265 Crittenden Boulevard CU 420708, Rochester, NY, USA. Email: Dongmei_Li@urmc.rochester.edu

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch6

Abstract: This chapter will review the statistical methods used in RNA sequencing data analysis, including bulk RNA sequencing and single-cell RNA sequencing. RNA sequencing data analysis has been widely used in biomedical and biological research to identify genes associated with certain conditions or diseases. Many statistical methods have been proposed to analyze bulk and single-cell RNA sequencing data. Several studies have compared the performance of different statistical methods for RNA sequencing data analysis through simulation studies and real data evaluations. This chapter will summarize the statistical methods and the evaluation results for comparing different statistical analysis methods used for RNA sequencing data analysis. It will cover the statistical models, model assumptions, and challenges encountered in the RNA sequencing data analysis. It is hoped that this chapter will help researchers learn more about the statistical perspective of the RNA sequencing data analysis and enable them to choose appropriate statistical analysis methods for their own RNA sequencing data analysis.

Keywords: bulk RNA sequencing; data analysis; differential analysis; RNA sequencing; single-cell RNA sequencing

In: *Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

INTRODUCTION

RNA sequencing, including bulk RNA sequencing and single-cell RNA sequencing, is a popular technology used in biological and biomedical fields (1, 2). Figure 1 shows the analysis flow of RNA sequencing data. In RNA sequencing experiments, RNAs of interest need to be extracted first from the cells and then converted to complementary DNA (cDNA) to be sequenced by high-throughput platforms. Next, the sequenced short cDNA fragments are mapped to a genome or a transcriptome, and the summarized count data are derived to estimate the expression levels for each gene or isoform (3–5). Finally, statistical methods or

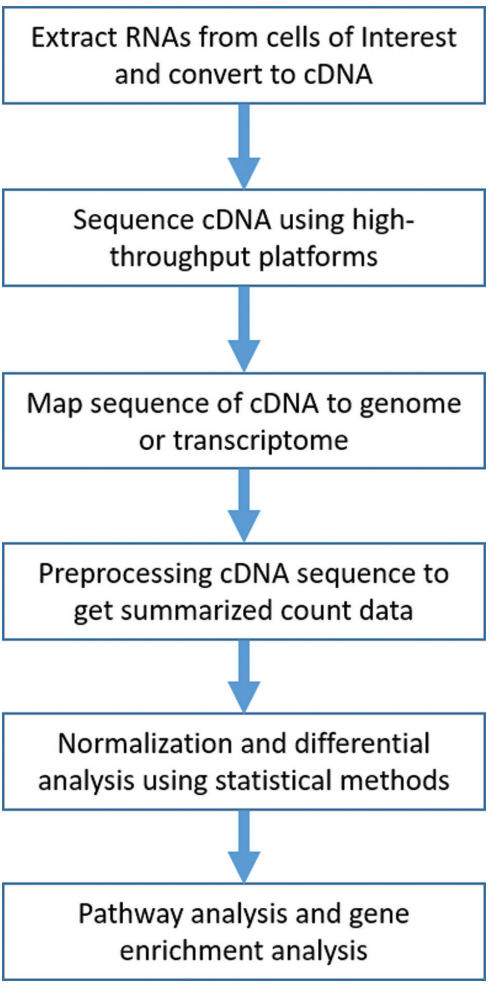

Figure 1 **Analysis flow of RNA sequencing data.**

machine learning methods are applied to the summarized count data after normalization to evaluate transcription levels under different biological and biomedical conditions, to discover novel transcripts and isoforms, and to detect alternative splicing and splice junctions (6). The single-cell RNA sequencing, in addition, allows to understand gene expression pattern within the cell; to identify cell heterogeneity, cell population, and sub-population; and to examine the effects of low copy mRNA distribution and transcriptional regulation (7). Pathway analysis and gene enrichment analysis are usually performed further on selected significant genes after differential analysis (8, 9).

RNA sequencing has been widely used to study the mechanism of complex disease, identify potential biomarkers for clinical indications and infer gene pathways (10–12). Similar to bulk RNA sequencing, single-cell RNA sequencing has been applied to identify cell populations, infer gene regulatory networks, and track different cell lineages (13–15). Single-cell RNA sequencing also has the potential to identify drug-resistant clones, assist non-invasive biopsy diagnosis, and infer stem cell regulatory networks (16–18).

As the sequencing technology advances rapidly, the cost of both bulk RNA sequencing and single-cell RNA sequencing also dramatically decreased (18, 19). With this massive amount of RNA sequencing data now available, it is very challenging to obtain accurate information from the data and further transform this information into useful knowledge (20, 21). Differential gene expression analysis also has its own challenges. The distribution of read coverage might be different along the genome attributed to the variation of genome compositions. Meanwhile, larger genes have more mapped reads than smaller genes although their expression levels might be the same. Furthermore, many biological variations sometimes cannot be accounted for in the data analysis due to relatively small sample sizes for each experimental condition. This chapter focuses on the statistical analysis methods used for differential analysis in both bulk RNA sequencing and single-cell RNA sequencing data. Commonly used statistical methods, their model assumptions, and tests for RNA sequencing differential analysis are discussed (Table 1). The simulation results of comparing different statistical methods and challenges encountered in the data analysis are summarized. Recommendations on the selection of appropriate statistical methods for RNA sequencing differential analysis are also provided.

STATISTICAL METHODS FOR BULK RNA SEQUENCING DIFFERENTIAL ANALYSIS

Current popular methods for bulk RNA-seq differential analysis methods could be classified into four categories based on the type of statistical methods used for differential analysis: (i) *t*-test analogical methods (Cuffdiff and Cuffdiff2) (22, 23), (ii) Poisson or negative binomial model-based methods (edgeR, DESeq, DESeq2, baySeq, EBSeq) (24–29), (iii) non-parametric methods (SAMseq and NOIseq) (30–32), and (iv) linear models (voom and sleuth) (33, 34).

TABLE 1 **Summary of gene differential expression analysis methods for bulk RNA and single-cell RNA sequencing data**

Bulk RNA sequencing data			
Method	**Read count distribution assumption/model**	**Differential analysis test**	**Reference**
Cuffdiff and Cuffdiff2	Similar to *t*-distribution on log-transformed data	*t*-test analogical method	(22, 23)
edgeR	Negative binomial distribution	Exact test analogous to Fisher's exact test or likelihood ratio test	(24, 25)
DESeq	Negative binomial distribution	Exact test analogous to Fisher's exact test	(26)
DESeq2	Negative binomial distribution	Wald test	(27)
baySeq	Negative binomial distribution	Posterior probability through Bayesian approach	(28)
EBSeq	Negative binomial-beta empirical Bayes model	Posterior probability through Bayesian approach	(29)
SAMseq	Non-parametric method	Wilcoxon rank statistics based permutation test	(30)
NOIseq	Non-parametric method	Corresponding logarithm of fold change and absolute expression differences have a high probability than noise values	(31, 32)
voom	Similar to *t*-distribution with empirical Bayes approach	Moderated *t*-test	(33)
Sleuth	Additive response error model	Likelihood ratio test	(34)
Single-cell RNA sequencing data			
Method	**Read count distribution assumption/model**	**Differential analysis test**	**Reference**
SCDE	Two-component mixture model with Poisson and negative binomial distributions	Posterior probability of being differentially expressed through Bayesian approach	(40)
MAST	Hurdle model with indicator variable and logistic regression	Differences in summarized regression coefficients between groups through bootstrap method	(41)
scDD	Bayesian modeling approach	Bayes factor score through permutation method	(42)
DEsingle	zero-inflated negative binomial model	Likelihood ratio test	(43)
SigEMD	Logistic regression and Wald test for selecting genes with zero count and then impute zero counts using the Lasso regression model	Non-parametric test based on Earth Mover's Distance (EMD) through permutation method	(44)

Table continued on following page

TABLE 1 Summary of gene differential expression analysis methods for bulk RNA and single-cell RNA sequencing data (Continued)

Single-cell RNA sequencing data			
Method	**Read count distribution assumption/model**	**Differential analysis test**	**Reference**
SINCERA	Exact or normal distribution	Welch's *t*-test or Wilcoxon rank sum test	(46)
D^3E	Discrete distribution	Cramér-von Mises test, Kolmogorov–Smirnov test or likelihood ratio test	(47)
EMDomics	Distribution functions are different	EMD-based permutation test	(48)
Monocle2	Generalized linear model approach	Likelihood ratio test	(45, 51)
Linnorm	*t*-distribution with empirical Bayes approach	Moderated *t*-test	(49)
Discriminative Learning	Multiple logistic regression model	Likelihood ratio test	(50)

Cuffdiff and Cuffdiff2

Both the Cuffdiff and Cuffdiff2 methods use the *t*-test analogical method to test the changes in gene expression levels between different groups (22, 23). The mean gene expression level for each gene is determined using the maximum likelihood estimating method for different groups. Then, the mean difference of the logarithm-transformed gene expression levels of the estimated gene expression levels is used as the numerator in the *t*-test analogical method, and the estimated variance of the mean differences in logarithm is estimated using the delta method. The power of the *t*-test analogical method in Cuffdiff and Cuffdiff2 depends on the length of the transcripts tested as longer transcripts yield more reads. Thus, the results from Cuffdiff and Cuffdiff2 are biased toward a higher probability of identifying longer transcripts or genes. The major differences between Cuffdiff and Cuffdiff2 are methods used to extrapolate the estimated gene expression levels. Cuffdiff determines the estimated gene expression levels using the maximum likelihood method with the Bayesian approach and Poisson distribution assumption, while the Cuffdiff2 method improves the estimation of gene expression levels through modeling cross-replicate variability in transcript-level counts and adopts the negative binomial distribution assumption for those estimated counts.

edgeR

For each gene in each sample, edgeR assumes that the summarized count follows a negative binomial distribution with mean equal to the multiplication of library size and relative abundance (the gene expression levels), and the variance for each

gene is a function of the mean (24, 25). The genewise dispersion is estimated using a conditional maximum likelihood method through the empirical Bayes approach. For gene differential expression testing, edgeR uses either an exact test analogous to Fisher's exact test with consideration of overdispersion or a likelihood ratio test within a negative binomial generalized log-linear model framework.

DESeq and DESeq2

DESeq uses a modified negative binomial model implemented in edgeR (26). DESeq estimates the variance based on the relative abundance of the gene through a data-driven approach. DESeq tests gene expression differences between groups using an exact test analogous to Fisher's exact test with test statistics as the sum of total count within each group and across groups. DESeq2 takes a generalized linear model approach to model the group differences in relative abundance, which can also accommodate more complex study designs (27). DESeq2 assumes that the dispersion follows a log normal prior distribution with means being a function of normalized counts for each gene. DESeq2 uses an empirical Bayes approach to integrate the dispersion and fold change estimates and tests the gene differential expression using the Wald test.

baySeq

baySeq assumes that the summarized count data follow a negative binomial distribution and use the whole dataset to obtain a prior distribution for the estimated model parameters (28). The data dispersion is approximated using the maximum likelihood method. The baySeq method uses a posterior probability of non-differential expression between groups and a Bayesian FDR estimate to select significantly differentially expressed genes between groups.

EBSeq

EBSeq assumes that within each biological condition, the expected count from each gene follows a negative binomial distribution (29). Within each group, the mean of gene expressions is a function of the variance of gene expressions. The variance of gene expressions follows a beta distribution with the two parameters estimated using the expectation-maximization (EM) algorithm. For the gene expression differential tests between groups, EBSeq obtains a posterior probability of genes being differentially expressed between groups through Bayes' rule using the EM algorithm within the negative binomial-beta empirical Bayes model framework. EBSeq also uses a Bayesian FDR estimate to assist the selection of significantly differentially expressed genes.

SAMseq

SAMseq is a non-parametric method proposed for differential gene expression testing between groups (30). For between-group comparisons, SAMseq uses the two-sample Wilcoxon rank statistics. SAMseq uses a re-sampling procedure to

account for different sequencing depths in the differential data analysis. The null distribution of the Wilcoxon rank statistic and FDR are estimated using the permutation method.

NOIseq

NOIseq is also a non-parametric method for testing differential gene expression between groups through ratio of fold change and absolute expression differences (31, 32). NOIseq uses sequencing-depth corrected and normalized RNA sequencing count data and models the noise distribution by contrasting the logarithm of fold change and absolute expression differences between groups. NOIseq considers a gene to be differentially expressed between groups if the corresponding logarithm of fold change and absolute expression difference values have a high probability to be higher than noise values.

voom

voom takes a linear modeling strategy to model the count data (33). It determines the mean–variance relationship based on the delta rule and Taylor's theorem and obtains the estimate for variance through the piecewise linear function defined by the fitted LOWESS curve. voom also generates a weight for each observation and uses the estimated variance and weight as the input in the limma empirical Bayes analysis pipeline. The gene expression differential analysis between groups is tested using the moderated *t*-statistics.

Sleuth

Sleuth uses an additive response error model with the total between-sample variability being an additive of biological variance and inferential variance (34). The biological variance is composed of between-sample variation and variation during the library preparation process. The inferential variance includes variation due to random sequencing of fragments and variation coming from computational inference procedures. Sleuth tests gene differential expression between groups using the likelihood ratio test.

STATISTICAL METHODS COMPARISONS FOR BULK RNA SEQUENCING DIFFERENTIAL ANALYSIS

In 2013, Soneson conducted an extensive comparison of 11 methods used for bulk RNA sequencing differential analysis through both simulation studies and real RNA sequencing data examples (35). The methods Soneson compared include edgeR, DESeq, baySeq, EBSeq, SAMseq, and voom, described before. The comparison of those methods showed that all methods had low power with small sample sizes, and there was no optimal method applicable for all conditions. voom performed well under many conditions and was robust to outliers and computationally efficient. However, voom performed worse when the variances were

unequal between groups. SAMseq requires larger sample sizes (at least 4–5 samples per group) to detect significantly differentially expressed genes. The comparison also found that DESeq was often overly conservative, and edgeR was too liberal with a larger number of false positives. Both baySeq and EBSeq were computationally less efficient. baySeq showed highly variable results when significant genes were all modulated in one direction, and the results were largely affected by outliers. EBseq had a poor false discovery rate (FDR) control in most situations and was relatively robust to outliers.

Previous experimental validation of selected differentially expressed genes from multiple RNA sequencing differential expression analysis methods (Cuffdiff2, edgeR, DESeq2) found a high FDR of the Cuffdiff2 method and high false negative rates of the DESeq2 method (36). The edgeR method had relatively higher sensitivity and specificity than the Cuffdiff2 and DESeq2 methods. In addition, the experimental validation also showed that pooled samples in the experiments suffered from lower positive predictive values than individual samples.

Using results from qRT-PCR as the gold standard, an extended review of eight RNA sequencing differential analysis methods (baySeq, DESeq, DESeq2, EBSeq, edgeR, voom, NOIseq, and SAMseq) was conducted to determine their precision, accuracy, and sensitivity (37). By comparing the results from qRT-PCR and selected differentially expressed genes from each of the eight methods, it was found that voom, NOIseq, and DESeq2 showed more consistent results than the other methods. In addition, the significantly differentially expressed genes selected by consensus of baySeq, DESeq2, voom, and NOISeq had the best performance indicators on precision, accuracy, and sensitivity. Furthermore, the investigation also found that mapping methods in the pre-processing step of RNA sequencing data analysis had minimal effect on downstream RNA sequencing gene differential analysis, given that a reference genome for the RNA sequencing data was available.

A recent investigation of six RNA sequencing differential analysis methods (DESeq, DESeq2, edgeR, SAMseq, EBSeq, and voom) focused on their stability measured by the area under the correlation curve (38). Among the explored factors that have a potential to affect the stability of RNA sequencing differential analysis methods, fold changes of truly differentially expressed genes and their variability seem largely to affect the stability of those methods. Larger sample size is associated with increased stability, and a sample size of 10 or larger in each group results in a plateau on stability. DESeq2 and edgeR were less likely to be affected by outliers on their stability measurements.

STATISTICAL METHODS FOR SINGLE-CELL RNA SEQUENCING DIFFERENTIAL ANALYSIS

Single-cell RNA sequencing is becoming popular in recent years to better understand the stochastic process and gene regulations in a granular resolution (13, 15, 16, 39). The commonly used gene differential expression analysis in single-cell RNA sequencing can be classified into two categories, with one category modeling excess zeros (SCDE, MAST, scDD, DEsingle, and SigEMD) (40–44) and the other category without modeling the excess zeros in the single-cell RNA sequencing data (DESeq2, SINCERA, D^3E, EMDomics, Monocle2, Linnorm, and Discriminative Learning)

(12, 27, 45–50). DESeq2 is a popular method used for bulk RNA sequencing data analysis, which is also often used for analyzing single-cell RNA sequencing data for testing of differential expression between groups.

Single-cell differential expression (SCDE)

SCDE uses a two-component mixture model for the gene expression data from single-cell RNA sequencing experiments (40). The excess zero part (dropouts) is modeled by a Poisson distribution with user-specified thresholds for the mean (such as 0.1). The expressed genes are modeled by a negative binomial distribution technique. For gene differential expression analysis between groups, SCDE takes a Bayesian approach to obtain the posterior probability of a gene being expressed in one group and then uses a fold expression difference between groups as the test statistics with empirical P-values calculated to select differentially expressed genes.

MAST

MAST uses a hurdle model approach for single-cell RNA sequencing gene differential expression analysis (41). MAST assumes conditional independence between expression rate and expression levels for each gene. MAST uses an indicator variable to denote whether a gene is expressed in a cell and fits a logistic regression for the discrete indicator variable. For genes expressed in a cell, MAST fits a normally distributed linear model. The gene differential expression analysis between groups is tested using the differences in summarized regression coefficients between groups. The null distribution of the test statistics is estimated through a bootstrap method with empirical Bayes approach regularizing model parameters.

scDD

scDD is also based on a Bayesian modeling approach to detect differentially expressed genes between groups (42). ScDD models the excess zeros using a logistic regression and models the non-zero gene expressions using a conjugate Dirichlet process mixture model of normal distributions. For testing gene differential expressions, scDD calculates an approximate Bayes factor score that compares the probability of differential expression with the probability of non-differential expression. The empirical P-values for the differential expression tests are computed using a permutation method.

DEsingle

DEsingle uses a zero-inflated negative binomial (ZINB) model to characterize the read counts and excess zeros in single-cell RNA sequencing data (43). The ZINB model has two components, one modeling the excess zeros through an indicator variable multiplied by the proportion of constant zeros and the other modeling the positive gene expressions through a negative binomial model multiplied by the proportion of non-zeros. The gene differential expression analysis is conducted through likelihood ratio tests within the ZINB model framework.

SigEMD

Different from other excess zero modeling methods for single-cell RNA sequencing differential analysis, SigEMD takes an additional step in modeling the excess zeros (44). SigEMD first uses logistic regression and the Wald test to select genes with zero counts that are affecting gene expression distributions, then SigEMD imputes those zero counts through a Lasso regression model. The gene differential analysis between groups is conducted using a non-parametric test based on Earth Mover's Distance (EMD). The P-values are computed using a permutation method.

SINCERA

SINCERA is a pipeline developed for single-cell RNA sequencing data analysis (46). SINCERA can be used for the pre-processing of single-cell RNA sequencing data, identifying cell types and key gene expression regulators, selecting differentially expressed genes, and predicting gene signatures. For gene differential analysis between groups, SINCERA uses the Welch's *t*-test when the sample size of both groups is >5; otherwise, SINCERA uses the Wilcoxon rank sum test. SINCERA also includes the SAMseq algorithm as an optional method for selecting differentially expressed genes from single-cell RNA sequencing data.

D^3E

D^3E is a discrete distribution method used for single-cell RNA sequencing gene differential expression analysis (47). To identify genes differentially expressed between groups, D^3E uses either the Cramér-von Mises test, the Kolmogorov–Smirnov test or the likelihood ratio test. To test the hypothesis of the driving mechanism in apparent changes, D^3E fits a transcriptional burst model to the expression data for each gene through a method of moments or a Bayesian approach. Following the transcriptional burst model, parameter changes between groups will be calculated.

EMDomics

EMDomics detects significantly differentially expressed genes between groups for single-cell RNA sequencing data by comparing the two distribution functions of gene expressions between groups (48). EMDomics compares the differences between groups based on EMD, a commonly used approach to compare two histograms in imaging analysis. EMDomics measures the differences between two normalized distributions of the two groups through normalized total cost of transforming distributions between groups. Permutation test is used to compute the P-values for the EMD tests.

Monocle2

Using the census algorithm, Monocle2 converts the relative single-cell RNA sequencing expression levels into relative counts for each gene without experimental spike-in controls (45, 51). The census algorithm in Monocle2 estimates

the total number of mRNAs in each cell by calculating the ratio of the total number of single-mRNA genes to the fraction of the library contributed by them and then rescales the transcript per million (TPM) in single cell values into mRNA counts for each gene. Monocle2 tests gene differential expression between groups through a likelihood ratio test for comparing a full generalized linear model with additional effects to a reduced generalized linear model based on negative binomial distributions.

Linnorm

Linnorm proposes a new normalization and transformation method for single-cell RNA sequencing data analysis (49). The normalization and transformation parameters are calculated based on stably expressed genes across different cells. Linnorm uses the moderated *t*-statistics in the limma package for gene differential expression analysis through the empirical Bayes approach to centralize the estimated variances from the data.

Discriminative learning

Discriminative learning uses the multiple logistic regression framework (50). Different from previous single-cell RNA sequencing differential analysis methods, discriminative learning uses the group labels as the outcome variables and uses the gene expression levels and other characteristics of the samples as the predictor variables to identify genes significantly associated with the group labels through likelihood ratio tests.

STATISTICAL METHODS COMPARISON FOR SINGLE-CELL RNA SEQUENCING DIFFERENTIAL ANALYSIS

A previous comparison of six methods (SCDE, MAST, D^3E, Monocle, edgeR, DESeq) for single-cell RNA sequencing differential analysis examined the performance of those methods under different unimodal or bimodal distributions (52). The comparison found significant differences among those methods regarding precision, recall, empirical power, and overall performance. The investigation did not suggest an optimal method that performs better than other methods under all scenarios. A call for new differential analysis methods for single-cell RNA sequencing data was suggested as a result from the comparisons.

Another evaluation of 36 approaches for gene differential expression analysis in single-cell RNA sequencing data found remarkable differences in the performance of those approaches (53). They also found the gene differential expression analysis methods developed specifically for single-cell RNA sequencing data did not perform generally better than the methods developed for bulk RNA sequencing data.

A recent comprehensive evaluation of single-cell RNA sequencing differential analysis methods compared 11 differential analysis methods, including SCDE, MAST, scDD, DEsingle, SigEMD, SINCERA, D^3E, EMDomics, Monocle2, edgeR,

and DESeq2 (54). The gene expression values from real single-cell RNA sequencing experiments are multimodal with excess zeros, which makes the gene expression differential analysis challenging. Currently, there is no method available that can handle both multimodality and excess zeros. The comparison showed that no single method performs uniformly better than other methods under all circumstances. In general, non-parametric methods that could handle multimodality perform better than methods modeling excess zeros, while methods modeling excess zeros resulted in higher true positive rates and lower false positive rates. Gene differential expression analysis methods developed for single-cell RNA sequencing data had similar performance as those methods developed for bulk RNA sequencing data. In addition, low agreement was found among the selected genes from those differential analysis methods for single-cell RNA sequencing data. This recent evaluation also indicates the need of new differential analysis methods for single-cell RNA sequencing data.

CONCLUSION

As RNA sequencing technology is getting increasingly popular and more advanced in the biomedical and biological fields, coupled with a decrease of the cost for RNA sequencing experiments, more RNA sequencing differential analysis methods will be developed to identify differentially expressed genes between groups. For gene differential analysis methods used for both bulk RNA sequencing and single-cell RNA sequencing data, there is no consensus on an optimal method under all circumstances, although DESeq2 is currently very popular for gene expression differential analysis for bulk RNA sequencing data within the bioinformatics community. Remarkable differences were also found among different gene expression differential analysis methods in terms of numbers and characteristics of selected differentially expressed genes. Gene expression differential analysis methods specific for single-cell RNA sequencing data have a similar performance as methods developed for bulk RNA sequencing data, when both were used for single-cell RNA sequencing data. Evaluations of commonly used gene expression differential analysis methods for RNA sequencing data indicate a need for better differential analysis methods, especially for single-cell RNA sequencing data. Taking consensus of the selected differentially expressed genes from multiple methods could improve accuracy and reduce the false discovery rate, but it could also increase the false negative rate. New methods that integrate multiple approaches with both reduced false positives and reduced false negatives might be the direction for future differential analysis method development.

Acknowledgement: This work was supported by the National Cancer Institute of the National Institutes of Health (NIH) and the Food and Drug Administration (FDA) Center for Tobacco Products under Award Number U54CA228110. Dr. Li's time is supported in part by the University of Rochester CTSA award number UL1 TR002001 from the National Center for Advancing Translational Sciences of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the Food and Drug Administration (FDA).

Conflict of Interest: The author declares no potential conflict of interest with respect to research, authorship, and publication of this chapter.

Copyright and permission statement: To the best of my knowledge, the materials included in this chapter do not violate copyright laws. All original sources have been appropriately acknowledged and referenced. Where relevant, appropriate permissions have been obtained from the original copyright holder(s).

REFERENCES

1. Buzdin A, Sorokin M, Garazha A, Glusker A, Aleshin A, Poddubskaya E, et al. RNA sequencing for research and diagnostics in clinical oncology. Semin Cancer Biol. 2019. http://dx.doi.org/10.1016/j.semcancer.2019.07.010
2. Wang T, Johnson TS, Shao W, Lu Z, Helm BR, Zhang J, et al. BERMUDA: A novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. Genome Biol. 2019;20(1):165. http://dx.doi.org/10.1186/s13059-019-1764-6
3. Tian L, Dong X, Freytag S, Le Cao KA, Su S, JalalAbadi A, et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat Methods. 2019;16(6):479–87. http://dx.doi.org/10.1038/s41592-019-0425-8
4. Ferrall-Fairbanks MC, Ball M, Padron E, Altrock PM. Leveraging single-cell RNA sequencing experiments to model intratumor heterogeneity. JCO Clin Cancer Inform. 2019;3:1–10. http://dx.doi.org/10.1200/CCI.18.00074
5. Wang L, Felts SJ, Van Keulen VP, Pease LR, Zhang Y. Exploring the effect of library preparation on RNA sequencing experiments. Genomics. 2018. http://dx.doi.org/10.1016/j.ygeno.2018.11.030
6. Parker BJ. Statistical methods for transcriptome-wide analysis of RNA methylation by bisulfite sequencing. Methods Mol Biol. 2017;1562:155–67. http://dx.doi.org/10.1007/978-1-4939-6807-7_11
7. Petegrosso R, Li Z, Kuang R. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. Brief Bioinform. 2019. http://dx.doi.org/10.1093/bib/bbz063
8. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019;14(2):482–517. http://dx.doi.org/10.1038/s41596-018-0103-9
9. Siavoshi A, Taghizadeh M, Dookhe E, Piran M. Gene expression profiles and pathway enrichment analysis to identification of differentially expressed gene and signaling pathways in epithelial ovarian cancer based on high-throughput RNA-seq data. bioRxiv. 2019:566331. http://dx.doi.org/10.1101/566331
10. Costa V, Aprile M, Esposito R, Ciccodicola A. RNA-Seq and human complex diseases: Recent accomplishments and future perspectives. Eur J Hum Genet. 2013;21(2):134–42. http://dx.doi.org/10.1038/ejhg.2012.129
11. Akond Z, Alam M, Mollah MNH. Biomarker identification from RNA-seq data using a robust statistical approach. Bioinformation. 2018;14(4):153–63. http://dx.doi.org/10.6026/97320630014153
12. Xiong H, Guo H, Xie Y, Zhao L, Gu J, Zhao S, et al. RNAseq analysis reveals pathways and candidate genes associated with salinity tolerance in a spaceflight-induced wheat mutant. Sci Rep. 2017;7(1):2731. http://dx.doi.org/10.1038/s41598-017-03024-0
13. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50(8):96. http://dx.doi.org/10.1038/s12276-018-0071-8
14. Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. Elife. 2019;8. http://dx.doi.org/10.7554/eLife.43803

15. Haque A, Engel J, Teichmann SA, Lonnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9(1):75. http://dx.doi.org/10.1186/s13073-017-0467-4
16. Hedlund E, Deng Q. Single-cell RNA sequencing: Technical advancements and biological applications. Mol Aspects Med. 2018;59:36–46. http://dx.doi.org/10.1016/j.mam.2017.07.003
17. Pizzolato G, Kaminski H, Tosolini M, Franchini DM, Pont F, Martins F, et al. Single-cell RNA sequencing unveils the shared and the distinct cytotoxic hallmarks of human TCRVdelta1 and TCRVdelta2 gammadelta T lymphocytes. Proc Natl Acad Sci U S A. 2019;116(24):11906–15. http://dx.doi.org/10.1073/pnas.1818488116
18. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced applications of RNA sequencing and challenges. Bioinform Biol Insights. 2015;9(Suppl 1):29–46. http://dx.doi.org/10.4137/BBI.S28991
19. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. Front Genet. 2019;10:317. http://dx.doi.org/10.3389/fgene.2019.00317
20. Williams AG, Thomas S, Wyman SK, Holloway AK. RNA-seq data: Challenges in and recommendations for experimental design and analysis. Curr Protoc Hum Genet. 2014;83:11.3.1–20. http://dx.doi.org/10.1002/0471142905.hg1113s83
21. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis. Brief Funct Genom. 2015;14(2):130–42. http://dx.doi.org/10.1093/bfgp/elu035
22. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511–15. http://dx.doi.org/10.1038/nbt.1621
23. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31(1):46–53. http://dx.doi.org/10.1038/nbt.2450
24. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. http://dx.doi.org/10.1093/bioinformatics/btp616
25. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288–97. http://dx.doi.org/10.1093/nar/gks042
26. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106. http://dx.doi.org/10.1186/gb-2010-11-10-r106
27. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. http://dx.doi.org/10.1186/s13059-014-0550-8
28. Hardcastle TJ, Kelly KA. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11:422. http://dx.doi.org/10.1186/1471-2105-11-422
29. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, et al. EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013;29(8):1035–43. http://dx.doi.org/10.1093/bioinformatics/btt087
30. Li J, Tibshirani R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013;22(5):519–36. http://dx.doi.org/10.1177/0962280211428386
31. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: A matter of depth. Genome Res. 2011;21(12):2213–23. http://dx.doi.org/10.1101/gr.124321.111
32. Tarazona S, Furio-Tari P, Turra D, Pietro AD, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 2015;43(21):e140.
33. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29. http://dx.doi.org/10.1186/gb-2014-15-2-r29
34. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods. 2017;14(7):687–90. http://dx.doi.org/10.1038/nmeth.4324
35. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14:91. http://dx.doi.org/10.1186/1471-2105-14-91

36. Rajkumar AP, Qvist P, Lazarus R, Lescai F, Ju J, Nyegaard M, et al. Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. BMC Genom. 2015;16. http://dx.doi.org/10.1186/s12864-015-1767-y
37. Costa-Silva J, Domingues D, Lopes FM. RNA-seq differential expression analysis: An extended review and a software tool. PLoS One. 2017;12(12). http://dx.doi.org/10.1371/journal.pone.0190152
38. Lin BQ, Pang Z. Stability of methods for differential expression analysis of RNA-seq data. BMC Genom. 2019;20. http://dx.doi.org/10.1186/s12864-018-5390-6
39. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: A tutorial. Mol Syst Biol. 2019;15(6):e8746. http://dx.doi.org/10.15252/msb.20188746
40. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740–2. http://dx.doi.org/10.1038/nmeth.2967
41. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16:278. http://dx.doi.org/10.1186/s13059-015-0844-5
42. Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016;17(1):222. http://dx.doi.org/10.1186/s13059-016-1077-y
43. Miao Z, Deng K, Wang X, Zhang X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Bioinformatics. 2018;34(18):3223–4. http://dx.doi.org/10.1093/bioinformatics/bty332
44. Wang T, Nabavi S. SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data. Methods. 2018;145:25–32. http://dx.doi.org/10.1016/j.ymeth.2018.04.017
45. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381–6. http://dx.doi.org/10.1038/nbt.2859
46. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: A pipeline for single-cell RNA-seq profiling analysis. PLoS Comput Biol. 2015;11(11):e1004575. http://dx.doi.org/10.1371/journal.pcbi.1004575
47. Delmans M, Hemberg M. Discrete distributional differential expression (D3E) – A tool for gene expression analysis of single-cell RNA-seq data. BMC Bioinformatics. 2016;17:110. http://dx.doi.org/10.1186/s12859-016-0944-6
48. Nabavi S, Schmolze D, Maitituoheti M, Malladi S, Beck AH. EMDomics: A robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. Bioinformatics. 2016;32(4):533–41. http://dx.doi.org/10.1093/bioinformatics/btv634
49. Yip SH, Wang P, Kocher JA, Sham PC, Wang J. Linnorm: Improved statistical analysis for single cell RNA-seq expression data. Nucleic Acids Res. 2017;45(22):e179. http://dx.doi.org/10.1093/nar/gkx828
50. Ntranos V, Yi L, Melsted P, Pachter L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. Nat Methods. 2019;16(2):163–6. http://dx.doi.org/10.1038/s41592-018-0303-9
51. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods. 2017;14(3):309–15. http://dx.doi.org/10.1038/nmeth.4150
52. Dal Molin A, Baruzzo G, Di Camillo B. Single-cell RNA-sequencing: Assessment of differential expression analysis methods. Front Genet. 2017;8:62. http://dx.doi.org/10.3389/fgene.2017.00062
53. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018;15(4):255–61. http://dx.doi.org/10.1038/nmeth.4612
54. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics. 2019;20(1):40. http://dx.doi.org/10.1186/s12859-019-2599-6