



# RNA-SEQ ANALYSIS: METHODS, APPLICATIONS AND CHALLENGES

EDITED BY: Filippo Geraci, Indrajit Saha and Monica Bianchini  
PUBLISHED IN: Frontiers in Genetics and Frontiers in Plant Science



# frontiers

## Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence.

The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714

ISBN 978-2-88963-705-8

DOI 10.3389/978-2-88963-705-8

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: [researchtopics@frontiersin.org](mailto:researchtopics@frontiersin.org)



# RNA-SEQ ANALYSIS: METHODS, APPLICATIONS AND CHALLENGES

Topic Editors:

**Filippo Geraci**, Institute for Informatics and Telematics, CNR, Pisa, Italy

**Indrajit Saha**, Department of Computer Science and Engineering,  
National Institute of Technical Teachers Training and Research, Kolkata, India

**Monica Bianchini**, DIISM, University of Siena, Siena, Italy

**Citation:** Geraci, F., Saha, I., Bianchini, M., eds. (2020). RNA-Seq Analysis: Methods, Applications and Challenges. Lausanne: Frontiers Media SA.  
doi: 10.3389/978-2-88963-705-8

# Table of Contents

- 05 Editorial: RNA-Seq Analysis: Methods, Applications and Challenges**  
Filippo Geraci, Indrajit Saha and Monica Bianchini

## SECTION 1

### RNA-SEQ ANALYSIS

- 08 Assessment of a Highly Multiplexed RNA Sequencing Platform and Comparison to Existing High-Throughput Gene Expression Profiling Techniques**  
Eric Reed, Elizabeth Moses, Xiaohui Xiao, Gang Liu, Joshua Campbell, Catalina Perdomo and Stefano Monti
- 22 Read Mapping and Transcript Assembly: A Scalable and High-Throughput Workflow for the Processing and Analysis of Ribonucleic Acid Sequencing Data**  
Sateesh Peri, Sarah Roberts, Isabella R. Kreko, Lauren B. McHan, Alexandra Naron, Archana Ram, Rebecca L. Murphy, Eric Lyons, Brian D. Gregory, Upendra K. Devisetty and Andrew D. L. Nelson
- 31 Integrative Differential Expression Analysis for Multiple EXperiments (IDEAMEX): A Web Server Tool for Integrated RNA-Seq Data Analysis**  
Verónica Jiménez-Jacinto, Alejandro Sanchez-Flores and Leticia Vega-Alvarado
- 47 DREAMSeq: An Improved Method for Analyzing Differentially Expressed Genes in RNA-seq Data**  
Zhihua Gao, Zhiying Zhao and Wenqiang Tang
- 61 CircCode: A Powerful Tool for Identifying circRNA Coding Ability**  
Peisen Sun and Guanglin Li

## SECTION 2

### SINGLE CELL RNA SEQUENCING

- 67 Single-Cell RNA-Seq Technologies and Related Computational Data Analysis**  
Geng Chen, Baitang Ning and Tielu Shi
- 80 Benchmark and Parameter Sensitivity Analysis of Single-Cell RNA Sequencing Clustering Methods**  
Monika Krzak, Yordan Raykov, Alexis Boukouvalas, Luisa Cutillo and Claudia Angelini
- 99 Reproducibility of Methods to Detect Differentially Expressed Genes From Single-Cell RNA Sequencing**  
Tian Mou, Wenjiang Deng, Fengyun Gu, Yudi Pawitan and Trung Nghia Vu
- 111 McImpute: Matrix Completion Based Imputation for Single Cell RNA-seq Data**  
Aanchal Mongia, Debarka Sengupta and Angshul Majumdar

## SECTION 3

### CASE STUDIES

- 123** *Expression Profile Analysis Identifies a Novel Five-Gene Signature to Improve Prognosis Prediction of Glioblastoma*  
Wen Yin, Guihua Tang, Quanwei Zhou, Yudong Cao, Haixia Li, Xianyong Fu, Zhaoping Wu and Xingjun Jiang
- 135** *Co-expression Network Analysis Identifies Four Hub Genes Associated With Prognosis in Soft Tissue Sarcoma*  
Zenhua Zhu, Zheng Jin, Yuyou Deng, Lai Wei, Xiaowei Yuan, Mei Zhang and Dahui Sun
- 145** *Long Noncoding RNA RAET1K Enhances CCNE1 Expression and Cell Cycle Arrest of Lung Adenocarcinoma Cell by Sponging miRNA-135a-5p*  
Chang Zheng, Xuelian Li, Yangwu Ren, Zhihua Yin and Baosen Zhou
- 156** *Analysis of Key Genes Involved in Potato Anthocyanin Biosynthesis Based on Genomics and Transcriptomics Data*  
Nie Tengkun, Wang Dongdong, Ma Xiaohui, Chen Yue and Chen Qin



# Editorial: RNA-Seq Analysis: Methods, Applications and Challenges

Filippo Geraci<sup>1\*</sup>, Indrajit Saha<sup>2</sup> and Monica Bianchini<sup>3\*</sup>

<sup>1</sup> Institute for Informatics and Telematics, CNR, Pisa, Italy, <sup>2</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, India, <sup>3</sup> DIISM, University of Siena, Siena, Italy

**Keywords:** RNA-seq, algorithm, software pipeline, method assessment, differential analysis

## Editorial on the Research Topic

### RNA-Seq Analysis: Methods, Applications and Challenges

## 1. INTRODUCTION

RNA-seq has revolutionized the research community approach to studying gene expression. In fact, this technology has opened up the possibility of quantifying the expression level of all genes at once, allowing an ex post (rather than ex ante) selection of candidates that could be interesting for a certain study. The continuous drop in costs and the independence of library preparation protocols from the model species, have convinced the stakeholders to invest in this technology, by creating consortia able to produce large disease-specific datasets that, in turn, fostered transcriptomic research at a population level. Among many others, a virtuous example in this sense is The Cancer Genome Atlas. In a short time RNA-seq has moved from a technology to merely quantify the expression of genes to a powerful tool to: discover new transcripts (via *de novo* transcriptome assembly), characterize alternative splicing variants or new cell types (through single cell RNA sequencing). Leveraging on RNA-seq for daily diagnostic activities is no longer a dream but a consolidated reality.

Although established best practices exist, managing RNA-seq data is not easy. Before sequencing, it is essential to carefully plan library preparation in order to minimize downstream analysis biases. Budget optimization is another important factor. Sequencing multiple samples increases statistical power and reduces undesired side effects due to noise and variability. However, more samples imply higher costs. Multiplexing has proved to be an effective tool to limit the budget without sacrificing the number of samples. DNA barcoding enables combining up to 96 samples into a single line, trading a lower sequencing depth for a higher number of sequenced samples. The downside of this technique is the increased burden of data analysis to achieve the same accuracy that would be achieved with a richer input.

Downstream sequencing, fastq data must be validated and processed to distill raw reads into a quantitative measure of gene expression. While validation is somehow a standard procedure, read count depends on the type of RNA (microRNA, etc.) and on the target application. Usually reads are: subjected to adapter removal, aligned against a reference genome, grouped by functional unit (e.g., transcripts, genes, microRNA, etc.), normalized and counted. Subsequent analyses can vary dramatically according to the application. In the simplest setting, the subset of genes responsible for the differences on the phenotype between two populations should be discovered. In other cases, one may want to build the co-expression (or reverse expression) network in order to find interacting genes or a pathway related to a certain phenotype. Other applications involve the discovery of unknown cell types, the organization of cell types in homogeneous families, the identification of

## OPEN ACCESS

### Edited and reviewed by:

Richard D. Emes,  
University of Nottingham,  
United Kingdom

### \*Correspondence:

Filippo Geraci  
filippo.geraci@iit.cnr.it  
Monica Bianchini  
monica@diism.unisi.it

### Specialty section:

This article was submitted to  
Bioinformatics and Computational  
Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 03 February 2020

**Accepted:** 24 February 2020

**Published:** 17 March 2020

### Citation:

Geraci F, Saha I and Bianchini M  
(2020) Editorial: RNA-Seq Analysis:  
Methods, Applications and  
Challenges. *Front. Genet.* 11:220.  
doi: 10.3389/fgene.2020.00220

new molecules (e.g., new microRNA, long non-coding RNA, etc.), or the annotation of new variants or alternative splicing.

## 2. RESEARCH TOPIC ORGANIZATION

This Research Topic is divided into three main sections: five articles cover the RNA-seq workflow, four papers discuss the most recent frontier of single cell RNA sequencing, while the last four contributions report on case studies, related to tumor profiling and plant science.

In the first part, we attempted to analyze the RNA-seq process (from experimental design to analysis and extraction of new knowledge) by highlighting the key choices of the state-of-the-art workflows. Although we have mainly focused on computational aspects, we believe that this Research Topic can catch the interest of those readers, specialized in the field of life science, who intend to become independent and autonomous in the analysis of their own data. Two papers of this section describe new methods: for the identification of differentially expressed genes and for the prediction of the circRNA coding ability.

The second section introduces a recent branch of RNA-seq data analysis: single cell sequencing (scRNA-seq). Although conceptually similar to sequencing cells in bulk, the single cell resolution of this technique introduces a lot of noise, that requires *ad hoc* analysis methods. Much of this section is dedicated to the introduction of basic single cell RNA sequencing concepts, from laboratory protocols to the most common analyses. In particular, the problems of assessing the results of clustering cell types and the reproducibility of differential expression experiments are discussed. Finally, this section concludes with the description of a new method to infer missing counts due to poor coverage of sequencing.

The last part of the Research Topic was dedicated to four case studies: three concerning tumors and one application in plant science. The rationale behind this choice was that of showing different types of analysis. In the conceptually simpler case, the goal of the analysis was to create a panel of genes prognostic of the onset of cancer. Next, an example of a co-expression network is shown. Finally, an example of interaction among different types of RNA (long non-coding, genes, microRNAs) has been reported, showing the complexity of the pathways that regulate the life of cells.

### 2.1. RNA-Seq Analysis

In Reed et al., the opportunity offered by Multiplexed RNA Sequencing is discussed. The study provides a comparison of several methods using real data from immortalized human lung epithelial cells.

In Peri et al., RMTA, an user-friendly analysis workflow, is proposed. RMTA was designed to provide standard pre-processing tools (i.e., read quality analysis, filters for lowly expressed transcripts, and read counting for differential expression analysis) in a scalable and easy to deploy environment.

In Jimenez-Jacinto et al., an integrative differential expression analysis web server (IDEAMEX) is described. The rationale

of IDEAMEX is that of freeing non-expert users from the (sometimes frustrating) experience of interacting with the UNIX-based environment for standard differential expression analyses.

In Gao et al., a new method for the identification of differentially expressed genes is reported. The key observation of this work is that the binomial distribution at the basis of the majority of the algorithms for differential expression analysis is unable to capture underdispersion characteristics of RNA-seq data.

In Sun and Li, the problem of predicting whether a given circular RNA can be translated or not is investigated. Circular RNAs differ from other types of RNA in that they are arranged as rings joining 3' and 5' endpoints. This characteristic makes hard to decide about their translation potential. The manuscript provides an algorithm to identify the coding ability of circRNAs with high sensitivity.

### 2.2. Single Cell RNA Sequencing

In Chen et al., an overview of currently available single-cell isolation protocols and scRNA-seq technologies is provided. In addition, several methods for scRNA-seq data analysis, from quality control to network reconstruction, are discussed.

In Krzak et al., the use of clustering to study heterogeneity of cells is dissected. In particular, this work aims at providing new insights into the advantages and drawbacks of scRNAseq clustering, highlighting open challenges.

In Mou et al., some issues connected to the reproducibility of differential expression studies is debated. The complexity of this type of analyses stands in the paucity of RNAs and in the consequent lower signal to noise ratio. The article shows pros and cons of standard and *ad-hoc* software for differential expression.

In Mongia et al., a method to impute dropouts in single cell expression data is detailed. Experiments on real data show that the proposed software is able to discriminate the real absence of reads from dropout events.

### 2.3. Case Studies

In Yin et al., differential expression analysis is used to pinpoint a small panel of genes potentially prognostic for the onset of Glioblastoma. The focus of the article is that of improving healthy/diseased classification regardless of the interaction among genes.

In Zhu et al., co-expressed genes are identified in order to build a network of interactions. Subsequently, the network is analyzed to select hub genes associated with soft tissue sarcomas.

In Zheng et al., the dynamics of the interaction among different molecules in lung adenocarcinoma is studied. The article reports on how the dysregulation of a long non-coding RNA triggers a sequence of dysregulations, causing the cell cycle arrest.

In Tengku et al., genomics and transcriptomics data are integrated in order to identify the crucial genes that affect anthocyanin biosynthesis transforming quantitative traits into quality traits.



## AUTHOR CONTRIBUTIONS

The authors all contributed equally to the Research Topic assembly and editing and to this editorial.

## FUNDING

IS was supported by a grant (DST/INT/POL/P-36/2016) from the Department of Science and Technology, India.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2020 Geraci, Saha and Bianchini. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*



# Assessment of a Highly Multiplexed RNA Sequencing Platform and Comparison to Existing High-Throughput Gene Expression Profiling Techniques

Eric Reed<sup>1,2</sup>, Elizabeth Moses<sup>2</sup>, Xiaohui Xiao<sup>2</sup>, Gang Liu<sup>2</sup>, Joshua Campbell<sup>1,2</sup>, Catalina Perdomo<sup>2</sup> and Stefano Monti<sup>1,2\*</sup>

<sup>1</sup> Bioinformatics Program, Boston University, Boston, MA, United States, <sup>2</sup> Section of Computational Biomedicine, School of Medicine, Boston University, Boston, MA, United States

## OPEN ACCESS

### Edited by:

Filippo Geraci,  
National Research Council (CNR), Italy

### Reviewed by:

Kashmir Singh,  
Panjab University, India  
Matteo Benelli,  
University of Trento, Italy  
Haibo Liu,  
Iowa State University, United States

### \*Correspondence:

Stefano Monti  
smonti@bu.edu

### Specialty section:

This article was submitted to  
Bioinformatics and Computational  
Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 06 September 2018

**Accepted:** 12 February 2019

**Published:** 05 March 2019

### Citation:

Reed E, Moses E, Xiao X, Liu G,  
Campbell J, Perdomo C and Monti S  
(2019) Assessment of a Highly  
Multiplexed RNA Sequencing  
Platform and Comparison to Existing  
High-Throughput Gene Expression  
Profiling Techniques.  
Front. Genet. 10:150.  
doi: 10.3389/fgene.2019.00150

The need to reduce per sample cost of RNA-seq profiling for scalable data generation has led to the emergence of highly multiplexed RNA-seq. These technologies utilize barcoding of cDNA sequences in order to combine multiple samples into a single sequencing lane to be separated during data processing. In this study, we report the performance of one such technique denoted as sparse full length sequencing (SFL), a ribosomal RNA depletion-based RNA sequencing approach that allows for the simultaneous sequencing of 96 samples and higher. We offer comparisons to well established single-sample techniques, including: full coverage Poly-A capture RNA-seq, microarrays, as well as another low-cost highly multiplexed technique known as 3' digital gene expression (3'DGE). Data was generated for a set of exposure experiments on immortalized human lung epithelial (AALe) cells in a two-by-two study design, in which samples received both genetic and chemical perturbations of known oncogenes/tumor suppressors and lung carcinogens. SFL demonstrated improved performance over 3'DGE in terms of coverage, power to detect differential gene expression, and biological recapitulation of patterns of differential gene expression from *in vivo* lung cancer mutation signatures.

**Keywords:** RNA sequencing, gene expression, microarray, multiplexing, platform comparison

## INTRODUCTION

Since its inception in 2008, RNA sequencing has become the gold-standard for whole-transcriptome high-throughput data generation (Mortazavi et al., 2008). In addition to RNA transcript expression quantification, RNA-seq allows for more advanced analyses including *de novo* transcriptome assembly (Robertson et al., 2010) and characterization of alternative splicing variants (Bryant et al., 2012). Furthermore, RNA-seq is species agnostic, such that the same library preparation technique may be utilized for humans, mouse, rat, kidney bean, etc. These represent clear advantages over hybridization-based microarray platforms in which individual microarray platforms are designed to quantify specific transcripts for a specific species (Wang et al., 2009). However, one persistent drawback of RNA-seq has been its relatively high cost.

The use of classic RNA-seq techniques for experimental designs that require profiling of many samples – especially when the marginal information value of each sample is relatively low, such as in medium- and high-throughput screening applications – can thus present a disqualifying cost burden.

Large-scale projects based on transcriptional profiling of chemical exposure experiments include the Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (Open TG-GATEs) (Igarashi et al., 2015), the DrugMatrix database (Ganter et al., 2006), and the Connectivity Map (CMap) (Subramanian et al., 2017), among others. Both the TG-GATEs and the DrugMatrix projects used microarrays for expression profiling, which was at the time significantly less costly than full coverage RNA-sequencing, yet still requiring multi-million budgets. Alternatively, the CMap project utilizes the Luminex-1000 (L1000) profiling platform, a bead-based analog expression assay which quantifies 1,058 human transcripts, which are used to impute the expression of 11,350 additional transcripts (Subramanian et al., 2017). This technique is among the least expensive expression assays available, but it is restricted to human screens and it directly profiles only a limited panel of genes. Given the flexibility of RNA-sequencing platforms, highly multiplexed techniques represent a viable alternative for generating transcriptional data from exposure screens, as well as from other experiments that require a large sample size. Therefore, evaluation of the technical validity of specific techniques serves to inform research strategies for a variety of biological inquiries.

The need to reduce the per sample cost of RNA-seq has led to the adoption of barcoding technologies, where cDNA sequences from individual samples are tagged and their libraries are combined and multiplex sequenced in a single lane (Wang et al., 2011). More recently, these techniques have been optimized to allow multiplex sequencing of 96 samples per lane or higher (Hou et al., 2015; Shishkin et al., 2015). Here, we report the results of our effort at optimizing and evaluating one such technique denoted as sparse full length (SFL) sequencing (Shishkin et al., 2015), a ribosomal RNA depletion-based RNA sequencing approach that allows for the simultaneous sequencing of 96 samples and higher. We offer comparisons to well established single-sample techniques, including: full coverage Poly-A capture RNA-seq and microarray, as well as another low-cost highly multiplexed technique known as 3' digital gene expression (3'DGE) (Asmann et al., 2009). Assessments include comparisons of coverage between the three RNA-sequencing techniques, as well as signal-to-noise and biological recapitulation of gene-level differential signals between treatment groups for the same samples profiled across SFL, microarray, and 3'DGE. For this evaluation study, we generated a set of exposure experiments on immortalized human lung epithelial (AALE) cells (Lundberg et al., 2002) in a two-by-two study design, in which samples received both genetic and chemical perturbations of known oncogenes/tumor suppressors and lung carcinogens (Figure 1). The goal of this report is not only to assess the performance of our optimized highly multiplexed technique, but to inform future research in terms of the strengths and

pitfalls of available cost-effective high throughput transcriptomic profiling techniques.

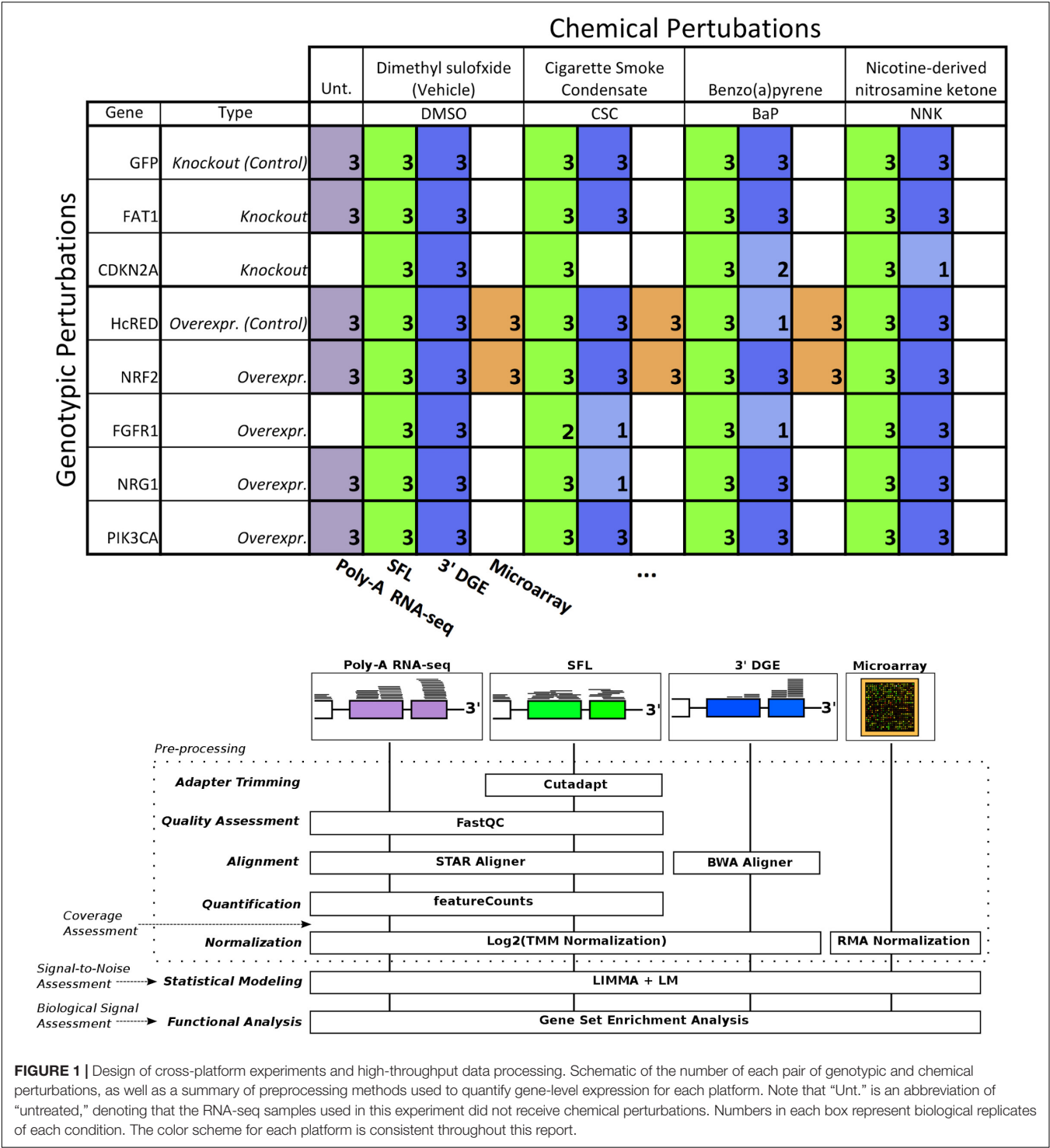
## MATERIALS AND METHODS

### Samples

Exposure experiments were performed on immortalized human bronchial epithelial cells (AALE). Cells were exposed to both chemical and genotypic perturbations with three replicates per perturbation combination. Cells were thawed from liquid nitrogen and grown up in SAGM small airway epithelial cell growth media (Lonza, Portsmouth NH). Cells were subcultured using Clonetics ReagentPack subculture reagents (Lonza, Portsmouth NH). In preparation for exposure, cells were plated into 24-well plates and allowed to reach confluency for 24 h. Cell culture media was then replaced, and compounds added at a concentration of 24 µg/ml CSC, 173 µM BaP, 490 µM NNK or DMSO. NNK and BaP compounds were obtained from Sigma-Aldrich (St. Louis, MO, United States) and CSC obtained from Murty Pharmaceuticals (Lexington, KY, United States). Genotypic perturbations included CRISPR knockouts of *FAT1*, and *CDKN2A*, as well as overexpression of *NRF2* (*NFE2L2*), *FGFR1*, *NRG1*, and *PIK3CA*. Cells transfected with a pSpCas9-EGFP (*GFP*) plasmid (PX458) in the absence of sgRNAs were used as controls for the CRISPR perturbations while overexpression of an empty vector containing the reporter HcRed served as control for the overexpression experiments. The same samples were profiled across SFL, microarray, and 3'DGE for a subset of combinations of exposures, though all samples were profiled by SFL. In addition, full coverage poly-A RNA-seq was performed on a separate set of samples for a subset of genotypic exposures, including CRISPR knockouts of *FAT1*, as well as overexpression of *NRF2*, *NRG1*, and *PIK3CA*. These samples did not receive any chemical exposures (Figure 1). Note that in a few cases there was not enough material to perform 3'DGE, as indicated by the sample numbers of certain perturbation combinations.

### Library Preparation

Library preparation for SFL sequencing was carried out based on the published protocol (Shishkin et al., 2015). An edited version of this protocol is available in the **Supplementary Material**. RNA was isolated using a standard Qiazol and Qiacube protocol from Qiagen (Valencia, CA, United States). RNA purity was assessed using a NanoDrop spectrophotometer and no samples were excluded from downstream analysis. The dual-barcoded SFL libraries were pooled from 96 individual samples and then sequenced on the Illumina® NextSeq 550 to generate more than 400 million single-end 75-bp reads. Poly-A RNA Sequencing libraries were prepared from total RNA samples using Illumina® TruSeq® RNA Sample Preparation Kit v2 and then sequenced on the Illumina® HiSeq 2500 to generate more than 5 million single-end 50-bp reads per sample. Microarray procedures were performed as described in GeneChip™ WT PLUS Reagent Kit manual and GeneChip™ WT Terminal



## Data Pre-processing

Affymetrix GeneChip Human Gene 2.0 ST Microarray CEL files were annotated to unique Entrez gene IDs, using a custom CDF file from BrainArray (hugene20st\_Hs\_ENTREZG\_21.0.0) and RMA-normalized. For SFL, adapter sequences were trimmed from raw sequence files using *Cutadapt v1.12*. Quality assessment of trimmed SFL sequence files as well as raw full coverage RNA-seq sequencing files was performed with *FastQC v0.11.5*. Both SFL and RNA-seq reads were aligned to human genome (UCSC RefSeq hg19) with *STAR v2.5.2b* with the non-default parameter, “*-outSAMtype BAM SortedByCoordinate*” (Dobin et al., 2013). Expression quantification in RefSeq genes was carried out with *featureCounts (subread) v1.5.0* (Liao et al., 2014). For 3'DGE, pre-quantified gene expression count matrices were obtained from the Broad Institute, Cambridge, MA, United States. These reads had been aligned to the transcriptome (UCSC RefSeq hg19), using *BWA aln v0.7.10* with the non-default parameter, “*-l 24*” (Li and Durbin, 2009). Considering that there are  $4^{10}$  ( $\sim 1.05 \times 10^6$ ) possible UMIs and the 3'DGE library sizes are on the order of  $10^6$  reads, it is highly unlikely for the same UMI to be added to multiple cDNA fragments from the same gene. Therefore, using a custom python program (Soumillon et al., 2014), reads with the same UMI and sample barcode were only counted once per gene. All further data processing and analysis were carried out in R.

## Coverage Assessment

Read coverage across the 82 samples, shared between SFL and 3'DGE, as well as all 18 full coverage RNA-seq samples was assessed for library size as well as percentage of the library size that was aligned, uniquely aligned (i.e., reads that only align once in the genome), and counted in the 22,233 genes which were annotated across all three platforms, i.e., the intersection of annotated genes. The full set of counted reads is hereafter referred to as the counted library. Unlike SFL and full coverage RNA-seq, 3'DGE reads are aligned directly to mRNA sequences, such that the reported numbers of counted reads and uniquely aligned reads are the same. To assess the relative distribution of reads across the total set of shared genes, we plotted the cumulative proportion of the sum of reads aligning to individual genes per samples ranked by relative expression across all three platforms. Saturation analysis of the estimated minimum percentage of the counted library size to maximize the number of genes quantified by each platform was performed using a loess fit the gene discovery of 20 subsamplings of the per sample counted libraries. All subsampling analysis was performed using *Subseq v1.8.0*.

Finally, we assessed the relative induction of noise introduced by subsampling progressively larger proportions of the original counted library sizes in each platform, as measured by the principal component error (Heimberg et al., 2016). In order to compare the three platforms assuming equally sized starting library, we repeated the assessment after first subsampling full coverage RNA-seq libraries and 3'DGE libraries to sizes matching that of SFL, the smallest library of the three platforms. This analysis was performed on the 18 samples of like genotypic

perturbations, with no chemical treatment in the case of full coverage RNA-seq samples and vehicle DMSO treatment in SFL and 3'DGE samples. Reported values reflect means across 20 iterations of the subsampling and principal component error calculation procedure.

## Signal-to-Noise Assessment

Signal-to-noise was compared among SFL, 3'DGE and microarrays based on four-group ANOVA analysis and two-group differential analysis. In order to estimate signal-to-noise as a means for assessing expected performance when applying standard statistical methods to the data, rather than differential gene expression analysis packages, classic ANOVA was performed for each gene using normalized data across all three platforms, using the *glm* function in R. In this analysis, the signal-to-noise was assessed across like samples undergoing exposure to CSC or DMSO vehicle, as well as genotypic perturbations of *NRF2* overexpression or HcRed control. Thus, the analysis included four independent groups of samples, receiving each combination of chemical (CSC or DMSO) and genotypic (*NRF2* or HcRed) perturbations, with three replicates in each group. Only genes with mean expression  $\geq 1$  across all 12 samples in both SFL and 3'DGE were included in the analysis (9,813 total genes). Expression levels across SFL and 3'DGE were normalized via trimmed mean of M values (TMM) (Robinson and Oshlack, 2010) scaling and  $\log_2$  counts-per-million transformation. Additionally, two-group differential gene expression analysis was performed for each stratified chemical and genotypic perturbation, using *LIMMA v3.30.7*. That is, differential expression of CSC- vs. DMSO-treated samples, within either HcRed or *NRF2* treatment, as well as differential expression of *NRF2*- vs. HcRed-treated samples, within either DMSO or CSC exposure, was performed. The SFL and 3'DGE count data were transformed for linear modeling based on *voom* (Ritchie et al., 2015). Following modeling, results were restricted to the top 10,000 genes as ranked by median-absolute-deviation (MAD). This heuristic gene filtering procedure was adopted because quantification-based filtering is not applicable to microarray data. This approach follows recommendations detailed in the *LIMMA* manual (Ritchie et al., 2015). All *p*-values reported from two-group differential analysis are two-sided. In both ANOVA and *LIMMA* analyses, nominal *p*-values for each gene were corrected for multiple comparisons using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995).

## Biological Signal Recapitulation

Two-group differential analysis signatures were compared by pre-ranked gene set enrichment analysis (GSEA) to gene sets derived from published signatures of smoking exposure in the airway from healthy volunteers (Spira et al., 2004; Beane et al., 2007), as well as to gene sets analytically derived from The Cancer Genome Atlas (TCGA) for patients with lung squamous cell carcinoma (LUSC) or lung adenocarcinoma (LUAD). The two smoking gene sets consist of genes reported as either up- or down-regulated in response to smoking in at least one of the two publications, while TCGA gene sets were



derived by probing differential expression of individual genes between patients with or without point mutations or copy number alterations (CNA) in genes of interest. These include mutations for the same panel of genes profiled for genotypic perturbations. In addition we include *KEAP1* mutations, a repressor of *NRF2* (Kansanen et al., 2013, 1). Specifically, point mutation signatures were derived from LUSC and LUAD, independently, by performing differential analysis of subjects with and without point mutations in genes of interest, matched for age, sex, and cancer stage. For *NRF2* and *PIK3CA* point mutations were defined at specific mutation hotspots of along the gene body (**Supplementary Figure S2**) (Campbell et al., 2016). Likewise, CNA gene signatures were assessed for amplification and deletions of genes of interest by differential analysis, using subjects with zero, one, or two additional copies or deletions of a gene of interest, respectively. All models for mutations and CNA were adjusted for tumor purity, as reported (Campbell et al., 2016). Differential signatures were derived using *LIMMA*. Genes associated with specific mutations or CNA were defined as those with significance and magnitude of the linear model's genetic alteration coefficient at FDR  $Q$ -value  $< 0.05$  and  $|\log_2 \text{fold-change}| > \log_2(1.5)$ , respectively.

Each of our genotypic perturbation signatures was compared by GSEA to the corresponding TCGA-derived gene sets. For example, the *PIK3CA* overexpression signatures were compared to the gene sets derived from *PIK3CA* mutation and CNA in the TCGA data. To assess the effect of read counts on gene discovery and biological recapitulation of each platform, we compared the differential analysis and GSEA results to that derived from subsampled libraries across full coverage RNA-seq, SFL, and 3'DGE. Similar to coverage assessment, this analysis was performed starting with full libraries across all three platforms, as well as initially subsampling the full coverage RNA-seq and 3'DGE libraries to sizes matching that of SFL. Reported values reflect means from 20

iterations of the subsampling followed by differential analysis and GSEA procedures.

## RESULTS

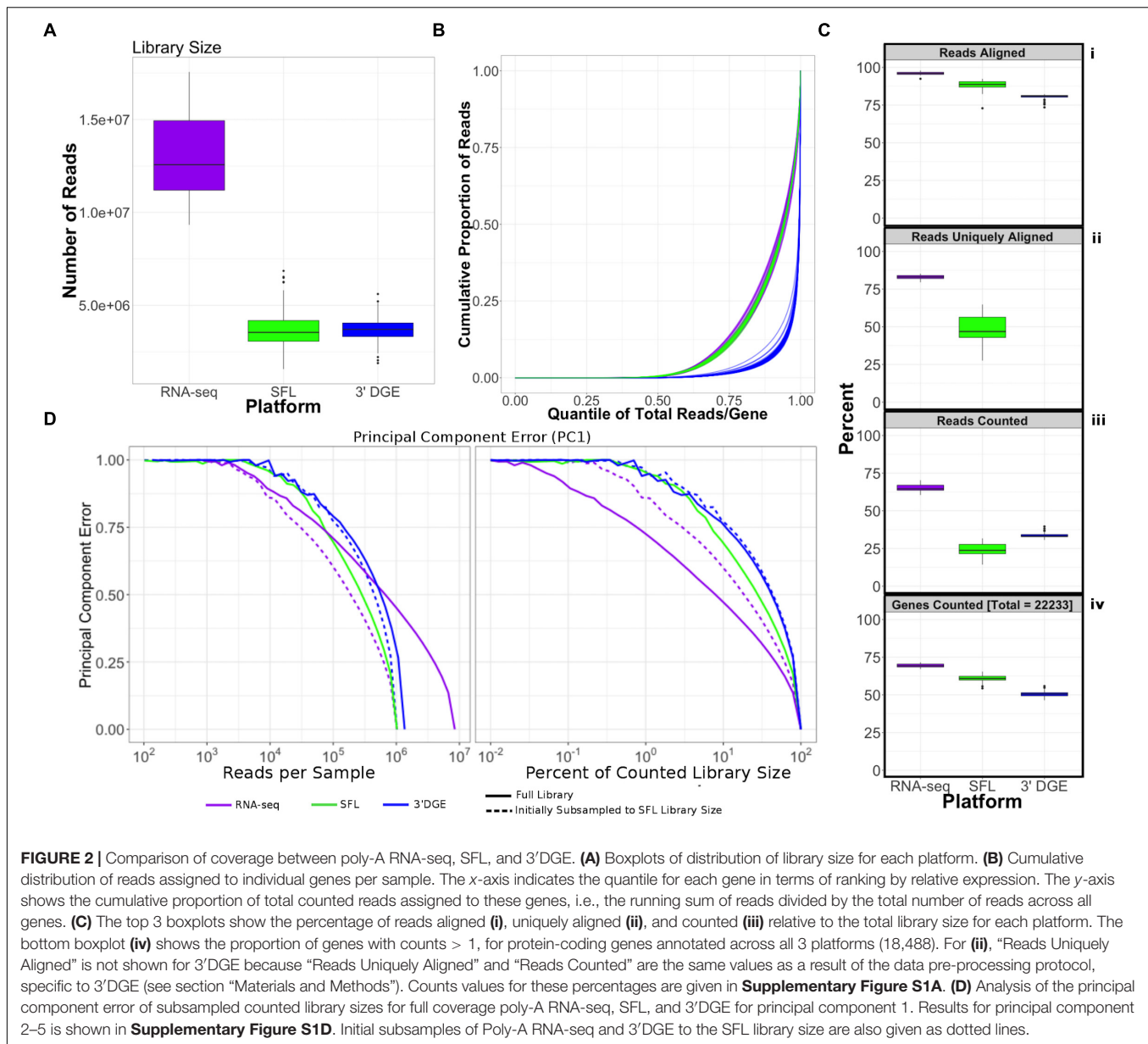
### Coverage Assessment

Comparison of coverage of the three sequencing platforms, full coverage poly-A RNA-seq, SFL, and 3'DGE, is summarized in **Table 1**, **Figure 2**, and **Supplementary Figure S1**. Comparison between SFL and 3'DGE included 82 samples each, while full coverage poly-A RNA-seq included all 18 available samples. None of the three platforms demonstrated differences in the library size variability (total number of assigned reads) across samples, although there was a notably high difference between the largest and smallest library size for the SFL samples, with a fold change of 4.3. Fold changes for full coverage RNA-seq and 3'DGE were 1.9 and 2.9, respectively (**Table 1** and **Figure 2A**).

Unsurprisingly, full coverage poly-A RNA-seq generated the largest library size, while the SFL and 3'DGE libraries were of comparable size (**Figure 2A**). Furthermore, full coverage poly-A RNA-seq yielded the highest percentage of reads aligned to the genome, followed by SFL and 3'DGE (**Table 1**, **Figure 2Ci**, and **Supplementary Figure S1A**). The lower mapping rate of 3'DGE is most likely due to the lower read quality scores of 3'DGE compared to full coverage RNA-seq and SFL (**Supplementary Figure S1B**). The mean percentage of reads with Phred quality scores greater than 20 (Q20) was only ~88% for 3'DGE, compared to ~100% for both full coverage RNA-seq and SFL. The relative 5'–3' transcript coverage for each sample across all three platforms is shown in **Supplementary Figure S1F**. As expected, reads alignments were skewed toward the 3' end of transcripts for 3'DGE, while we did observe relatively uniform coverage along the transcript for full coverage RNA-seq and SFL.

**TABLE 1** | Comparison of read assignment between full coverage poly-A RNA-seq, SFL, and 3'DGE.

	Counts (million)				Percent (value/library size)			
	Mean (SD)	Median	Minimum	Maximum	Mean (SD)	Median	Minimum	Maximum
<b>Poly-A RNA-seq (RNA-seq)</b>								
Library size (total reads)	13.0 (2.3)	12.6	9.3	17.6				
Aligned reads	12.4 (2.2)	12.0	9.0	16.9	95.9 (1.3)	96.0	92.4	97.9
Uniquely aligned reads	10.8 (1.9)	10.3	7.8	14.8	82.9 (1.5)	83.0	79.5	85.3
Counted reads	8.4 (1.5)	8.1	6.4	10.9	65.2 (2.7)	64.6	60.5	70.3
<b>Sparse full length sequencing (SFL)</b>								
Library size (total reads)	3.8 (1.1)	3.5	1.6	6.9				
Aligned reads	3.3 (1.0)	3.1	1.4	5.9	88.5 (2.9)	88.8	73.0	92.5
Uniquely aligned reads	1.8 (0.6)	1.8	0.7	3.2	48.5 (8.0)	46.8	27.6	64.8
Counted reads	0.9 (0.3)	0.9	0.3	1.6	24.5 (4.0)	23.8	14.3	31.7
<b>3' digital gene expression (3'DGE)</b>								
Library size (total reads)	3.7 (0.7)	3.7	1.9	5.6				
Aligned reads	3.0 (0.6)	3.0	1.5	4.5	80.6 (1.6)	81.0	73.5	82.2
Uniquely aligned reads								
Counted reads	1.2 (0.2)	1.2	0.7	1.8	33.3 (1.4)	33.0	30.5	38.6



For SFL there was a clear drop-off when going from percentage of aligned reads to percentage of uniquely aligned reads due to ribosomal RNA (rRNA) contamination of the SFL samples (**Figure 2Cii**). The majority of reads aligning to ribosomal regions specifically align to RNA28S (**Supplementary Figure S3**). For 3'DGE, unique UMIs are aligned directly to transcript sequences and not to the whole genome, such that the number of uniquely aligned reads and reads counted in transcripts are the same (**Figures 2Cii,iii**) (Morrissey et al., 2009). The percentage of reads that are counted in transcripts is greatest for full coverage poly-A RNA-seq (mean percentage of total library size: 65.2%), followed by 3'DGE (33.3%), and SFL (24.5%). However, while the counted read library size is greater for 3'DGE than for SFL, more genes were quantified by SFL than by 3'DGE (**Figure 2Civ**) (counts > 0 across all samples for 22,233 genes shared across

all three platforms.). A median of 60.9 and 50.5% genes were quantified by SFL and 3'DGE, respectively. The number of genes quantified was near the saturation point for each platform, such that this discrepancy is not due to read depth of each platform (**Supplementary Figure S1C**). The reason for the low gene discovery of 3'DGE is further illustrated in **Figure 2B**, where it is shown that the reads are more evenly distributed across the 22,233 genes by SFL than by 3'DGE, with the cumulative distribution of reads counted in individual genes nearly identical in SFL and full coverage poly-A RNA-seq.

The principal component (PC) error was estimated for each platform for different subsamples of the full counted library size. The first PC is shown in **Figure 2D**, while the second through the fifth PCs are shown in **Supplementary Figure S1D**. We observe that as the counted library size increases, the PC error decreases

at the fastest rate for full coverage RNA-seq, followed by SFL, then 3'DGE. Although these differences are considerably more prominent when comparing full coverage RNA-seq to either SFL or 3'DGE, we do observe that when down-sampling from 10 to 100% of the counted library size, the PC error decreases at a consistently faster rate for SFL than for 3'DGE. Initially subsampling full coverage RNA-seq and 3'DGE to match the full SFL counted library size does not change the results. The same trend is also observed in the cumulative variance explained by each successive PC across full coverage RNA-seq, SFL, and 3'DGE (**Supplementary Figure S1E**).

In summary, despite lower overall counted library size due to ribosomal RNA contamination, SFL demonstrates greater coverage in low-to-medium expressed genes than 3'DGE, comparable to full coverage poly-A RNA-seq. Consequently, the transcriptional signal captured by the SFL libraries are more robust to subsampling of the data compared to 3'DGE as measured by the principal component error.

## Signal-to-Noise Evaluation

Differential expression models comparing experimental groups of matched samples was performed in SFL, microarray, and 3'DGE and the corresponding signal-to-noise scores were compared pairwise between platforms (**Figure 3**). Samples shared across the three platforms include three replicates for each of four experimental groups, corresponding to *NRF2* overexpression or HcRed vehicle, as well as CSC chemical exposure or DMSO vehicle (**Figure 1**). Signal-to-noise was assessed by a four-group comparison with classic ANOVA (**Figures 3A–D**), as well as by stratified two-group differential analyses using *LIMMA* (**Figures 3E,F**).

We compared the  $\log_{10}$  *F*-statistics between ANOVA models across all three platforms (**Figure 3A**). Overall, the distribution of *F*-statistics is most similar between SFL and microarrays, with a Pearson correlation of 0.291. Though statistically significant ( $p < 0.01$ ), the corresponding mean difference between  $\log_{10}$  *F*-statistics is only 0.026. The mean differences of the  $\log_{10}$  *F*-statistics between SFL and 3'DGE, and between 3'DGE and microarray are 0.328 and 0.302, respectively, and the corresponding Pearson correlations are 0.160 and 0.216, respectively. These results are consistent with the discovery rates estimated for different FDR *Q*-value thresholds (**Figure 3B**). For example, at the FDR *Q*-value threshold of 0.05, the discovery rates of SFL and microarray are almost identical, 0.214 (2083 genes), 0.209 (2038 genes), respectively, while the discovery rate of 3'DGE is much smaller 0.032 (310 genes).

Loess regression of the  $\log_{10}$  *F*-statistics as a function of mean gene expression shows that the statistical signal increases with mean normalized expression. This trend is consistently positive for both SFL and 3'DGE, while leveling off at the most highly expressed genes in microarrays (**Figure 3C**). Furthermore, SFL signal is greater than 3'DGE signal at all levels of mean expression (**Figure 3C**). In agreement with the results from coverage comparison, the distribution of mean normalized expressions in 3'DGE is smaller than that of SFL, while SFL is comparable to that of microarray (**Figure 3D**). Adherence to assumption of normality, assessed through a Shapiro–Wilk

test, is also associated with higher mean normalized expression (**Supplementary Figure S4**).

The results of the comparisons of the two-group differential analyses across all three platforms were generally congruous with those of the four-group ANOVA analyses (**Figures 3E,F** and **Supplementary Figures S5, S6**). In all four two-group comparisons, the correlation of test statistics is closest between microarray and SFL results, followed by 3'DGE versus microarray results, and 3'DGE versus SFL. For example, in the DMSO-stratified, *NRF2* versus HcRed analysis, estimates of the Pearson correlations of test statistics are 0.66, 0.45, and 0.43, respectively (**Figure 3E**). The discovery rate of 3'DGE is the lowest across all four differential analyses, while the discovery rate of SFL is higher in three out of four of these analyses (**Figure 3F** and **Supplementary Figures S5, S6**).

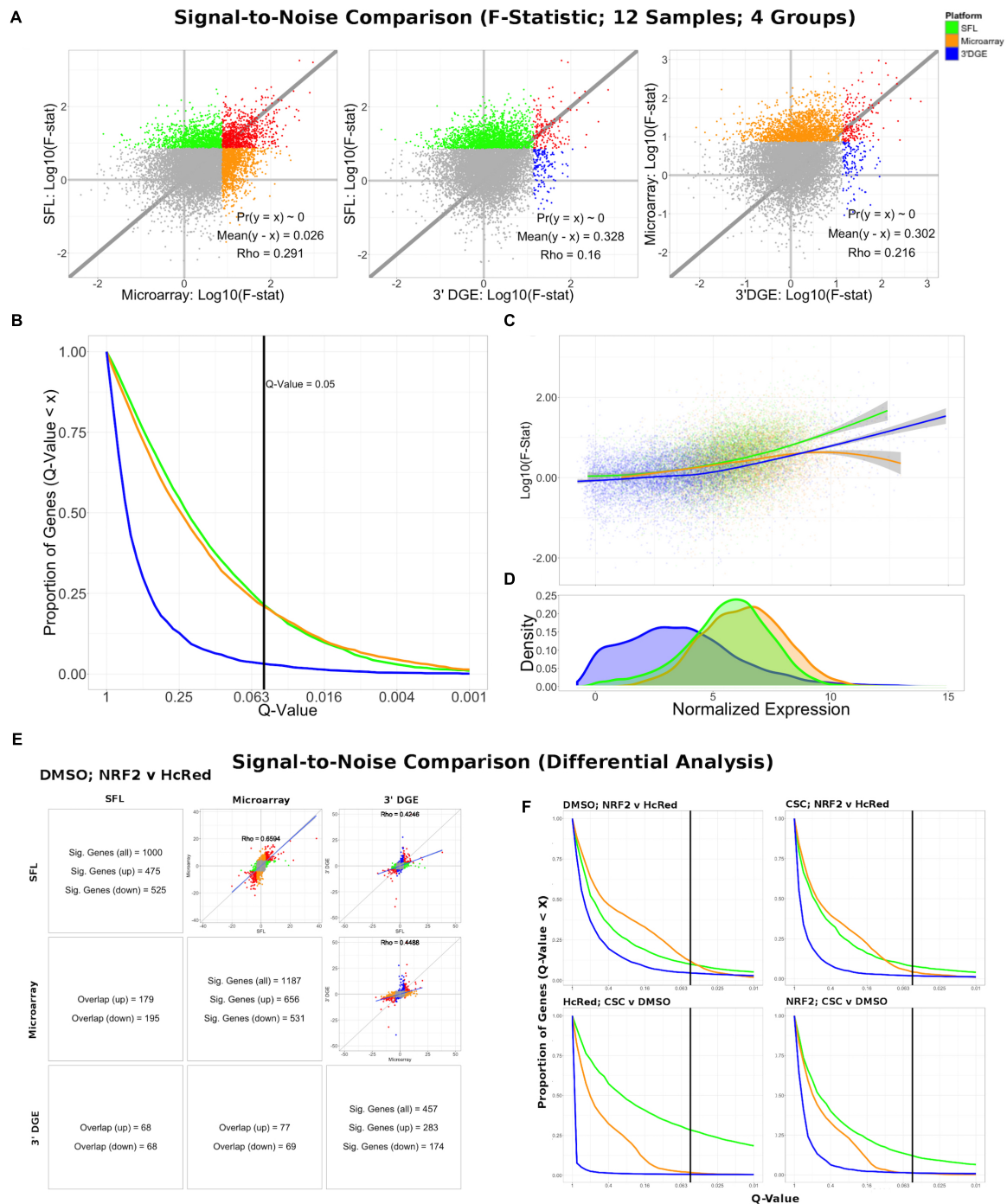
In summary SFL demonstrated greater statistical power than 3'DGE to detect differentially expressed genes, and its results more closely matched those in microarrays.

## Biological Signal Recapitulation Evaluation

To evaluate the ability of each platform to recapitulate biologically relevant results, we utilized previously published signatures of smoking exposure in lung (Spira et al., 2004; Beane et al., 2007), as well as differential signatures derived from the TCGA LUSC and LUAD datasets associated with mutations of the genes over-expressed in our experiments. From each of these signatures two gene sets were extracted, one of genes positively associated and one of genes negatively associated to the variable of interest. These gene sets were then tested via pre-ranked gene set enrichment analysis against each of our differential analysis results (CSC vs. DMSO, stratified by *NRF2* or HcRed perturbation; *NRF2* vs. HcRed, stratified by CSC or DMSO perturbation). The enrichment results with respect to both the smoking exposure signatures and the TCGA mutations are summarized in **Figure 4A**, and further detailed in **Supplementary Figure S7**, and confirm the highest sensitivity of microarrays, followed by SFL and 3'DGE.

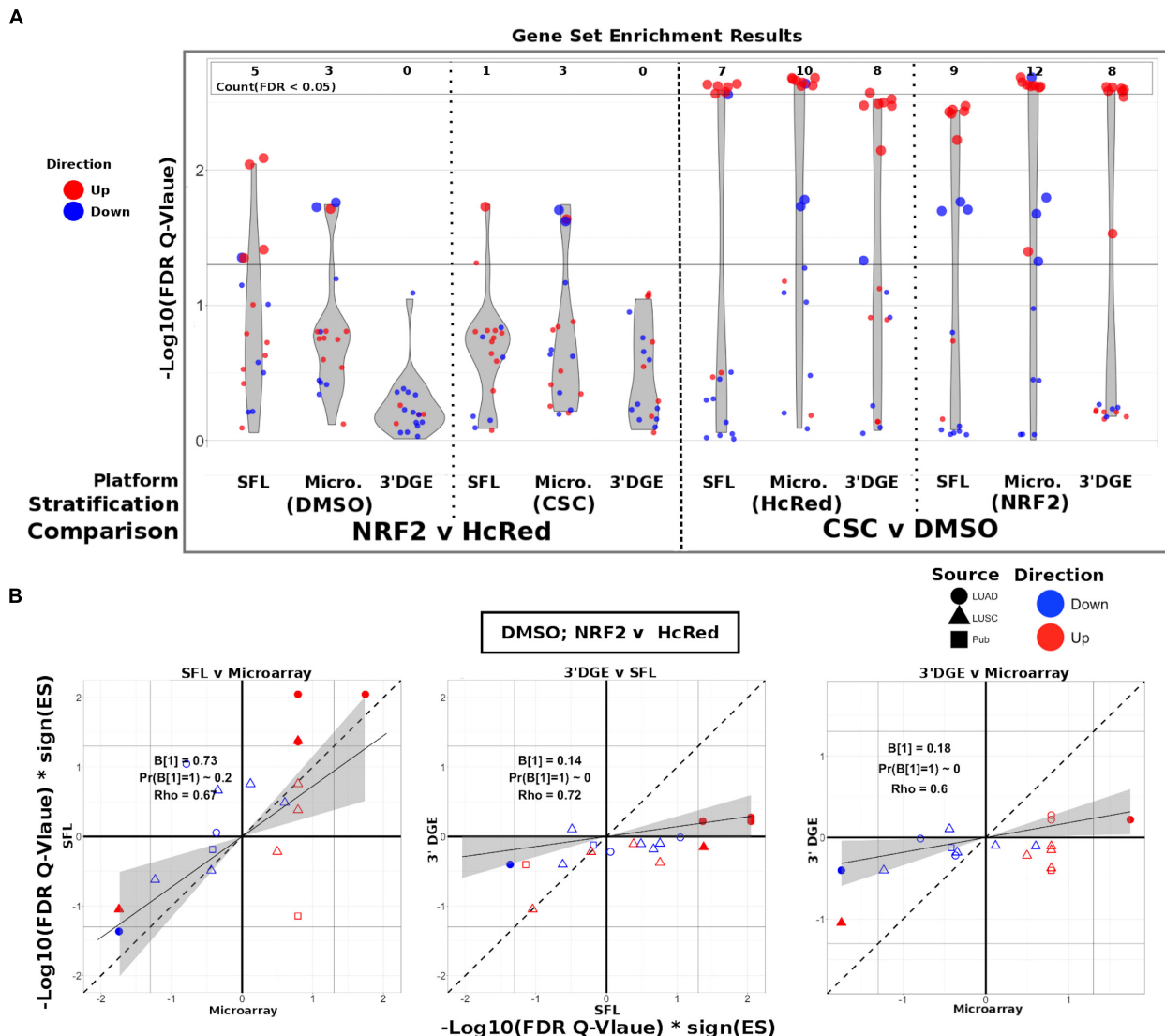
The set of genes up-regulated in “smokers vs. non-smokers” was found to be significantly (FDR *Q*-value  $< 0.05$ ) enriched in all “CSC vs. DMSO” signatures, within both genotypic stratifications for all three platforms. Conversely, the set of down-regulated genes in “smokers vs. non-smokers” was only enriched in the microarray signature of “*NRF2* over-expressed; CSC vs. DMSO” (**Supplementary Figure S7**).

The enrichment results of TCGA-derived gene sets with respect to differential signatures of genotypic perturbations were in agreement with the gene-level results, in that they consistently demonstrated smaller discovery rates by 3'DGE than by SFL or by microarrays (**Figure 4A**). For example, the significantly enriched gene sets in “DMSO-treated; *NRF2* vs. HcRed” differential signatures across all three platforms are highlighted in **Supplementary Figure S7**. The number of gene sets enriched in microarray, SFL, and 3'DGE platforms are five, three, and zero, respectively.



**FIGURE 3 |** Signal-to-noise comparison between SFL, microarray, and 3'DGE. **(A)** Scatterplots comparing the  $\log_{10}(\text{F-Statistics})$  from ANOVA models comparing four  $n = 3$  groups (HcRed:DMSO, HcRed:CSC, NRF2:DMSO, and NRF2:CSC). The gray line shows  $y = x$ . The platform with the higher mean  $\log_{10}(\text{F-Statistic})$  is plotted on the y-axis. Also, included are the  $p$ -value and difference in mean between each bi-platform comparison from paired  $t$ -testing, as well as the squared correlation coefficient.  $P$ -values  $\sim 0$  are less than 0.01. Color of indicate genes discovered by individual platforms (green, orange, or blue), neither platform (gray), and both platforms (red). **(B)** Plot of the Discovery Rate versus FDR Q-Value from threshold for each platform from four group ANOVA models. The x-axis is plotted on a  $-\log_{10}$  scale. The vertical line is indicative of a Q-value threshold of 0.05. **(C)** Loess fit of the  $\log_{10}(\text{F-Statistic})$  versus median normalized expression from four group ANOVA models. **(D)** Distribution of mean normalized expression across all three platforms. **(E)** Comparison of gene discovery (FDR Q-Value  $< 0.05$ ) by differential analysis with limma, comparing normalized gene expression between DMSO:NRF2 and DMSO:HcRed, including the raw discovery rates, discovered gene overlap, and linear fits, comparing test statistics from each platform. Genes that are discovered by more than 1 platform are shown in red in the scatterplots. Additional comparisons are shown in **Supplementary Figure S5**. **(F)** Plot of the Discovery Rate versus FDR Q-Value from threshold for each platform from two group differential analyses. The x-axis is plotted on a  $-\log_{10}$  scale. The vertical line is indicative of a Q-value threshold of 0.05.





**FIGURE 4 |** Comparison of gene-set enrichment of smoking and gene mutation signatures across SFL, 3'DGE and microarray. **(A)** Violin plots of the  $-\text{Log}_{10}(\text{FDR Q-Value})$  from gene set enrichment analysis of TCGA-derived gene-sets with respect to genotypic perturbations (left) and chemical perturbations (right) differential signatures across like samples within SFL, Microarray, and 3'DGE. Each column corresponds to differential signatures comparing genotypic or chemical perturbation groups, stratified by a single chemical or genotypic perturbation group, respectively, e.g., the left-most column shows the enrichment results with respect to the “DMSO-treated; NRF2 vs. HcRed” signature within the samples (*stratum*) in SFL data. Specific results for TCGA-derived gene sets are shown in **Supplementary Figure S7**. **(B)** Comparison of the gene set enrichment results between SFL, microarray and 3'DGE with respect to the “DMSO-treated; NRF2 vs. HcRed” differential signature. Shown are the transformed FDR Q-values of the TCGA-derived gene sets corresponding to mutations of NRF2 and CNA of KEAP1. The  $|\text{Log}_{10}(\text{FDR Q-Value})|$  corresponding to the  $\text{FDR} < 0.05$  significance thresholds are shown as vertical and horizontal gray lines for the y and x-axes, respectively. Points of gene sets whose enrichment meets this threshold in either of the two platforms are filled in. Colors and shape of points denote direction and source of the gene set, respectively. Additional results for chemical and genotypic perturbation signatures are shown in **Supplementary Figure S8**.

In addition to comparing which gene sets were significantly enriched in individual differential signatures, we compared the relative statistical signal of these enrichments. To this end, we transformed the permutation-based FDR Q-values by taking the negative  $\text{Log}_{10}$  and multiplying by the direction of the enrichment score (ES),  $-\text{Log}_{10}(\text{FDR Q-values}) * \text{sign}(\text{ES})$ . For each two-platform comparison, we fit a regression model through the origin. Since consistent results across platforms would result in a model fit close to the identity line,  $y = x$ , we tested

whether the slope coefficient equaled 1 (i.e.,  $B_1 = 1$ ). **Figure 4B** shows these results for each of the three comparisons of the NRF2 and KEAP1 mutation-based gene sets enrichment against the “DMSO-treated; NRF2 vs. HcRed” signatures. In all three comparisons, microarrays have the highest measured enrichment signal, followed by SFL and 3'DGE, however, the difference between microarray and SFL results is not significant,  $B_1 = 0.73$ ;  $p\text{-value} = 0.2$ . The coefficients for both of the comparisons to 3'DGE, are highly skewed in favor of microarray and SFL,



$B_1 = 0.18$  and  $0.14$ , respectively. Both of these comparisons are highly significant with  $p$ -values  $< 0.01$ . Comparison of the enrichment results for other differential signatures show similar trends (**Supplementary Figure S8**).

Next, we compared enrichment results with respect to all genotypic perturbation signatures between SFL and 3'DGE (**Figure 5A** and **Supplementary Figure S9A**). Each comparison (i.e., each point in the plot) denotes gene set enrichment results with respect to genotypic perturbations within each of the four chemical exposures, DMSO, CSC, BaP, and NNK. Gene sets were tested for enrichment against concordant differential signatures, e.g., the *PIK3CA* mutation-derived gene set was tested against the “*PIK3CA* vs. HcRed” signatures. As in the previous analysis, the permutation-based enrichment FDR  $Q$ -values were transformed by  $-\log_{10}(\text{FDR } Q\text{-values}) \times \text{sign}(\text{ES})$ . In the “DMSO-treated; genotypic perturbation vs. control” signatures, we observe that the gene set enrichment is generally more significant for SFL than for 3'DGE ( $B_1 = 0.63$ ;  $p$ -value  $< 0.01$ ; **Figure 5A**). The results obtained in CSC- and NNK-treated signatures, demonstrate concordance to these results ( $B_1 = 0.65$ ;  $p$ -value =  $0.03$  and  $B_1 = 0.60$ ;  $p$ -value =  $0.01$ , respectively). The BaP-treated results are less comparable since only one genotypic perturbation signature, “*FAT1* vs. GFP,” is available for this stratification (**Supplementary Figure S9A**).

Additionally, we compared our differential signatures to available full coverage poly-A RNA-seq genotypic perturbations (**Supplementary Figure S9B**), although these results are considered less comparable because of differences in experimental set-up. In particular, in the full coverage poly-A RNA-seq experiments the genotypic perturbations were performed on untreated rather than DMSO-treated cell lines (**Figure 1**).

The effect on discovery rate by subsampling the data across all three platforms is shown in **Figure 5B**. Generally, we did not observe a plateauing of discovery rate, where the number of detected genes plateaus near full counted library size. When comparing the correlation between GSEA results on subsampled data we observe similar trends across full coverage RNA-seq, SFL, and 3'DGE (**Figure 5C**). Initial subsampling of full coverage RNA-seq and 3'DGE to the SFL counted library size did not change the analysis results.

In summary, differential analysis of molecular and genotypic perturbations with SFL recapitulates biologically meaningful signal of gene sets derived from high coverage *in vivo* data sets. This performance is comparable to both 3'DGE and microarray.

## DISCUSSION

The goal of this study was to evaluate the performance of SFL sequencing, a low-cost method for performing highly multiplexed RNA-seq, and to compare it to other high-throughput gene expression profiling platforms. The development of such methods would be instrumental to the generation of large-scale perturbation screens based on *in vitro* models. The reduction of the cost per profile would make it feasible to significantly increase the number of replicates

and conditions to be profiled, including multiple time points, concentrations, and biological models, and thus would support a more in-depth investigation of the heterogeneity of the biological response to different exposures. It would also support the development of more accurate predictive models of the adverse or therapeutic outcomes of various exposures. Finally, insights gained from our study will also inform the design of protocols for single cell RNA-sequencing (Eberwine et al., 2014), given their reliance on highly multiplexed libraries.

In addition to SFL, the platforms included in this analysis were 3'DGE, an alternative highly multiplexed sequencing platform, Affymetrix GeneChip Human Gene 2.0 ST Microarray, an analog expression platform, and full coverage poly-A capture RNA-seq. The cost per sample for SFL and 3'DGE was  $\sim \$50$ , a 10-fold decrease from that of full coverage RNA-seq,  $\$500$ , and a 7-fold decrease from that of the microarray,  $\$350$  USD. Throughout this analysis we demonstrate comparable performances of SFL and 3'DGE to these more expensive platforms. Furthermore, in this analysis we consistently find evidence that SFL outperforms 3'DGE.

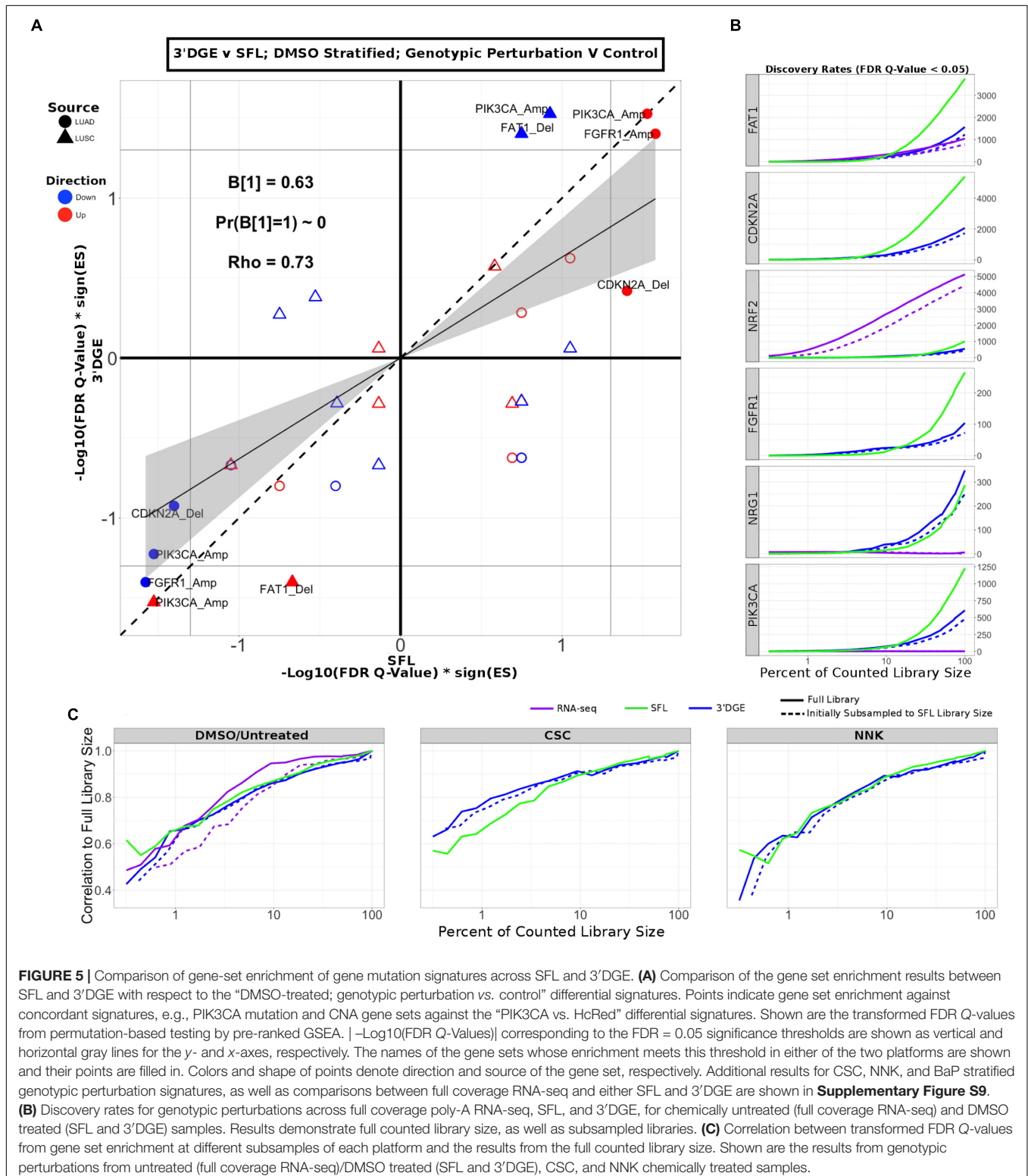
Performance was assessed in terms of coverage, signal-to-noise, and recapitulation of expected biological signal derived from independently generated, publicly available data collected from human subjects. Coverage was assessed by comparing the three digital expression platforms, while signal-to-noise and biological recapitulation was assessed by comparing SFL, 3'DGE, and microarrays. Microarray expression quantification has been shown to be highly correlated with qRT-PCR, especially when processed with updated probe set annotations, utilized in this analysis (Sandberg and Larsson, 2007). Chemical and molecular perturbations were carried out in the same samples, and concurrently profiled by SFL, 3'DGE, and microarrays. We also leveraged previously generated full coverage poly-A RNA-seq profiles from similar perturbations of AALE cell lines.

For coverage assessment, performance was evaluated in terms of the distribution of total reads, or library size, that were aligned to the human genome, and further quantified in annotated genes. The best performance was expected in full coverage poly-A RNA-seq, given that this is the most well-established technique and has by far the highest sequencing depth. This was confirmed, as full coverage poly-A RNA-seq was measured to have the highest per sample library size, percentage of aligned reads, percentage of uniquely aligned reads, and percentage of counted reads (**Figure 2** and **Supplementary Figure S1**). The coverage performance of SFL suffered as a result of rRNA contamination, where as many as 53% of the total library size per sample was assigned to ribosomal regions of the genome (**Supplementary Figure S3**).

3'DGE is a poly-A capture technique, therefore ribosomal depletion is not a possible pitfall. 3'DGE generates a short nucleotide tags from transposon-based fragmentation, which are enriched for 3' adjacent sequences of a given transcript (Soumillon et al., 2014). Since many transcripts of the same gene generate identical sequence tags, unique molecular identifiers (UMIs) are used to distinguish between unique reads and duplicate reads generated from PCR amplification. Although

mRNA fragment duplication occurs with any RNA-seq protocol, the impact of this artifact on downstream analyses is negligible for techniques, such as SFL, which generate more complex sequence libraries (Parekh et al., 2016).

3'DGE sequences were aligned directly to human mRNAs, rather than the whole genome. Therefore, percentages of reads aligned and reads counted (Figures 2Ci,iii) reflect the percentages of these non-unique UMIs that align to at least one



gene and the number of unique UMIs that align to only one gene, respectively. We observe that the percentage of counted reads is greater for 3'DGE than SFL, which is explained by a loss of reads to rRNA contamination in SFL. However, we observe notably more genes quantified by SFL than by 3'DGE (Figures 2B,Civ), which indicates that more reads are assigned to fewer genes in 3'DGE compared to SFL, as well as to full coverage RNA-seq (Figure 2C). Although rRNA contamination is a potential drawback of any ribosomal depletion RNA-sequencing technique, the extent of ribosomal contamination is variable, and could be potentially improved by further optimization of the library preparation protocol.

The difference in distribution of reads across shared genes between SFL and 3'DGE likely explains the difference in information retained by subsampling as measured by principal component error. Although full coverage poly-A RNA-seq clearly outperforms both SFL and 3'DGE for principal component assessment, we consistently observe that, as the counted library size increases, the rate of principal component error decreases faster for SFL than 3'DGE (Figure 2D and Supplementary Figure S1D). This is unsurprising considering that not only are considerably fewer genes quantified by SFL compared to 3'DGE, but there is also no discernable difference between the rate of genes counted as a function of counted library size between the two platforms (Supplementary Figure S1C). As we subsample the counted libraries, though we may lose the same number of genes between SFL and 3'DGE, the percent of genes lost, and consequently the information lost, will be greater for 3'DGE than SFL. Furthermore, this more even read distribution likely explains the improved performance of SFL over 3'DGE in statistical signal. In particular, our signal-to-noise evaluation shows consistently higher gene-level statistical signal from SFL and microarray experiments than from 3'DGE experiments (Figure 3). These differences appear to be driven by the differences in the relative quantification of genes, given that statistical signal is positively associated with mean gene expression for each platform, and 3'DGE experiments showed lower gene-level quantification than SFL and microarrays (Figures 3C,D). We observe similar cross-platform relationships in the two-group differential analyses (Figures 3E,F).

The gene set-based enrichment results are consistent with those from signal-to-noise analyses. In every comparison of enrichment scores between SFL and 3'DGE, we observe generally higher gene set enrichment with respect to the SFL-derived signatures (Figures 4, 5A and Supplementary Figures S8, S9). The gene sets were selected to represent known biological responses to the profiled perturbations, and thus their enrichment with respect to the perturbation signatures are expected to be true positives.

The enrichment results confirm this expectation. For example, in the signatures of *NRF2* overexpression, we consistently observe enrichment of the gene sets derived from *NRF2* amplifications and *KEAP1* deletions, each of which should increase *NRF2* activity (Supplementary Figure S7) (Kansanen et al., 2013). Similarly, we observe significant concordant enrichment of the gene sets derived from *NRF2* and *KEAP1*-dysregulated lung tumors in the signature of CSC exposure, suggesting

that the *NRF2* pathway is activated by CSC exposure *in vitro* (Supplementary Figure S7), which has been previously reported (Adair-Kirk et al., 2008). Interestingly, these results demonstrate that the activation of the *NRF2* pathway in normal airway epithelial cells *in vitro* (by ectopic expression of the gene or by CSC treatment) is concordant with the activation of *NRF2* by somatic genome alterations in lung tumors, a finding that, to the best of our knowledge, has not been previously observed.

Possible sources of technical variability in this study are the different sequencing platforms, service providers, and read lengths. However, when subsampling the 3'DGE and SFL counted libraries, we generally observe higher discovery rates at all percentages of the full counted libraries, and even more so when the 3'DGE counted libraries are initially subsampled to full SFL counted library sizes (Figure 5B), demonstrating that SFL shows improvements independent of the mapping rate. This result confirms previous reports showing that increasing read length above 50-bp does not improve read quantification (Chhangawala et al., 2015). Furthermore, similar results have been reported even when the same sequencing platform is used. A recent study reported a greater number of genes detected, as well as higher differential analysis discovery rates, in conventional RNA-seq than in 3'DGE at identical counted library sizes, using the Illumina HiSeq 2500 platform to generate both libraries (Xiong et al., 2017).

In summary, in this study we observe higher performance of SFL than 3'DGE, as measured by coverage, signal-to-noise, and biological recapitulation of known signal, with the performance of SFL often matching that of well-established "gold standards" (full coverage RNA-seq or microarrays). On the other hand, the fact that 3'DGE is shown to allocate a large number of reads to relatively fewer, highly expressed genes, makes this platform more suitable for problems where high accuracy in the differential quantification of highly expressed genes is needed. Furthermore, the ready availability of 3'DGE as a core-provided option, which allows for the out-sourcing of library preparation, sequence read pre-processing and gene quantification, is an additional value-added of the platform. Ultimately, the best-suited platform for a specific project will depend on the study goals, design, and availability of different resources. We believe our study presents useful results to make a more informed choice.

The utility of highly multiplexed RNA-seq crucially depends on the trade-off between cost and data quality, and on the nature of the experiments for which the platform would be ideally suitable. These will in general be experiments where the marginal information content of a single profile is relatively low, and thus justifies trading-off some data quality for reduced cost.

## DATA AVAILABILITY

Data for SFL, 3'DGE, and Microarray experiments is available through the Gene Expression Omnibus (GEO) at accession numbers: GSE118797, GSE118798, and GSE118799. Reviewers

may access the data prior to publication using the tokens: mfbiskyuxpudxmz, gdpyscyerrolvad, and cxojquwahlgttsl.

## AUTHOR CONTRIBUTIONS

SM, CP, and JC designed the experiments. EM performed the wet-lab experiments. ER performed all data pre-processing and analysis. ER, SM, JC, and CP interpreted the analysis results. XX and GL refined, implemented, and recorded the SFL protocol. ER, SM, JC, and CP wrote the manuscript. SM oversaw the whole project.

## FUNDING

This work was supported in part by a Superfund Research Program grant P42ES007381 to SM, a Find the Cause Breast Cancer Foundation (<http://findthecause.org>) grant to SM, an

Evans Foundation pilot grant to SM and CP, and a LUNgevity Career Development award to JC.

## ACKNOWLEDGMENTS

Dr. Alexander A. Shishkin for his feedback during the development of the SFL protocol. 3' Digital Gene Expression libraries were prepared by the Broad Technology Labs and sequenced by the Broad Genomics Platform, using SCRBS-Seq library preparation techniques.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.00150/full#supplementary-material>

## REFERENCES

- Adair-Kirk, T. L., Atkinson, J. J., Griffin, G. L., Watson, M. A., Kelley, D. G., DeMello, D., et al. (2008). Distal airways in mice exposed to cigarette smoke: Nrf2-regulated genes are increased in clara cells. *Am. J. Respir. Cell Mol. Biol.* 39, 400–411. doi: 10.1165/rcmb.2007-0295OC
- Asmann, Y. W., Klee, E. W., Thompson, E. A., Perez, E. A., Middha, S., Oberg, A. L., et al. (2009). 3' tag digital gene expression profiling of human brain and universal reference RNA using illumina genome analyzer. *BMC Genomics* 10:531. doi: 10.1186/1471-2164-10-531
- Beane, J., Sebastiani, P., Liu, G., Brody, J. S., Lenburg, M. E., and Spira, A. (2007). Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol.* 8:R201. doi: 10.1186/gb-2007-8-9-r201
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300.
- Bryant, D. W., Priest, H. D., and Mockler, T. C. (2012). "Detection and quantification of alternative splicing variants using RNA-seq," in *RNA Abundance Analysis*, eds H. Jin and W. Gassmann (Totowa, NJ: Humana Press), 97–110.
- Campbell, J. D., Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., et al. (2016). Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* 48, 607–616. doi: 10.1038/ng.3564
- Chhangawala, S., Rudy, G., Mason, C. E., and Rosenfeld, J. A. (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biol.* 16:131. doi: 10.1186/s13059-015-0697-y
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635
- Eberwine, J., Sul, J.-Y., Bartfai, T., and Kim, J. (2014). The promise of single-cell sequencing. *Nat. Methods* 11, 25–27. doi: 10.1038/nmeth.2769
- Ganter, B., Snyder, R. D., Halbert, D. N., and Lee, M. D. (2006). Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix<sup>®</sup> database. *Pharmacogenomics* 7, 1025–1044. doi: 10.2217/14622416.7.7.1025
- Heimberg, G., Bhatnagar, R., El-Samad, H., and Thomson, M. (2016). Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* 2, 239–250. doi: 10.1016/j.cels.2016.04.001
- Hou, Z., Jiang, P., Swanson, S. A., Elwell, A. L., Nguyen, B. K. S., Bolin, J. M., et al. (2015). A cost-effective RNA sequencing protocol for large-scale gene expression studies. *Sci. Rep.* 5:9570. doi: 10.1038/srep09570
- Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T., et al. (2015). Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.* 43, D921–D927. doi: 10.1093/nar/gku955
- Kansanen, E., Kuosmanen, S. M., Leinonen, H., and Levenon, A.-L. (2013). The Keap1-Nrf2 pathway: mechanisms of activation and dysregulation in cancer. *Redox Biol.* 1, 45–49. doi: 10.1016/j.redox.2012.10.001
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324
- Liao, Y., Smyth, G. K., and Shi, W. (2014). Feature counts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656
- Lundberg, A. S., Randell, S. H., Stewart, S. A., Elenbaas, B., Hartwell, K. A., Brooks, M. W., et al. (2002). immortalization and transformation of primary human airway epithelial cells by gene transfer. *Oncogene* 21, 4577–4586. doi: 10.1038/sj.onc.1205550
- Morrissey, A. S., Morin, R. D., Delaney, A., Zeng, T., McDonald, H., Jones, S., et al. (2009). Next-generation tag sequencing for cancer gene expression profiling. *Genome Res.* 19, 1825–1835. doi: 10.1101/gr.094482.109
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226
- Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., and Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* 6:25533. doi: 10.1038/srep25533
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007
- Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., et al. (2010). De novo assembly and analysis of RNA-seq data. *Nat. Methods* 7, 909–912. doi: 10.1038/nmeth.1517
- Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25. doi: 10.1186/gb-2010-11-3-r25
- Sandberg, R., and Larsson, O. (2007). Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics* 8:48. doi: 10.1186/1471-2105-8-48
- Shishkin, A. A., Giannoukos, G., Kucukural, A., Ciulla, D., Busby, M., Surka, C., et al. (2015). Simultaneous generation of many RNA-seq libraries in a single reaction. *Nat. Methods* 12, 323–325. doi: 10.1038/nmeth.3313
- Soumilion, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A., and Mikkelsen, T. S. (2014). Characterization of directed

- differentiation by high-throughput single-cell RNA-Seq. *bioRxiv* [Preprint]. doi: 10.1101/003236
- Spira, A., Beane, J., Shah, V., Liu, G., Schembri, F., Yang, X., et al. (2004). Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* 101, 10143–10148. doi: 10.1073/pnas.0401422101
- Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171, 1437.e17–1452.e17. doi: 10.1016/j.cell.2017.10.049
- Wang, L., Si, Y., Dedow, L. K., Shao, Y., Liu, P., and Brutnell, T. P. (2011). A low-cost library construction protocol and data analysis pipeline for illumina-based strand-specific multiplex RNA-seq. *PLoS One* 6:e26426. doi: 10.1371/journal.pone.0026426
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484
- Xiong, Y., Soumillon, M., Wu, J., Hansen, J., Hu, B., van Hasselt, J. G. C., et al. (2017). A comparison of mRNA sequencing with random primed and 3'-directed libraries. *Sci. Rep.* 7:14626. doi: 10.1038/s41598-017-14892-x

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Reed, Moses, Xiao, Liu, Campbell, Perdomo and Monti. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Read Mapping and Transcript Assembly: A Scalable and High-Throughput Workflow for the Processing and Analysis of Ribonucleic Acid Sequencing Data

Sateesh Peri<sup>1</sup>, Sarah Roberts<sup>2</sup>, Isabella R. Kreko<sup>3</sup>, Lauren B. McHan<sup>3</sup>, Alexandra Naron<sup>3</sup>, Archana Ram<sup>3</sup>, Rebecca L. Murphy<sup>4</sup>, Eric Lyons<sup>1,2</sup>, Brian D. Gregory<sup>5</sup>, Upendra K. Devisetty<sup>2</sup> and Andrew D. L. Nelson<sup>6\*</sup>

<sup>1</sup> Genetics Graduate Interdisciplinary Group, University of Arizona, Tucson, AZ, United States, <sup>2</sup> CyVerse, University of Arizona, Tucson, AZ, United States, <sup>3</sup> LIVE-for-Plants Summer Research Program, School of Plant Sciences, University of Arizona, Tucson, AZ, United States, <sup>4</sup> Biology Department, Centenary College of Louisiana, Shreveport, LA, United States, <sup>5</sup> Department of Biology, University of Pennsylvania, Philadelphia, PA, United States, <sup>6</sup> Boyce Thompson Institute, Cornell University, Ithaca, NY, United States

## OPEN ACCESS

### Edited by:

Filippo Geraci,  
Italian National Research Council  
(CNR), Italy

### Reviewed by:

Cuncong Zhong,  
University of Kansas, United States  
Eve Syrkin Wurtele,  
Iowa State University, United States

### \*Correspondence:

Andrew D. L. Nelson  
an425@cornell.edu

### Specialty section:

This article was submitted to  
Bioinformatics and  
Computational Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 16 July 2019

**Accepted:** 12 December 2019

**Published:** 24 January 2020

### Citation:

Peri S, Roberts S, Kreko IR, McHan LB, Naron A, Ram A, Murphy RL, Lyons E, Gregory BD, Devisetty UK and Nelson ADL (2020) Read Mapping and Transcript Assembly: A Scalable and High-Throughput Workflow for the Processing and Analysis of Ribonucleic Acid Sequencing Data. *Front. Genet.* 10:1361. doi: 10.3389/fgene.2019.01361

Next-generation RNA-sequencing is an incredibly powerful means of generating a snapshot of the transcriptomic state within a cell, tissue, or whole organism. As the questions addressed by RNA-sequencing (RNA-seq) become both more complex and greater in number, there is a need to simplify RNA-seq processing workflows, make them more efficient and interoperable, and capable of handling both large and small datasets. This is especially important for researchers who need to process hundreds to tens of thousands of RNA-seq datasets. To address these needs, we have developed a scalable, user-friendly, and easily deployable analysis suite called RMTA (Read Mapping, Transcript Assembly). RMTA can easily process thousands of RNA-seq datasets with features that include automated read quality analysis, filters for lowly expressed transcripts, and read counting for differential expression analysis. RMTA is containerized using Docker for easy deployment within any compute environment [cloud, local, or high-performance computing (HPC)] and is available as two apps in CyVerse's Discovery Environment, one for normal use and one specifically designed for introducing undergraduates and high school to RNA-seq analysis. For extremely large datasets (tens of thousands of FASTq files) we developed a high-throughput, scalable, and parallelized version of RMTA optimized for launching on the Open Science Grid (OSG) from within the Discovery Environment. OSG-RMTA allows users to utilize the Discovery Environment for data management, parallelization, and submitting jobs to OSG, and finally, employ the OSG for distributed, high throughput computing. Alternatively, OSG-RMTA can be run directly on the OSG through the command line. RMTA is designed to be useful for data scientists, of any skill level, interested in rapidly and reproducibly analyzing their large RNA-seq data sets.

**Keywords:** RNA-seq, transcriptomics, high throughput (-omics) techniques, bioinformatics, workflow

## INTRODUCTION

RNA-sequencing (RNA-seq) provides scientists with the ability to monitor genome-wide transcription across numerous cells or tissues and between experimental conditions in a rapid and affordable manner. Data generated from RNA-sequencing are incredibly powerful for differential gene expression analysis (Mortazavi et al., 2008; Li et al., 2016; Schlackow et al., 2017), novel gene discovery (Martin et al., 2013; Nelson et al., 2017), transcriptome-wide structural analysis (Gosai et al., 2015; Anderson et al., 2018), and even transcriptome-wide association studies (Galpaz et al., 2018; Gusev et al., 2019). In addition to generating and examining novel RNA-seq data, scientists are re-examining the hundreds of thousands of publicly available archived datasets to make novel discoveries (Lachmann et al., 2018), an analytical feat that represents a bottleneck for most researchers. The popularity of RNA-sequencing is perhaps most apparent by examining the dramatic increase in the number of RNA associated sequence read archives (SRAs) deposited in National Center for Biotechnology Information (NCBI's) SRA (Leinonen et al., 2011; **Figure 1**) over the last 10 years.

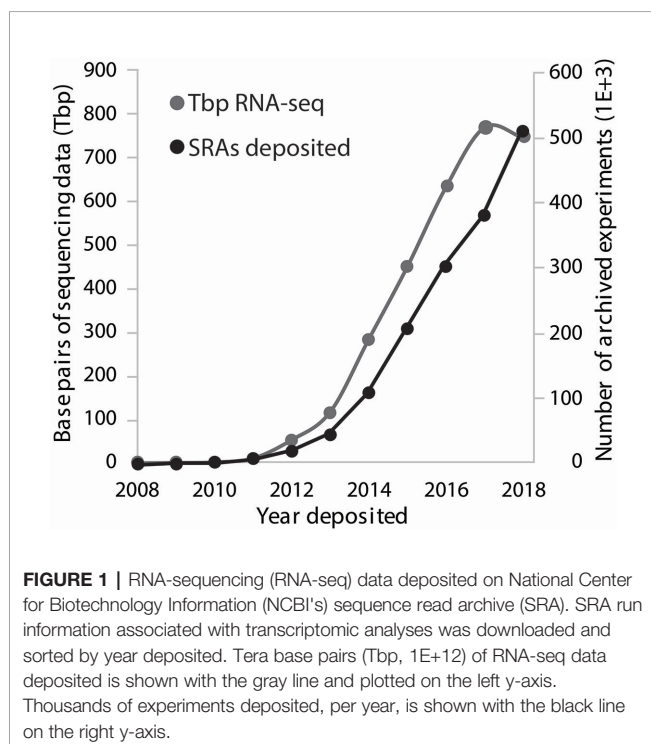
Alignment-based processing of these massive volumes of RNA-seq data typically involves two computationally intensive steps: mapping reads against a reference genome and transcript assembly. Reference genome based read mapping is performed using splice-aware algorithms such as STAR (Dobin et al., 2013) or HISAT2 (Pertea et al., 2016). The computational cost associated with mapping reads is dependent on the size of the genome and the number of reads to be mapped but typically

takes hours to days on a standard lab server. The mapped reads are then used to assemble transcripts using programs such as StringTie or Cufflinks. Transcript assembly is less computationally intensive than read mapping but can still require several hours to complete. In addition to the computational requirements, both of these steps require substantial data storage resources and technical skills in transferring and manipulating large files, further increasing the technological burden for the researcher.

Successful assembly of RNA-seq data is insufficient to achieve the ultimate experimental goal: extraction of meaningful data. Data extraction usually involves differential expression analyses, isoform analysis, or novel gene identification. Each of these analyses requires different input file types and the use of different applications—each with their own intricacies surrounding installation, use, and preference for a Linux environment. In addition, preparing data files and then organizing them into the appropriate file structure for these next steps rapidly becomes tedious when performed on hundreds to thousands of files. Thus, despite the wealth of computing resources, extracting meaningful knowledge from RNA-seq data is still a non-trivial task.

Cloud-computing cyber-infrastructure platforms such as CyVerse (Merchant et al., 2016) and Galaxy (Afgan et al., 2016) have lifted the computational and data management burdens and made RNA-seq analysis more accessible to non-traditional data scientists. In contrast to fee-based services such as the Cancer Genomics Cloud (Lau et al., 2017) or FireCloud (Chet et al., 2017 – doi 10.1101/209494), CyVerse and Galaxy are free to users and provide long-term data storage solutions integrated with limited on-demand cloud compute resources. CyVerse and Galaxy also offer graphical user interface (GUI) platforms which allow researchers with minimal programming experience to easily deploy and handle large volumes of jobs in parallel. A complement to single-source resources like CyVerse and Galaxy is the Open Science Grid [OSG (Pordes et al., 2007)], a distributed computing resource capable of handling hundreds of thousands of jobs and transferring hundreds of petabytes of data per day. Thus, these computational resources make large dataset analysis and re-analysis feasible in a reasonable time-frame and cost-effective way.

Here we introduce RMTA (Read Mapping, Transcript Assembly), a high throughput RNA-seq read mapping and transcript assembly workflow. RMTA is easy to use and incorporates features that move beyond the standard RNA-seq workflow, allowing data scientists to focus their time on downstream analyses. For users with access and familiarity with high-performance computing (HPC) command-line operations, RMTA is packaged as a Docker container for one-step installation (**Table 1**). In contrast to other containerized RNA-seq analysis tools (Polarin et al., 2015; Jensen et al., 2018), RMTA is also installed as an app in CyVerse's Discovery Environment, which obviates computing and data storage requirements while providing a GUI for users less familiar with the command-line. Finally, for users querying extremely large data sets, OSG-RMTA marries the computational resources



**TABLE 1 |** Deployment options for read mapping and transcript assembly (RMTA).

Platform	App Name	Size of Datasets That Can Be Handled	Data Storage Available	Genome Services Available
DE	RMTA v2.5.1.2	1–100	Yes	Yes
DE	OSG-RMTA v2.5.1.2	100–1000s	Yes	Yes
DE	RMTA-Instructional	1–10	Yes	Yes
Local	RMTA in Docker	Restricted to user capacity	No	No
OSG*	OSG-RMTA	100–1000s	No	No

Platforms include the Discovery Environment, a local computer, or high performance computing center, or the Open Science Grid. \*Users wishing to utilize the Open Science Grid (OSG) outside of the Discovery Environment will need their own OSG account.

of the OSG with the job scheduling, data storage and management capabilities of CyVerse. Beyond read mapping and assembly, RMTA has a number of additional features that automate onerous data transformation and quality control steps, thus producing outputs that can be directly used for differential expression analysis or novel gene identification. In addition, the output from RMTA may be rapidly integrated in downstream transcriptomic data visualization platforms to help researchers extract meaningful knowledge. RMTA is both straightforward to install and use, and is meant to be used by both advanced and novice data scientists in their examination of their RNA-seq data.

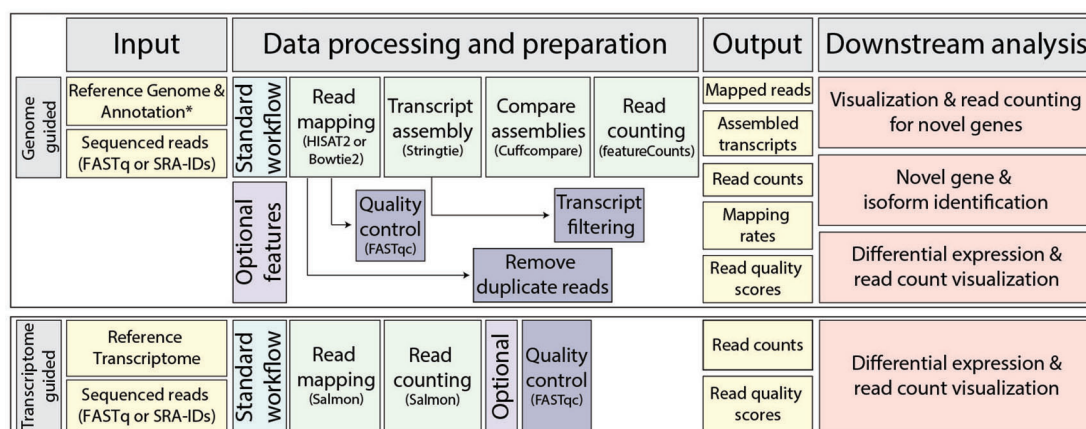
## MATERIALS AND METHODS

In this section we provide an overview of RMTA, its different features, and its deployment options.

## Overview of the Read Mapping and Transcript Assembly Workflow

RMTA automates the three critical steps of RNA-seq analysis: read mapping, transcript assembly, and read counting. For genome-guided read mapping, RMTA utilizes either the splice-aware algorithm HISAT2 or the splice-unaware algorithm Bowtie 2 (Langmead and Salzberg, 2012) for mapping and then StringTie (Pertea et al., 2016) for transcript assembly (**Figure 2**). Minimum input requirements include a reference genome (FASTA or pre-indexed), and RNA-seq reads as either compressed or uncompressed FASTq, or as a list of one to thousands of SRA IDs. A reference genome annotation file (in GFF/GFF3/GTF) is optional and allows for downstream novel gene identification. RMTA automatically builds a reference genome index (if it is not provided) from the user supplied reference genome, aligns reads to the genome, and then returns a binary encoded version of a sorted sequence alignment map (BAM) file for each input FASTq/SRA. This BAM file is then automatically used as input for StringTie, where it, along with the reference genome annotation, is used to assemble transcripts. Following transcript assembly, each BAM file is processed by featureCounts (Liao et al., 2014) to determine how many reads map back to each gene/exon in the reference genome annotation file.

As an alternative to genome-guided read mapping and transcript assembly, RMTA also allows for read alignment directly to a transcriptome using the quasi-aligner and transcript abundance quantifier Salmon (Patro et al., 2017; Srivastava et al., 2019). Minimum input for Salmon includes a reference transcriptome (in FASTA format) and then RNA-seq reads (as above). Salmon maps reads to the provided transcript assembly and then counts the number of reads associated with each transcript, generating an output file (quant.sf) that can immediately be used for differential expression. The utilization of



**FIGURE 2 |** Read mapping and transcript assembly (RMTA) workflow with suggested downstream analyses. The standard RMTA workflow consists of read mapping by either HISAT2 or Bowtie 2, transcript assembly by StringTie, assembly comparison to the reference annotation by Cuffcompare to identify novel transcripts, and then read counting by featureCounts. Several optional features are included, such as the ability to perform quality control on RNA-sequencing (RNA-seq) data with FastQC, filtering of lowly expressed transcripts, and removal of duplicate reads (Bowtie 2 only). Output is listed, and are ready for downstream analyses such as those shown.

Salmon is only appropriate when the user is wanting to rapidly test for differential expression and cannot facilitate the identification of novel genes or data visualization in a genome browser.

OSG-RMTA utilizes a similar workflow to RMTA. The primary difference is how the user plans on launching jobs and providing the necessary input data to the OSG. When launched directly from within the OSG through a user's personal account, the user must provide access to all necessary data (e.g. genomes, RNA-seq data, etc). Thus, we recommend users submit jobs to the OSG through CyVerse's Discovery Environment. When jobs are submitted *via* the Discovery Environment, it automatically prepares the information needed to run the job and submits it to the OSG *via* HTCondor (Thain et al., 2005) and requires no OSG account (Table 1). Once the job is launched OSG-RMTA uses the information provided by the Discovery Environment to retrieve input files, process the data, and upload the results back to the Data Store, allowing the user to submit and walk away.

RMTA is also available for implementation on a HPC, a public cloud-based computing system (i.e., XSEDE or Atmosphere), or a local compute system. For local or cloud-based computing, a Dockerized version of RMTA identical to that used in the Discovery Environment is available for use inside a Docker command line environment. However, the user will need to direct Docker to the location of the input files and assign the required “flags” that are hidden when using RMTA in the Discovery Environment. More information on how to run the Docker version of RMTA on a Linux/personal computer (PC)/Mac operating system (OS) and a list of all available flags are available here (<https://github.com/Evolinc/RMTA>). Docker requires root privileges and thus is not available for HPC where users are denied super user do “sudo.” For HPC systems, Docker can be used alongside Singularity (Kurtzer, et al., 2017; instructions found here: <https://sylabs.io/guides/3.4/user-guide/>).

## Additional Read Mapping and Transcript Assembly Features

Several additional features have been included in the RMTA workflow to facilitate data discovery and quality control. For users wishing to call single nucleotide polymorphisms from their RNA-seq [or DNA-sequencing (DNA-seq)] data in a high throughput manner, the read aligner Bowtie 2 (Langmead and Salzberg, 2012) has been included as an optional aligner in the RMTA workflow. When the Bowtie option is selected, HISAT2 and StringTie are both removed from the workflow, but the additional option to remove duplicate reads (important for population level analyses) becomes available.

Poor quality RNA-seq reads, particularly at the 5' or 3' ends as a result of adaptor contamination or a drop in sequencing quality, can lead to a significant population of unmapped reads. To help the user identify issues resulting from poor read mapping rates, the quality control tool FastQC (Andrews, 2010) is available as an additional option in the RMTA workflow for both genome or transcriptome-guided read mapping approaches. FastQC provides the user with both an overview

of potential issues with their data, as well as summary graphs highlighting issues such as per base sequence quality and Kmer content. Because FastQC works on read files in FASTq format, and we envision many users running RMTA directly on SRAs, FastQC has been placed downstream of read mapping (Figure 2). When the FastQC option has been selected, BAM files are converted back into FASTq with mapped and unmapped reads, along with their associated quality score, retained. This FASTq file is then used as input for FastQC, and then deleted afterward to reduce disk usage. If issues are detected at the 5' or 3' of sequencing reads, RMTA includes additional options for specifically trimming bases off of either end during the next analysis. Sequencing reads of overall poor quality will simply not be mapped and therefore do not need to be trimmed, but will still be highlighted in the FastQC results.

RMTA is also designed to aid in the identification of novel genes such as long non-coding RNAs from genome-guided transcriptome assemblies. To help the user remove transcript assembly artifacts that can arise from low expression, and therefore improve their attempts at novel gene identification, RMTA has two options for filtering lowly expressed transcripts. The user can decide to filter based on low expression [denoted as fragments per kilobase of transcript per million mapped read (FPKM)], low/incomplete read coverage (read per base), or use both filters in combination. We find that applying both filters (e.g., setting them both to one) helps to remove a large percentage of poorly assembled transcripts.

## Output From RMTA

The RMTA workflow produces a number of files that are designed to be immediately useful for downstream analyses such as differential expression, novel gene identification, and single-nucleotide polymorphism (SNP) discovery. Directly within the RMTA\_Output folder the user will find the sorted BAM files and the filtered transcript assembly files (in GTF). The naming convention of these files reflects the SRA or FASTq from which they were derived (i.e., the input ID will be prepended to the output files). The filtered transcript assembly file is prepared for immediate use in the novel long non-coding RNA (lncRNA) identification package, Evolinc (Nelson et al., 2017), whereas the sorted BAM file is ready for import and visualization within a genome browser such as EPIC-CoGe (Nelson et al., 2018) or Integrative Genomics Viewer (IGV) (J. T. Robinson et al., 2011). The user will also find a “mapped.txt” file in the RMTA\_Output folder, which contains information about alignment rates for each input FASTq/SRA. Within the RMTA\_output folder is a subfolder labeled “Feature\_counts” which contains a featureCounts summary.txt file and a tab-delimited file containing the number of reads assigned to each gene/exon for each of the RNA-seq data sets analyzed. If using the transcriptome-guided mapping approach (i.e., Salmon), a single quant.sf file will be generated that will contain the counts of all reads mapped to each transcript in each of the RNA-seq datasets processed. If the user selected the FastQC option, there will be a subfolder within the Output folder called “FastQC\_out.” This folder will contain a FastQC.html file for each data set examined. Clicking on this file within the Discovery



Environment will open up a new tab in the user's browser where all of FastQC's output information will be displayed. If the user chose Bowtie as the read aligner and "remove duplicate reads" as an additional option, then the RMTA\_Output folder will only contain a sorted BAM file with duplicates removed for each SRA/FASTq input file, as well as a mapped.txt file. No additional files will be generated. A similar file/folder structure is generated no matter how an RMTA job has been launched (DE/OSG/HPC).

## Deployment Options

The different deployment options for RMTA and the benefits associated with each are summarized in **Table 1**. RMTA is freely available as an app (RMTA v2.6.3) within CyVerse's Discovery Environment (<https://wiki.cyverse.org/wiki/display/DEapps/RMTA+v2.6.3>). Running RMTA within the Discovery Environment allows the user to take advantage of CyVerse's simplified data management and storage options through the Data Store. In addition, integrated in the Discovery Environment are a number of virtual interactive computing environment (VICE) apps, such as the DESeq2 RStudio app, that allow users to examine their data start to finish completely in the cloud (<https://learning.cyverse.org/projects/vice/en/latest/>). OSG-RMTA (v2.6.3) is available as a separate app within the Discovery Environment. Although the OSG-RMTA app outwardly looks identical to RMTA, jobs are submitted to the OSG by CyVerse on behalf of the user, while also automating data management and transfer between the Data Store and OSG. RMTA is available as a Docker image <https://hub.docker.com/r/evolinc/osg-rmta/> for easy installation in a command line environment (e.g. XSEDE or PC) where Docker is already installed or where the user has the necessary privileges to install Docker. Additionally, Docker can run within Singularity (Kurtzer et al., 2017), which enables launching RMTA within an HPC environment. Having RMTA packaged within a Docker container abrogates the need for installation of prerequisite software. For users with an OSG account and for whom a CyVerse account is unnecessary, OSG-RMTA is already present on the OSG as a Docker image for immediate use. A brief tutorial on how to use RMTA and OSG-RMTA in the command line and OSG, respectively, can be found in the README.md at (<https://github.com/Evolinc/RMTA>). Finally, a stripped down version of RMTA (few visible options) aimed at introducing undergraduates to the concepts of RNA-seq is also available in the Discovery Environment (RMTA\_Instructional) with instructions at ([https://wiki.cyverse.org/wiki/display/DEapps/RMTA\\_Instructional](https://wiki.cyverse.org/wiki/display/DEapps/RMTA_Instructional)).

## Additional Discovery Environment-Specific Features to Simplify Ribonucleic Acid Sequencing Analysis

Although RMTA and OSG-RMTA are packaged as Docker images for use outside of CyVerse's Discovery Environment (e.g. OSG or an HPC), we highly recommend using the Discovery Environment integrated RMTA apps to take advantage of both the Discovery Environment's GUI and CyVerse's integrated Data Store. The Data Store makes data

management relatively easy [drag 'n' drop as opposed to shipping hard drives to Amazon Web Services (AWS) (Zhao et al., 2013)]. A number of up-to-date genomes are available in the community Data Store and the Discovery Environment has an application programming interface (API) that can acquire any of the 50,000 additional genomes from CoGe (Lyons et al., 2008) or public/private databases if needed. A Discovery Environment app has also been developed to retrieve GTF and BAM files from subdirectories generated for each SRA (File\_Select v1.0) and place them into a single, user-specified folder, making data management even easier.

Researchers running OSG-RMTA in the Discovery Environment can take advantage of two features that facilitate a "divide and conquer" approach to job submission to the OSG. Long (> 1,000s) lists of SRAs can be divided up into smaller lists using the File\_Split v1.0 app. The Discovery Environment's HT Analysis Path List file feature then uses these lists to parallelize their job submissions to the OSG. ([https://wiki.cyverse.org/wiki/display/TUT/Parallel+execution,+DE+\(Discovery+Environment\)+style](https://wiki.cyverse.org/wiki/display/TUT/Parallel+execution,+DE+(Discovery+Environment)+style)). Thus, a thousand SRAs can be processed in roughly the same time it would take to process 100. All of this happens with a few clicks of a button.

## Data and Software Availability

RMTA and OSG-RMTA are freely available to use as an app on CyVerse's Discovery Environment or on the Open Science Grid (<https://hackmd.io/s/rJrqyAAQ>). Detailed instructions on how to use RMTA in the Discovery Environment can be found at (<https://wiki.cyverse.org/wiki/display/DEapps/RMTA+v2.6.3>). Working within the Discovery Environment requires a modern hypertext markup language 5 (HTML5) capable browser and a free CyVerse user account ([user.cyverse.org](http://user.cyverse.org)). Users wishing to use OSG-RMTA on the OSG directly (not through the Discovery Environment) will need an account (<http://osgconnect.net/>). The source code of the workflow is available at <https://github.com/Evolinc/RMTA> and <https://github.com/Evolinc/OSG-RMTA> and the Docker images for users wishing to adapt RMTA to novel environments are available at <https://hub.docker.com/r/evolinc/rmta> and <https://hub.docker.com/r/evolinc/osg-rmta/>. Test data for RMTA are present in the Discovery Environment and on GitHub.

## Data Visualization in EPIC-CoGe, Long Non-Coding Ribonucleic Acid Identification With Evolinc, and Analysis of Gene Expression

Two sorted.bam files, SRR2240264 (flower) and SRR2240265 (root) from an RMTA run on 100 paired-end (PE) SRAs were uploaded to EPIC-CoGe from CyVerse's Data Store using the LoadExp+ tool (Grover et al., 2017) in CoGe. Expression data were associated with the *Arabidopsis thaliana* (Col-0) genome (v10.02, id 16911). These two datasets are publicly available in the RMTA folder (id 2568) at [www.genomevolution.org](http://www.genomevolution.org). To identify lncRNAs, all 100 "filtered.gtf" files from the 100 PE RMTA analyses were added to an HTPathlist file in the Discovery Environment. This HTPathlist file was then used as the input



for a single Evolinc analysis in the Discovery Environment (Evolinc v1.7.5; Nelson et al., 2017). The updated annotation file from each Evolinc job were merged using the Evolinc\_merge app (v1.0). FASTA sequence for all identified lncRNAs were extracted using the gffread utility in the Cufflinks package. GC content and length of all *Arabidopsis* protein-coding genes and the newly identified lncRNAs were calculated using a custom Perl script (**File S1**). Principal component analyses were generated in R (code in **File S2**) using the read count data from RMTA.

## RESULTS

To demonstrate the utility of RMTA, we used our workflow (**Figure 2**) to process 1,000 *A. thaliana* SRAs (single-end reads) and 100 SRAs (paired-end reads) directly from NCBI's data repository, representing 1.27 terabases of RNA-seq data (**Table 2** and **Table S1**). SRA IDs were obtained from NCBI's SRA by searching the term "*Arabidopsis thaliana*" and then exporting all summary results to a tab-delimited file using NCBI's "Send to" API. For downstream analysis of the PE data, specific RNA-seq data from root ( $n = 68$ ) and flower ( $n = 32$ ) tissue were chosen from these summary results (**Table S1**). PE and single-end (SE) SRA IDs were copied into new list files in the Discovery Environment, partitioned into lists of 10 and 100, respectively, using File\_Split-1.0. These 10 list files were then added to an HT Analysis Path List that subsequently became the input for the RMTA app. Two analyses were launched in the Discovery Environment (one for PE and one for SE) whereupon they were automatically divided up and submitted simultaneously as 10 jobs each. Specific options selected for these analyses were: HISAT2 for the aligner, a FPKM, and coverage cut-off threshold of 1, and Run FastQC selected. All other options were left as default.

Mapping rates and time to completion are shown in **Table 2** and **Table S1**. While the mapping rates for most (76%) of the PE SRAs were >90% (avg = 92.7%, **Table 2**), six SRAs displayed rates <75%. FastQC results were interrogated to identify potential reasons for why these mapping rates might be low and if 5' or 3' trimming of reads might facilitate better mapping. FastQC results revealed a significant enrichment of adaptor sequence for these samples. A subsequent relaunching of

RMTA with 15 nts trimmed from the 5' end (an option within RMTA) resulted in improved mapping rates for all six samples. This demonstrates the utility of being able to analyze hundreds of SRAs at once with default settings, and then follow up with adjusted parameters for problematic samples.

We then demonstrated three ways in which RMTA can facilitate downstream analysis: 1) by visualizing the RMTA generated BAM files in the EPIC-CoGe genome browser, 2) by testing for variation between datasets using the RMTA featureCounts output, and 3) identifying novel lncRNAs using RMTA's filtered genome annotation output. Users often wish to sanity check their RNA-seq data by viewing them in a genome browser. A benefit of performing RMTA in CyVerse's Discovery Environment is the ability to immediately import the large mapped read (BAM) files from the Data Store into the EPIC-CoGe genome browser (Nelson et al., 2018). Genomes for over 19,000 organisms are available on CoGe, meaning that the user will not only be able to visualize their RNA-seq, but can also import genomes from CoGe into the Discovery Environment to supplement the genomes already available. Two of the *Arabidopsis* PE-SRAs were imported into EPIC-CoGe (publicly available in the CoGe folder "RMTA," ID: 2568) for public browsing (**Figure 3A**). For users performing RMTA locally (i.e., in a Docker container), genome browsers such as IGV (J. T. Robinson et al., 2011) are freely available and easy to use.

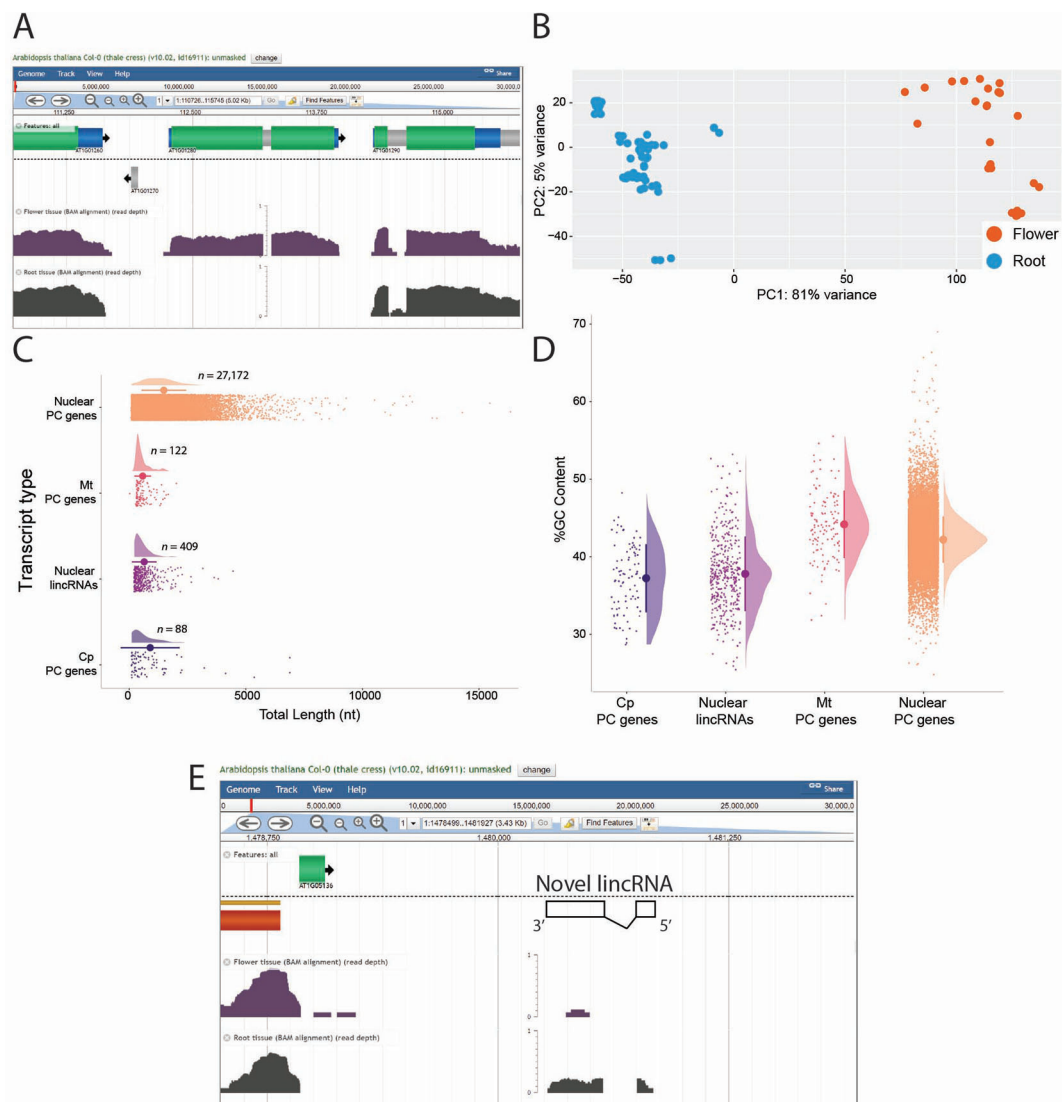
Sample variation within the 100 *Arabidopsis* PE-SRAs, consisting of RNA-seq data from 68 root and 32 flower samples (see **Table S1** for IDs) was examined using the RMTA-produced table of exon associated read counts (feature\_counts.txt). While these analyses can occur using Discovery Environment RStudio VICE app deployments of DESeq2 (Love et al., 2014) or EdgeR (M. D. Robinson et al., 2010); [https://learning.cyverse.org/projects/vice/en/latest/user\\_guide/quick-rstudio.html](https://learning.cyverse.org/projects/vice/en/latest/user_guide/quick-rstudio.html), as the output file from featureCounts is small and manageable, it was downloaded and manipulated in a local R environment (i.e., RStudio; RStudio Team, 2015; R-code available in **File S2**). A principal component analysis (PCA) demonstrated that, as expected, the largest amount of variation between samples (PC1) could be explained by tissue (**Figure 3B**). This analysis demonstrates the ease with which researchers can validate their RNA-seq data and proceed to differential expression analyses using the RMTA workflow.

The filtered genome annotation produced by RMTA is perfect for novel gene identification without any additional data transformation. To describe this, the RMTA output file "filtered.gtf," with transcripts with an FPKM or read/base <3 removed, was used as input in the long non-coding RNA identification pipeline, Evolinc (v1.7.5; Nelson et al., 2017) in the Discovery Environment. Like RMTA, Evolinc is also packaged as a Docker image for local discovery. Evolinc was used to identify putative lncRNAs expressed in the root and flower RNA-seq data. The number of lncRNAs identified and some basic characteristics, such as average length and GC content relative to nuclear and organellar protein-coding genes, are shown in **Figures 3C, D**, with the R code necessary to recapitulate the images available in **File S3**. Reads mapped to these lncRNAs, and other novel genes, can also be visualized in a

**TABLE 2 |** Mapping rates and time to completion for the example read mapping and transcript assembly (RMTA) analyses.

	Mapping rates				Gbp mapped	Mbp/minute
	> 90%	90–75%	75–50%	<50%		
SE samples	63%	16%	9%	12%	863	45
PE samples	76%	15%	5%	4%	406	26

RMTA was used to process 100 paired-end (PE) and 1,000 single-end (SE) *Arabidopsis* sequence read archives (SRAs). The percentage of these SRAs with mapping rates >90%, 90–75%, etc., are shown. Gbp =  $1 \times 10^9$  base pairs mapped. Mbp/minute = million base pairs mapped per minute.



**FIGURE 3 |** Examples of downstream analyses facilitated by the read mapping and transcript assembly (RMTA) workflow. The output generated by RMTA are immediately useful for the usual analyses performed following an RNA-sequencing experiment. **(A)** EPIC-CoGe screenshot of *Arabidopsis* root and flower RNA-sequencing (RNA-seq) data processed by RMTA highlighting a gene, *AT1G01280*, that is highly expressed in flower tissue but not roots. **(B)** Principal component analysis (PCA) of 100 *Arabidopsis* sequence read archives (SRAs) generated in R using the read count output file from RMTA. **(C)** Comparison of the length of Evolinc identified long non-coding RNA (lncRNAs) relative to other nuclear and organellar genes. PC, protein-coding gene; Mt, mitochondria; Cp, chloroplast. **(D)** Comparison of GC content of Evolinc identified lncRNAs relative to other nuclear and organellar genes. **(E)** EPIC-CoGe visualization of the expression of a locus identified by Evolinc as a lncRNA. The boundaries of the lncRNA and its orientation have been added to the EPIC-CoGe screenshot.

genome browser using the BAM files generated by RMTA (**Figure 3E**). In sum, RMTA is not only a simple and intuitive means of processing large amounts of RNA-seq data, but also facilitates commonly performed downstream analyses.

## DISCUSSION

As the technical and financial barriers to generating raw RNA-seq data are reduced, the barrier to discovery will be shifted

toward the computational steps required to analyze those data and the integration with other software for extracting high-value knowledge and novel scientific insights. RMTA was designed with the goal of alleviating many of the tedious or time consuming steps of RNA-seq processing and downstream data analysis. This goal was primarily accomplished by incorporating the three main steps of RNA-seq processing (read mapping, assembly, and counting) into a very approachable, yet scalable and interoperable tool, and ensure that the output files from RMTA are easily ingested by other platforms and analysis tools.

The usefulness of RMTA is most apparent when utilized within CyVerse's Discovery Environment. Access to public or private genomes (through CoGe and the Data Store), automatic data retrieval from NCBI's SRA, data management through the Data Store, and job submission within the Discovery Environment or direct to the Open Science Grid, means that data scientists can perform all of their analyses in the cloud. In addition, users can take advantage of parallelizable job submission options that are available to divide and conquer their large datasets. Once finished, RMTA produces output files that are ready for immediate analysis (e.g., differential expression), visualization (e.g., in a genome browser), or novel gene identification (e.g., long non-coding RNAs), all of which can also occur in the cloud. In sum, large-scale RNA-seq analysis is no longer limited to data scientists with HPC access or a high-end local computer.

RMTA was also designed for users who prefer to perform analyses locally. By packaging RMTA in a Docker container we have removed the tedious task of installing prerequisite software and made RMTA capable of running on any operating system. Thus, processing and analysis of RNA-seq data is no longer restricted to a Linux machine but can now also be performed on a machine utilizing Windows or Mac OS. In addition, recognizing data storage limitations, RMTA removes unnecessary files generated during the analysis that would rapidly fill up most storage allotments.

Not all data scientists have the same needs in terms of available features or in the amount of data to be processed. To this end we developed variants of RMTA targeting undergraduate or high school instructors (RMTA\_instructional), users processing 1–100s of data files (RMTA), and users processing 1,000s or more data files (OSG-RMTA). RMTA\_instructional is available as an app in the Discovery Environment with minimal fields exposed, with example input files added to the appropriate fields, and with entry level descriptions of the purpose behind each field. RMTA and OSG-RMTA are available as both Discovery Environment apps and Docker images, with OSG-RMTA already available on the OSG. RMTA and OSG-RMTA offer the same features, differing only in where and how jobs are submitted.

In summary, RMTA opens up the task of RNA-seq processing and data analysis to anyone with access to a web browser, thereby democratizing data discovery. It also enables analysis of all transcripts, not just the ones matching already annotated genes, thus encouraging a more inclusive view of what genomic regions are actually transcribed. Finally, RMTA serves as a useful tool for savvy data scientists wishing to reduce the time and effort necessary to process large data sets.

## REFERENCES

- Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., et al. (2016). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 44 (W1), W3–10. doi: 10.1093/nar/gkw343
- Anderson, S. J., Kramer, M. C., Gosai, S. J., Yu, X., Vandivier, L. E., Nelson, A. D. L., et al. (2018). N6-Methyladenosine inhibits local

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. Accession numbers can be found here: **Table S1**.

## AUTHOR CONTRIBUTIONS

ADN, UD, EL, and BG designed the workflow. ADN, UD and SP wrote the code. SR and UD integrated RMTA into the OSG. SP, IK, LM, AN, AR, and RM tested RMTA, resolved bugs, and wrote the tutorials. SP, IK, LM, AN, AR, and RM analyzed the data. SP, RM, EL, and ADN wrote the manuscript. All authors read and approved the manuscript.

## FUNDING

This work has been supported by the National Science Foundation grants IOS—1758532 (to ADN, RM, and UD), IOS—1743442 (to CyVerse), IOS—1444490 (to EL and BG), NSF Research Experience for Undergraduates (REU to IK, LM, and AN), and an NSF Research Assistantship for High School Students (RAHSS to AR). As AR is a minor, parental consent has been given to include her as an author on this manuscript.

## ACKNOWLEDGMENTS

We would like to thank CyVerse for technical advice and application implementation. We would like to thank Jennifer Meneghin for posting her custom Perl script to the web in 2010. We would also like to thank Dr. Nirav Merchant (University of Arizona) for feedback on the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.01361/full#supplementary-material>

**TABLE S1** | SRA IDs, tissue information, and mapping rates of example data.

**FILE S1** | Script used to calculate transcript length and GC content.

**FILE S2** | R code necessary to recapitulate PCA of example PE data.

**FILE S3** | R code necessary to recapitulate visualization of data.

- ribonucleolytic cleavage to stabilize mRNAs in Arabidopsis. *Cell Rep.* 25, 1146–1157. doi: 10.1016/j.celrep.2018.10.020
- Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*, Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- Chet, B., Hanna, M., Salinas, E., Neff, J. C., Saksena, G., Livitz, D., et al. (2017). FireCloud, a scalable cloud-based platform for collaborative genome analysis: strategies for reducing and controlling costs. *bioRxiv*. doi: 10.1101/209494

- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-Seq Aligner. *Bioinformatics* 29 (1), 15–21. doi: 10.1093/bioinformatics/bts635
- Folarin, A. A., Dobson, R. J., and Newhouse, S. J. (2015). NGSeasy: a next generation sequencing pipeline in Docker containers. *F1000Res*. doi: 10.12688/f1000research.7104.1
- Galpaz, N., Gonda, I., Shem-Tov, D., Barad, O., Tzuri, G., Lev, S., et al. (2018). Deciphering genetic factors that determine melon fruit-quality traits using RNA-Seq-based high-resolution QTL and eQTL mapping. *Plant J. Cell Mol. Biol.* 94 (1), 169–191. doi: 10.1111/tpj.13838
- Gosai, S. J., Foley, S. W., Wang, D., Silverman, I. M., Selamoglu, N., Nelson, A. D. L., et al. (2015). Global analysis of the RNA-protein interaction and RNA secondary structure landscapes of the Arabidopsis nucleus. *Mol. Cell* 57 (2), 376–388. doi: 10.1016/j.molcel.2014.12.004
- Grover, J. W., Bomhoff, M., Davey, S., Gregory, B. D., Mosher, R. A., and Lyons, E. (2017). CoGe LoadExp+: a web-based suite that integrates next-generation sequencing data analysis workflows and visualization. *Plant Direct* 1 (2), 1–7. doi: 10.1002/pld3.8
- Gusev, A., Lawrenson, K., Lin, X., Lyra, P. C. Jr, Kar, S., Vavra, K. C., et al. (2019). A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. *Nat. Genet.* 51 (5), 815–823. doi: 10.1038/s41588-019-0395-x
- Jensen, T. L., Frasketi, M., Conway, K., Villaruel, L., Hill, H., Krampis, K., et al. (2018). RSEQREP: RNA-Seq Reports, an open-source cloud-enabled framework for reproducible RNA-Seq data processing, analysis, and result reporting. *F1000Res*. 2 (2162). doi: 10.12688/f1000research.13049.2
- Kurtz, G. M., Sochat, V., and Bauer, M. W. (2017). Singularity: scientific containers for mobility of compute. *PLoS One* 12 (5), e0177459. doi: 10.1371/journal.pone.0177459
- Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., et al. (2018). Massive mining of publicly available RNA-Seq data from human and mouse. *Nat. Commun.* 9 (1), 1366. doi: 10.1038/s41467-018-03751-6
- Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9 (4), 357–359. doi: 10.1038/nmeth.1923
- Lau, J. W., Lehnert, E., Sethi, A., Malhotra, R., Kaushik, G., Onder, Z., et al. (2017). The cancer genomics cloud: collaborative, reproducible, and democratized—a new paradigm in large-scale computational research. *Cancer Res.* 77 (21), e3–e6. doi: 10.1158/0008-5472.CAN-17-0387
- Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. *Nucleic Acids Res.* 39 (Database issue), D19–D21. doi: 10.1093/nar/gkq1019
- Li, S., Yamada, M., Han, X., Ohler, U., and Benfey, P. N. (2016). High-resolution expression map of the Arabidopsis root reveals alternative splicing and lincRNA regulation. *Dev. Cell* 39 (4), 508–522. doi: 10.1016/j.devcel.2016.10.012
- Liao, Y., Smyth, G. K., and Shi, W. (2014). Feature Counts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30 (7), 923–930. doi: 10.1093/bioinformatics/btt656
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi: 10.1186/s13059-014-0550-8
- Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., et al. (2008). Finding and comparing syntenic regions among Arabidopsis and the outgroups Papaya, Poplar, and Grape: CoGe with Rosids. *Plant Physiol.* 148 (4), 1772–1781. doi: 10.1104/pp.108.124867
- Martin, L. B. B., Fei, Z., Giovannoni, J. J., and Rose, J. K. C. (2013). Catalyzing plant science research with RNA-Seq. *Front. Plant Sci.* 4, 66. doi: 10.3389/fpls.2013.00066
- Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., et al. (2016). The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biol.* 14 (1), e1002342. doi: 10.1371/journal.pbio.1002342
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5 (7), 621–628. doi: 10.1038/nmeth.1226
- Nelson, A. D. L., Devisetty, U. K., Palos, K., Haug-Baltzell, A. K., Lyons, E., and Beilstein, M. A. (2017). Evolinc: a tool for the identification and evolutionary comparison of long intergenic non-coding RNAs. *Front. Genet.* 8, 52. doi: 10.3389/fgene.2017.00052
- Nelson, A. D. L., Haug-Baltzell, A. K., Davey, S., Gregory, B. D., and Lyons, E. (2018). EPIC-CoGe: managing and analyzing genomic data. *Bioinformatics* 34 (15), 2651–2653. doi: 10.1093/bioinformatics/bty106
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14 (4), 417–419. doi: 10.1038/nmeth.4197
- Perteau, M., Kim, D., Perteau, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-Seq experiments with HISAT, stringtie and ballgown. *Nat. Protoc.* 11 (9), 1650–1667. doi: 10.1038/nprot.2016.095
- Pordes, R., Petravick, D., Kramer, B., Olson, D., Livny, M., Roy, A., et al. (2007). The open science grid. *J. Physics. Conf. Ser.* 78, 012057. doi: 10.1088/1742-6596/78/1/012057
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26 (1), 139–140. doi: 10.1093/bioinformatics/btp616
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29 (1), 24–26. doi: 10.1038/nbt.1754
- RStudio Team (2015). *RStudio: integrated development for R*. RStudio, Inc (Boston, MA).
- Schlackow, M., Nojima, T., Gomes, T., Dhir, A., Carmo-Fonseca, M., and Proudfoot, N. J. (2017). Distinctive patterns of transcription and RNA processing for human lincRNAs. *Mol. Cell* 65 (1), 25–38. doi: 10.1016/j.molcel.2016.11.029
- Srivastava, A., Malik, L., Sarkar, H., Zakeri, M., Sonesson, C., Love, M. I., et al. (2019). Alignment and mapping methodology influence transcript abundance estimation. *BioRxiv*, 657874. doi: 10.1101/657874
- Thain, D., Tannenbaum, T., and Livny, M. (2005). Distributed computing in practice: the Condor experience. *Concurr. Comput.* 17, 323–356. doi: 10.1002/cpe.938
- Zhao, S., Prenger, K., and Smith, L. (2013). Stormbow: a cloud-based tool for reads mapping and expression quantification in large-scale RNA-Seq studies. *ISRN Bioinf.*, 481545S. doi: 10.1155/2013/481545

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Peri, Roberts, Kreko, McHan, Naron, Ram, Murphy, Lyons, Gregory, Devisetty and Nelson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Integrative Differential Expression Analysis for Multiple EXperiments (IDEAMEX): A Web Server Tool for Integrated RNA-Seq Data Analysis

Verónica Jiménez-Jacinto<sup>1</sup>, Alejandro Sanchez-Flores<sup>1\*</sup> and Leticia Vega-Alvarado<sup>2\*</sup>

<sup>1</sup> Unidad Universitaria de Secuenciación Masiva y Bioinformática, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Mexico, <sup>2</sup> Instituto de Ciencias Aplicadas y Tecnología, Universidad Nacional Autónoma de México, Mexico City, Mexico

## OPEN ACCESS

### Edited by:

Monica Bianchini,  
University of Siena, Italy

### Reviewed by:

Zeeshan Ahmed,  
University of Connecticut,  
United States  
Gaurav Sablok,  
Finnish Museum of Natural History,  
Finland

### \*Correspondence:

Alejandro Sanchez-Flores  
alexsf@ibt.unam.mx  
Leticia Vega-Alvarado  
leticia.vega@icat.unam.mx

### Specialty section:

This article was submitted to  
Bioinformatics and Computational  
Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 15 December 2018

**Accepted:** 13 March 2019

**Published:** 29 March 2019

### Citation:

Jiménez-Jacinto V,  
Sanchez-Flores A and  
Vega-Alvarado L (2019) Integrative  
Differential Expression Analysis  
for Multiple EXperiments (IDEAMEX):  
A Web Server Tool for Integrated  
RNA-Seq Data Analysis.  
Front. Genet. 10:279.  
doi: 10.3389/fgene.2019.00279

The current DNA sequencing technologies and their high-throughput yield, allowed the thrive of genomic and transcriptomic experiments but it also have generated big data problem. Due to this exponential growth of sequencing data, also the complexity of managing, processing and interpreting it in order to generate results, has raised. Therefore, the demand of easy-to-use friendly software and websites to run bioinformatic tools is imminent. In particular, RNA-Seq and differential expression analysis have become a popular and useful method to evaluate the genetic expression change in any organism. However, many scientists struggle with the data analysis since most of the available tools are implemented in a UNIX-based environment. Therefore, we have developed the web server IDEAMEX (Integrative Differential Expression Analysis for Multiple EXperiments). The IDEAMEX pipeline needs a raw count table for as many desired replicates and conditions, allowing the user to select which conditions will be compared, instead of doing all-vs.-all comparisons. The whole process consists of three main steps (1) Data Analysis: that allows a preliminary analysis for quality control based on the data distribution per sample, using different types of graphs; (2) Differential expression: performs the differential expression analysis with or without batch effect error awareness, using the bioconductor packages, NOISeq, limma-Voom, DESeq2 and edgeR, and generate reports for each method; (3) Result integration: the obtained results the integrated results are reported using different graphical outputs such as correlograms, heatmaps, Venn diagrams and text lists. Our server allows an easy and friendly visualization for results, providing an easy interaction during the analysis process, as well as error tracking and debugging by providing output log files. The server is currently available and can be accessed at <http://www.uusmb.unam.mx/ideamex/> where the documentation and example input files are provided. We consider that this web server can help other researchers with no previous bioinformatic knowledge, to perform their analyses in a simple manner.

**Keywords:** bioinformatics, RNA-Seq, differential expression, NGS, transcriptomics



## INTRODUCTION

Transcriptomics experiments have been used widely to measure the RNA levels expressed in tissues or cells from practically any organism. This approach has been used since the implementation of Northern blots hybridization analysis and was scaled up by the development of microarray technology. However, transcriptomics has been improved with the aid of sequencing technologies which recently have been replacing microarrays by using RNA sequencing (RNA-Seq) experiments to evaluate gene expression at a genome-wide scale. Therefore, either microarrays or RNA-Seq technologies have generated a massive amount of data results that demands *ad hoc* methods to fully analyze and compare gene expression between different conditions, tissues or cell populations for a given organism.

To quantify the transcription levels and identify differential expressed genes under different conditions, using RNA-Seq data from high-throughput sequencing technologies, a general workflow can be described: (1) quality control of RNA-Seq reads (Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data); (2) read trimming or filtering (Chen et al., 2017; Roser et al., 2018); (3) mapping trimmed/filtered reads to a reference (genome or transcriptome) (Li and Durbin, 2009; Langmead and Salzberg, 2012; Kim et al., 2013; Wu et al., 2016); (4) obtaining the read count for each gene (Quinlan and Hall, 2010; Li and Dewey, 2011; Roberts et al., 2011) and (5) differential expression analysis (Anders and Huber, 2010; Tarazona et al., 2011; McCarthy et al., 2012; Love et al., 2014; Ritchie et al., 2015). Currently, due to the size of datasets, steps 1 to 4 have to be performed by the user and many tools for each step are available and have been widely used and cited elsewhere. However, the differential expression analysis is probably the most important step that allows the user to interpret the biological information regarding the expression profiles of a given organism under different conditions.

The gene expression profile contains the information regarding genes related to the organism response to a certain condition. To retrieve such information, the differential expression analysis has to be performed and it requires statistical methods to differentiate between expression changes due to the tested conditions and biological “noise” or variability. Currently, several computational tools have been developed mainly in the programming language R and packages are available at the Bioconductor project repository (Huber et al., 2015). However, R language and packages have to be used mainly through a UNIX-based operating system and by command-line instructions which requires a certain level of programming skills. Therefore, non-bioinformatics researchers demand either a Graphical User Interface (GUI) in order to use differential expression tools or web-based applications. A GUI-based solution still requires a local installation of all packages needed for the differential expression analysis and this could remain challenging. The web-based applications are now emerging

(de Jong et al., 2015; Monier et al., 2018; Zhang et al., 2018) as friendlier option to perform the differential expression analysis in a more friendly way and without installing software in a local computer.

Here, we introduce the IDEAMEX web server (Integrative Differential Expression Analysis for Multiple EXperiments) that uses as input an RNA-Seq raw count table in text format and generates results using bioconductor packages NOISeq, limma-voom, DESeq2 and edgeR. These packages have been constantly benchmarked and presented the most reliable results with different datasets and gold-standards (Seyednasrollah et al., 2015; Costa-Silva et al., 2017). In this work, we demonstrate the functionality of IDEAMEX, using RNA-Seq data from a previous publication (Olvera et al., 2017) where the differential expression analysis in tilapia liver was performed, in addition to other datasets used as examples to test the website.

Our server has been used in several projects and has been visited from different world-wide locations as recorded in our site tracker. IDEAMEX is available and can be accessed at <http://www.uusmb.unam.mx/ideamex/> where the documentation and example input files are provided. Our server offers a web server-based analysis that can help researchers with no previous bioinformatic knowledge, to perform their transcriptomic analyses in a simple manner, in order to interpret the biological data contained in their RNA-Seq experiments.

## MATERIALS AND METHODS

### Web Server Description

The web page is hosted by the “Unidad Universitaria de Secuenciación Masiva y Bioinformática” core lab facility, at the “Instituto de Biotecnología” of the “Universidad Nacional Autónoma de México, Campus Morelos located in Cuernavaca, Morelos, México.” A Linux box computer with Ubuntu 14.04 LTS with the following hardware main characteristics: Intel Core i7 4770 processor; 32 Gbytes of DDR3 RAM and 1 Tbyte of hard disk storage.

The deployment was implemented using the Apache HTTP server version 2.4.7 with a PHP v5.5.9 front-end that coordinates the writing of the input and output files to a SQL database through a POSTGRES Relational Data Base Manager (RDBM) server (psql version 9.3.22. The installed R version is 3.5.2. The web server can be accessed at <http://www.uusmb.unam.mx/ideamex/>.

The web server interface has been tested using different web browsers and different operative systems. Using Microsoft Windows 10: Microsoft EdgeHTML 17.17134; Google Chrome version 72.0.3626.109 (Official Build) (64-bit); Mozilla Firefox Quantum 63.0 (64-bit). Using MacOS Sierra 10.13.6: Safari 12.0.3; Google Chrome 71.0.3578.98 (64-bit). Using Linux Ubuntu 16.04 LTS: Mozilla Firefox Quantum 65.0.

Additionally, the scripts and binaries used in the web server can be found in the public repository <https://github.com/leticiaVega/IDEAMEX>

## RNA-Seq Examples and Data From Tilapia Liver Experiment

We used as example to test our website two datasets. The first example contains data from the Pasilla Bioconductor library (Brooks et al., 2010), taking in account only the gene level counts. This dataset contains RNA-Seq count data for treated and untreated cells from the S2-DRSC cell line. The second example file which can be used to test the batch effect error awareness, was taken from the NBPSec CRAN package (Di et al., 2014). This dataset contains the *Arabidopsis thaliana* RNA-Seq data (Cumbie et al., 2011), comparing  $\Delta$ hrcC challenged and mock-inoculated samples. In this case, the samples were collected in three batches.

We also obtained RNA-Seq publicly available data already reported (Olvera et al., 2017) that was generated to determine the effect of 3,5-di-iodothyronine (T2) and 3,5,3'-tri-iodothyronine (T3) exogenous treatment on the transcriptome of tilapia (*Oreochromis niloticus*) liver. For control and each hormone treatment, two biological replicates were generated. The FASTQ raw data can be found under the following SRA identifiers: SRX2630485, SRX2630486, SRX2630487, SRX2630488, SRX2630489, and SRX2630490.

Briefly, the quality control (QC) and filtering for the raw data was performed using the FASTQC software (Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data) and contamination and adapter removal was carried out using in-house Perl scripts. QC'd reads were mapped using the Bowtie 1.1.234 aligner (Langmead et al., 2009) to the annotated *Oreochromis niloticus* (Orenil1.0.cds.all, 21,437 coding genes) CDS dataset downloaded from Ensembl repository database (Aken et al., 2016) using the BioMart utility. Quantification and repetitiveness normalization were carried out using eXpress software 1.535 (Roberts et al., 2011). Total effective counts for each sample were merged; a matrix was generated using the "abundance\_estimates\_to\_matrix.pl" Perl script included in the Trinity pipeline (Grabherr et al., 2011; Roberts et al., 2011). The resulting matrix was used as input for the differential expression analysis in the IDEAMEX web server. The select parameters were:  $p\text{-adj}/\text{FDR} = 0.05$ ;  $\log\text{FC} = 2$ ;  $\text{CPM} = 1$ .

## Differential Expression Packages

Based on the parameters defined by the user, 4 different R (version 3.5.2) packages for differential expression analysis are run: edgeR version 3.24.3 (Anders and Huber, 2010), using TMM normalization method (works with or without replicates); limma-Voom version 3.38.3 (Ritchie et al., 2015), using log2-counts per million normalization method (works with replicates only); DESeq2 version 1.22.2 (Love et al., 2014), with DESeq2-default normalization method (works with or without replicates) and NOISeq version 2.26.1 (Tarazona et al., 2011), with TMM normalization method (works with or without replicates). Other packages used in the server are: VennDiagram 1.6.20; ggplot2 3.1.0; UpSetR 1.3.3; corrplot 0.84 and ComplexHeatmap 1.20.0. The packages can change depending on the R programming language version, but all changes are reported to the user in log

files that contain all details about the commands and parameters used for the analysis.

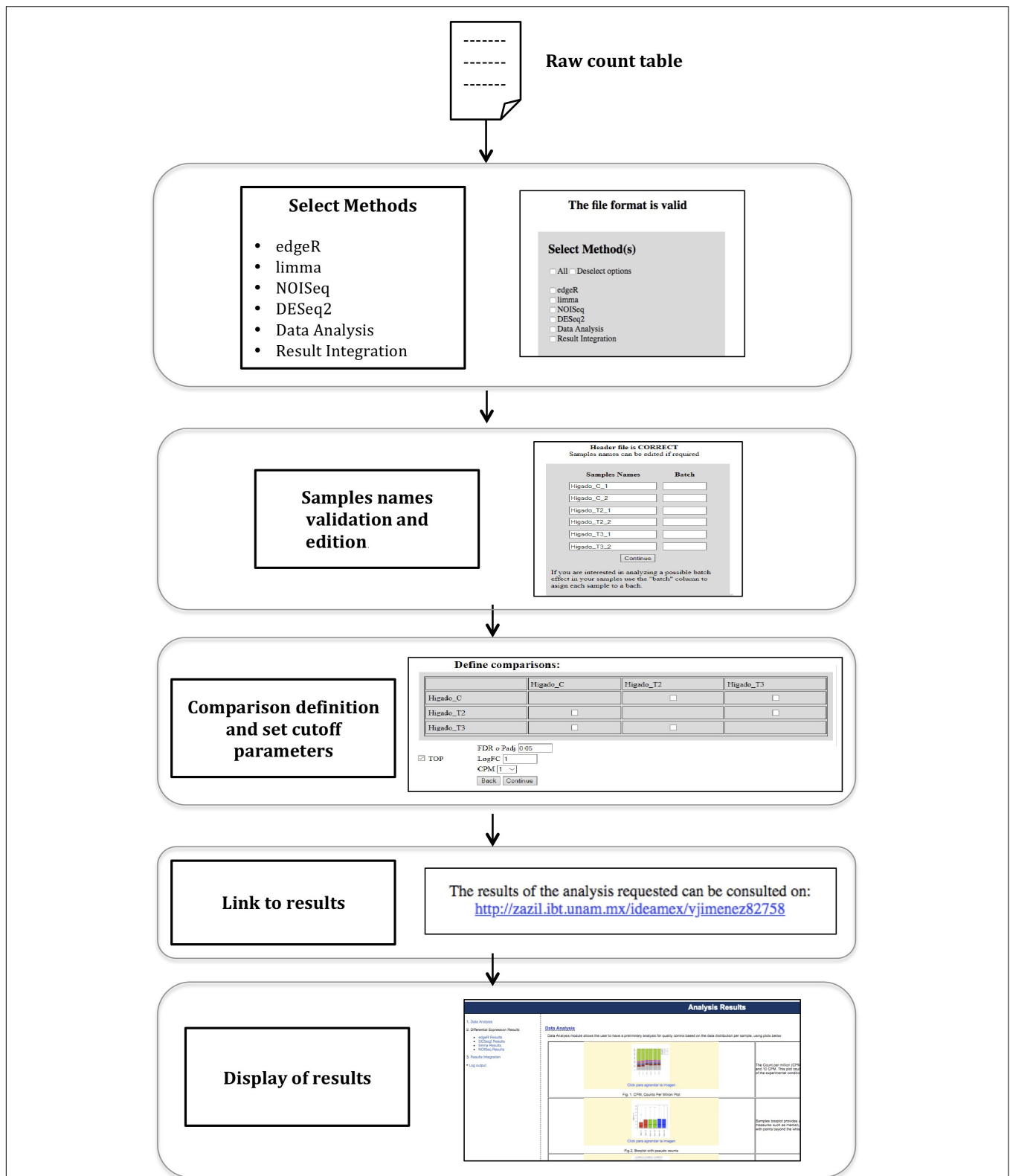
## RESULTS

### The IDEAMEX Web Server Implementation

The general workflow used in the IDEAMEX web server can be observed in **Figure 1**. First, the user has to enter a valid email address that will be used to report the follow up or the differential expression analysis to the user. In a nutshell, the pipeline starts with a raw count table for as many desired replicates and conditions, allowing the user to select which conditions will be compared, instead of doing all-vs.-all comparisons. After the web server validates the input format, the user can edit the sample names select one or more differential expression methods and the parameters to filter results. Additionally, the user can indicate if the samples belong to different batches so the selected differential expression methods, can correct any possible batch effect. Then, the data analysis step is performed where a preliminary quality control report is generated, based on the data distribution per sample. Next, the differential expression analysis is performed using one or more selected methods. Finally, the result from the different selected methods are integrated and are reported using Venn diagrams, a upset bar plot graph and text files for further filtering and analysis. Several additional plots are generated including correlograms to check the consistency between some calculations and heatmaps. Further details and study cases for dataset examples are described in the IDEAMEX User Manual that can be downloaded from the website. To demonstrate the functionality of our web server, we used a dataset generated from an RNA-Seq experiment to compare the effect of thyroid hormones in tilapia liver (see Materials and Methods).

Optionally, the user can perform a full registration at the IDEAMEX homepage, in order to keep track of all projects results. The sample name format should have a suffix\_[0-9] structure: nameCond1\_1, nameCond1\_2, ..., nameCond1\_n, nameCond2\_1, nameCond2\_2, ..., nameCond2\_m. Once the input file is validated, the server can infer the replicates from the suffix before the underscore symbol and the replicate number will be the digit after the underscore symbol. However, during the input loading, the user can edit these names. In case of samples being prepared in different batches, this information can be specified in the same window the sample names are edited. Indicating samples in different batches will turn on the batch effect error correction of different methods. Importantly, use this option only if you have knowledge of samples from a given condition, being prepared in a different batch which can give the experiment an extra variability. The user manual has a case of study for samples with batch effect.

In this work, the samples were named liverC\_1, liverC\_2 for replicates of control condition (no treatment) and liverT2\_1, liverT2\_2, liverT3\_1, liverT3\_2 for replicates that correspond to each of the 3,5-T2 and 3',3,5-T3 (T2 and T3) thyroid hormones treatments. A raw count table (**Supplementary Material S1**) in tab-separated text format, was generated and fed to the



**FIGURE 1 |** IDEAMEX workflow diagram. The web server workflow starts with the loading and validation of the raw count table as input. Then, the user selects one or more methods for differential expression analysis, data analysis and results integration. An optional step to edit the sample names is available. The user designs the comparison matrix by selecting which conditions will be compared. A link to the results is generated and after a few minutes, the results are presented in the Analysis Results web page.

**TABLE 1** | Raw count table example.

	LiverC_1	LiverC_2	LiverT2_1	LiverT2_2	LiverT3_1	LiverT3_2
ENSONIT00000002512	6.816486	5.866294	11.949044	7.285873	14.838847	7.979772
ENSONIT00000002995	0.000000	0.000000	0.001585	0.009734	0.000334	0.752950
ENSONIT00000006143	33.849657	109.674115	127.148250	141.191874	181.345619	132.397050
ENSONIT000000026691	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ENSONIT00000008087	74.458461	359.525580	149.166187	161.170914	235.990094	237.782394
ENSONIT000000021608	59.602367	101.722543	255.731580	259.076778	364.441300	329.630108
ENSONIT00000008926	0.000000	8.473091	33.032248	28.360464	21.724295	14.028806
ENSONIT00000011237	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ENSONIT000000021761	59.306830	135.526032	162.356881	113.464849	238.652733	233.360459

The sample names denote the condition naming and replicate number. *liverC\_N*, RNA-Seq counts for liver tissue with no treatment. *liverT2\_N*, RNA-Seq counts for liver tissue with T2 hormone treatment. *liverT3\_N*, RNA-Seq counts for liver tissue with T3 hormone treatment. N, replicate number. Raw count table should be in simple text format.

IDEAMEX web server. A snippet of the input raw count table is shown in **Table 1**.

## Input and Data Quality Control

The next step is to select the differential analysis method(s), the data quality analysis and the result integration by clicking on each box. It is recommended to click on the “select all” box to perform a full analysis. Afterward, the cut-off values for statistical confidence ( $p$ -adj and False Discovery Rate [FDR]), normalization (CPM) and transcript abundance difference (logFC) can be selected. Also, the comparison matrix can be defined to establish which samples or conditions will be compared.

A link to the Analysis Results web page will be generated, where the user results can find a link to the “(1) Data Analysis” section. A series of plots are displayed, allowing the user to have a preliminary analysis for quality control based on the data distribution per sample. All conditions defined in the raw count table are depicted as boxplots, CPM bar plots, density plots, principal components analysis (PCA) plots and multi-dimensional scaling (MDS) plots. Inspection and evaluation of these plots are essential steps for the interpretation of the differential expression analysis.

## CPM Plot Evaluation

In gene expression analysis, only a fraction of genes is expected to show differential expression between experimental conditions. The Count per million (CPM) plot shows the number of genes within each sample, having no counts (CPM = 0) or more than 1, 2, 5, or 10 CPM. This plot could help the user to decide the threshold to remove very low expressed genes in any of the experimental conditions. The default CPM cut-off value of 1 can be changed according to the user judgment, but it has to be done by re-running the analysis.

As observed in **Figure 2**, there is an increase of genes with CPM > 10 in the T2 and T3 samples, compared to the C condition. Also, the group of genes with CPM = 2 were decreased in T2 and T3 compared to the C condition. Approximately, ~70% of the genes presented no counts. This plot is the first glance to the expression profile for the compared conditions. For this

particular case, CPM = 1 is a convenient cut-off value which was the default option.

## Boxplot Evaluation

**Figure 3** presents the boxplots which provide an easy way to visualize the count distribution in each sample. If the count values distribution is highly skewed, then data transformation can be applied to roughly normalize the distribution. **Figure 3A** presents the log2 normalized data (pseudo-counts) and **Figure 3B** depicts the normalized data using the Trimmed Mean of  $M$ -values (TMM) method which is used for the differential expression analysis in edgeR and NOIseq packages. As observed, TMM normalization adjust the data according to the sequencing yield of each sample. The boxplot is an easy way to visualize the data distribution since it shows statistical measures such as median, quartiles, minimum and maximum values. Whiskers are also drawn extending beyond each end of the box with points beyond the whiskers typically indicating count outliers. In the log2 boxplot, the sequencing yield difference per sample is very evident. In this case, the control samples have fewer reads than the other samples. However, TMM normalization can fix this problem and this is why several differential expression methods have implemented this normalization procedure.

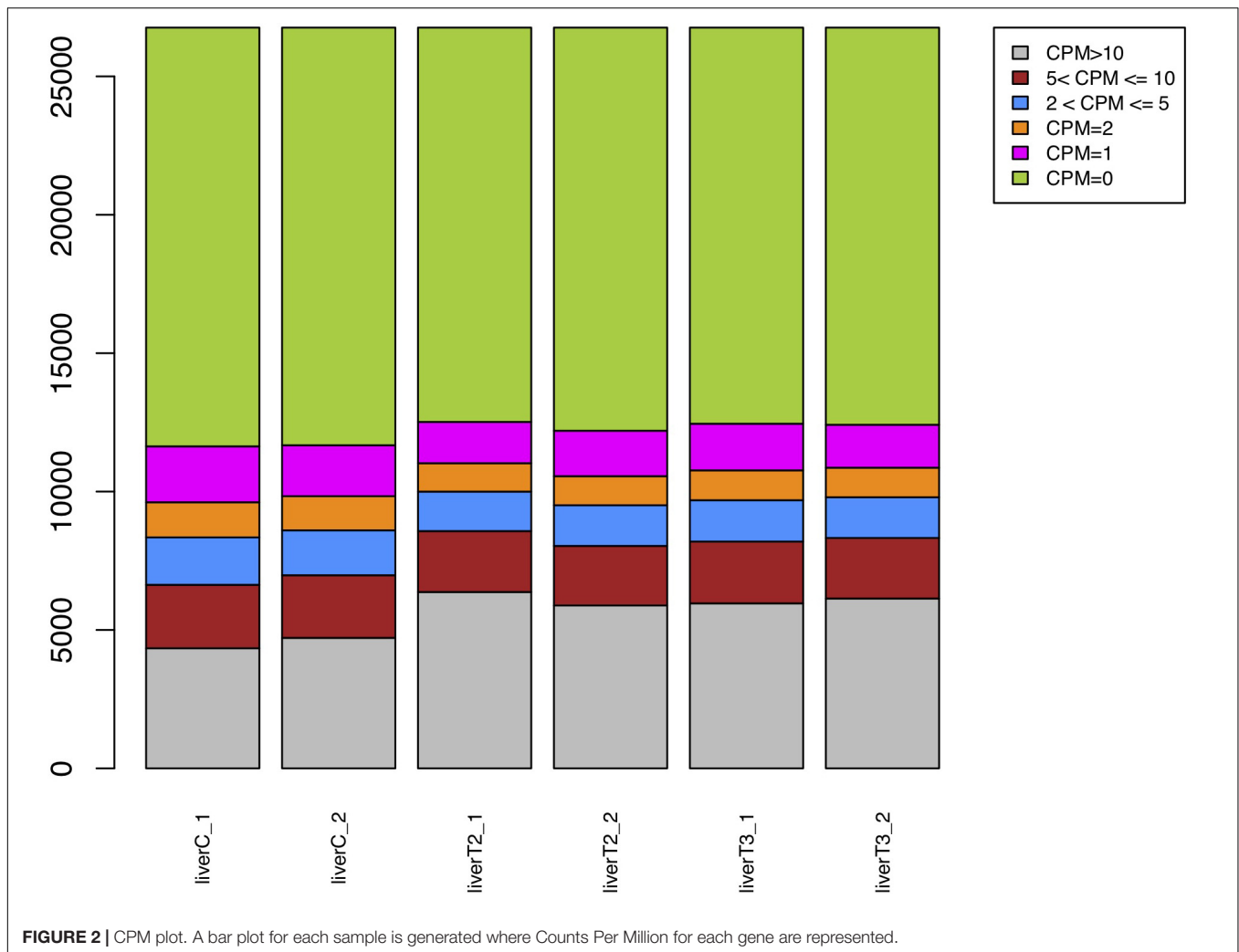
It is important to mention that the user will find a pair of boxplots, PCA and MDS graphs, since the data is plotted using pseudo-counts and TMM values.

## Density Plot Evaluation

The normalized count distributions can also be summarized by means of a density plot. Density plot provide more detail by enabling the detection of a dissimilarity in replicate count distribution. Ideally, the density plot for each replicate for a given condition, should greatly overlap indicating lower variability between replicates. **Figure 4** shows a density plot for the samples where replicates for the C condition, indicating certain dissimilarity in replicates for that condition.

## PCA Plot Evaluation

This type of plot is useful for visualizing the overall effect of experimental covariates and batch effects. In the context of RNA-Seq analysis, PCA shows groups of samples that ideally



will correspond to each condition. Clustering first by the most significant group, then by progressively less significant groups. **Figure 5** depicts how the 3 conditions (C, T2, and T3) form separate clusters, although some dispersion between replicates can be observed. This suggests that the variability among individuals was high, but due to the cluster separation it shouldn't affect the analysis. When a replicate is grouped with other samples from different conditions, it is recommended to remove it from the analysis if there are enough replicates left (at least two). Also, this plot could indicate if there is a batch effect problem, where samples in a same condition are very dispersed in the plot. In that case, the user can rerun the analysis indicating which samples could belong to a different batch. However, we recommend to confirm this with records from the preparation of the samples in the wet lab.

### MDS Plot Evaluation

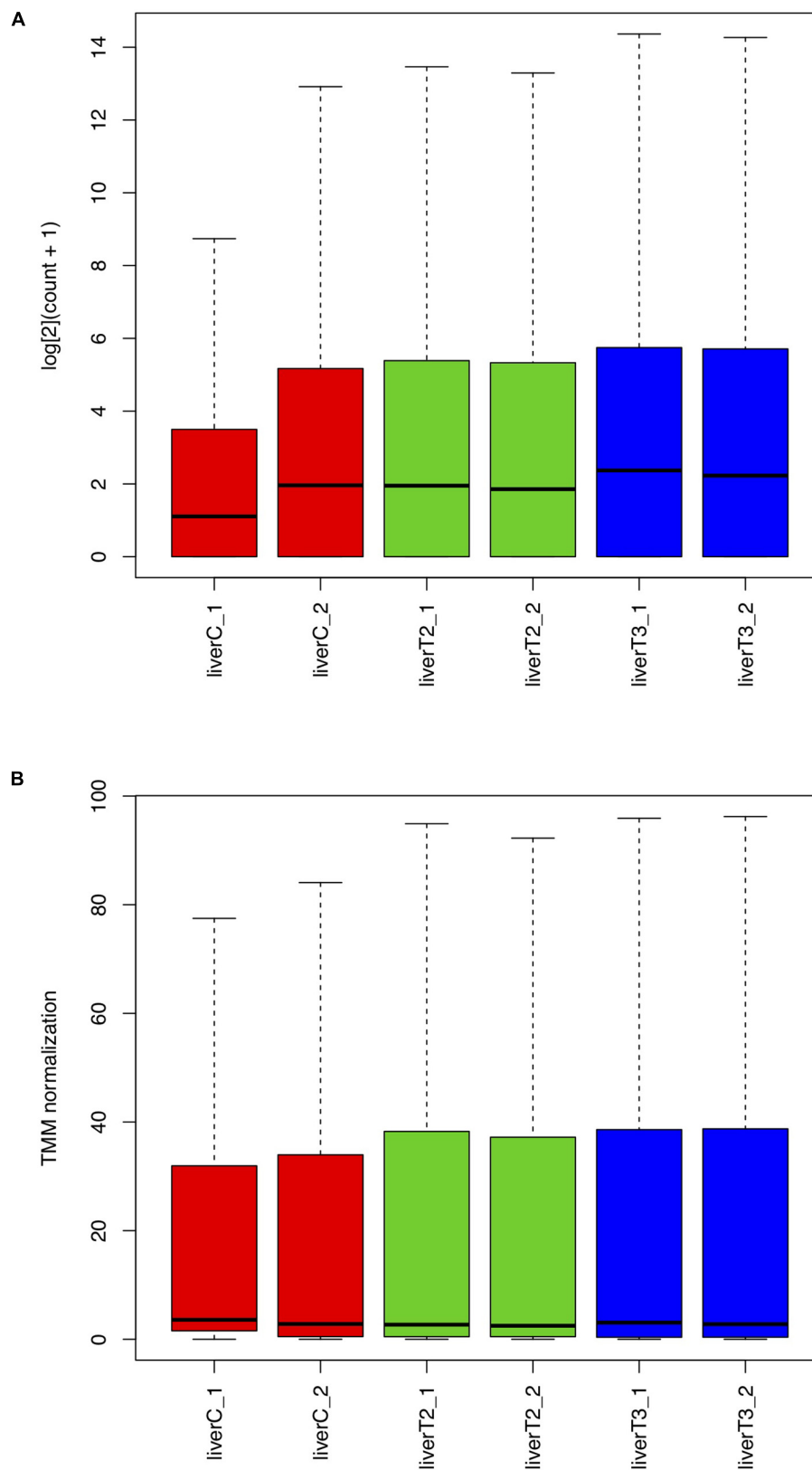
Multi-dimensional scaling (MDS) is a technique that is used to create a visual representation of the pattern of proximities (similarities, dissimilarities, or distances) among a set of objects. In the context of RNA-Seq analysis, MDS plot shows variation

among RNA-Seq samples, the more is the distance between sample, the higher is their dissimilarity. Therefore, samples belonging to the same condition or treatment should be closer to each other and distant to other conditions. However, if different conditions are grouped together, this could mean that those treatments or conditions have a very similar effect. Worst-case scenario, the user can suspect of a sample mislabeling. Conceptually, MDS and PCA plots can provide the same information and as observed in **Figure 6**, samples belonging to C, T2, and T3 form separate clusters with a certain dispersion among replicates. Similarly, to the PCA plot, this plot could indicate if there is a batch effect problem, where samples in a same condition are very dispersed in the plot. Again, we recommend to confirm the preparation of the samples, by checking records from the preparation of the samples in the wet lab.

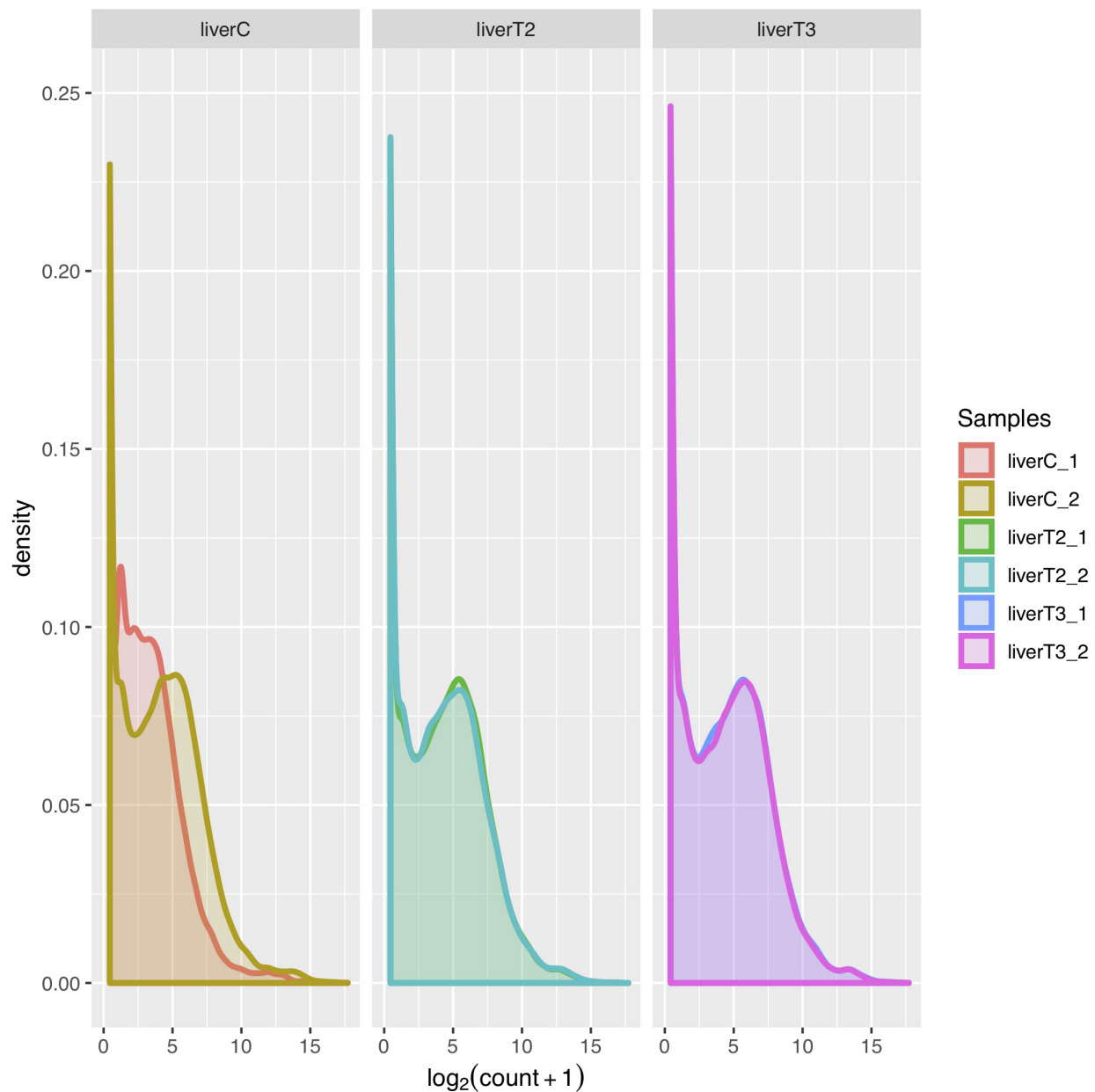
### Differential Expression Analysis

The "(2) Differential Expression Results" section has links with the name of each selected method, where the user can display the analysis output. A detailed description of each method output can be found in the User Manual at the IDEAMEX web page.





**FIGURE 3 |** Boxplot with normalized counts. The frequency distribution and some statistics like mean, median and outliers are represented in these plots. **(A)** log<sub>2</sub> normalized counts. **(B)** TMM normalized counts.



**FIGURE 4 |** Density plots. The count distribution between replicates and conditions.

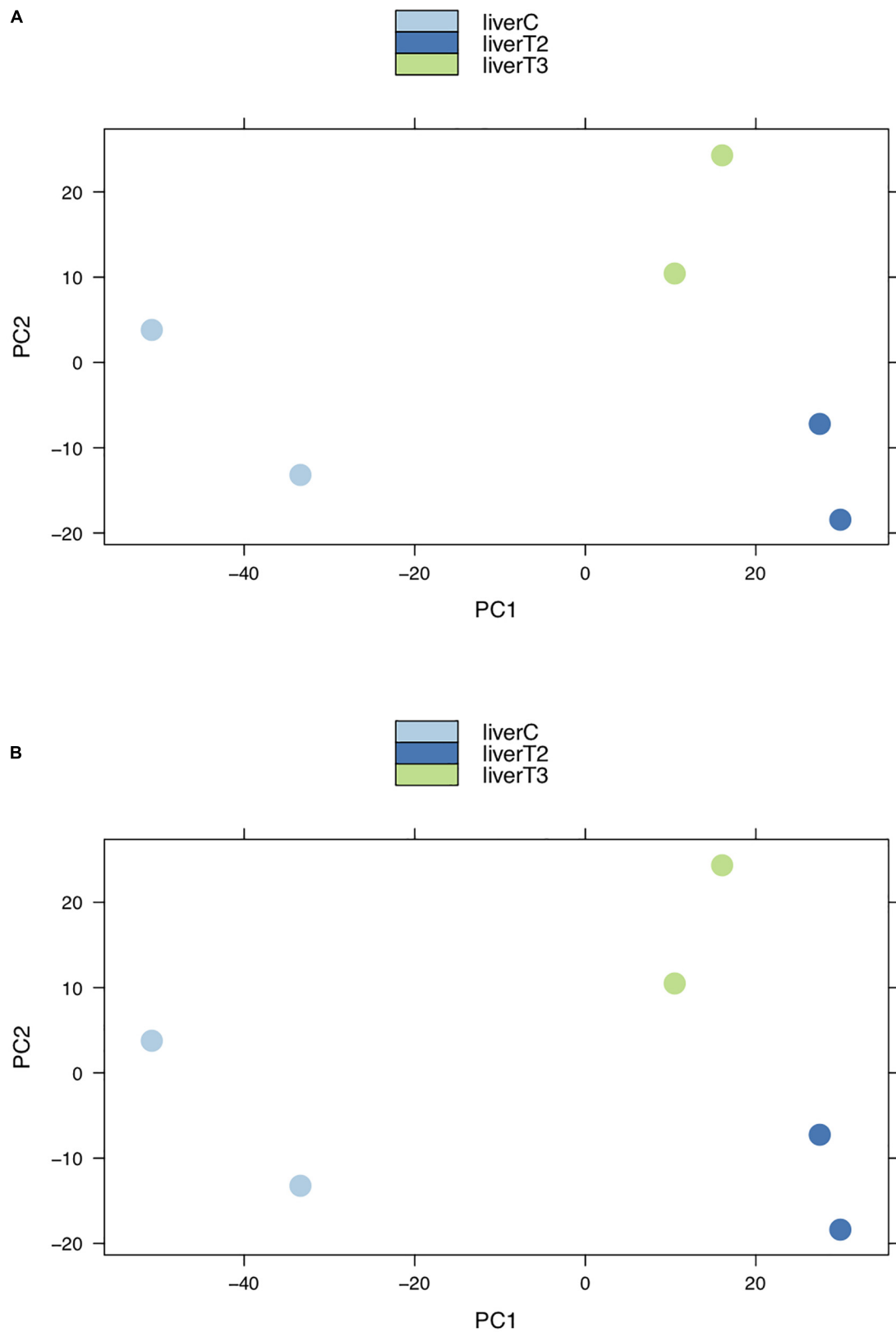
However, here we describe the generated graphs for a better interpretation. **Table 2** shows the output plots generated by each method, contributing with different representations of the genes that were differentially expressed. Some of these plots were already used in the “(1) Data Analysis” section (PCA and MDS plots). If the user indicated that samples for a given condition belonged to different batches, the batch error effect correction for several methods will be applied.

### Expression, MA, MD and Smear Plots

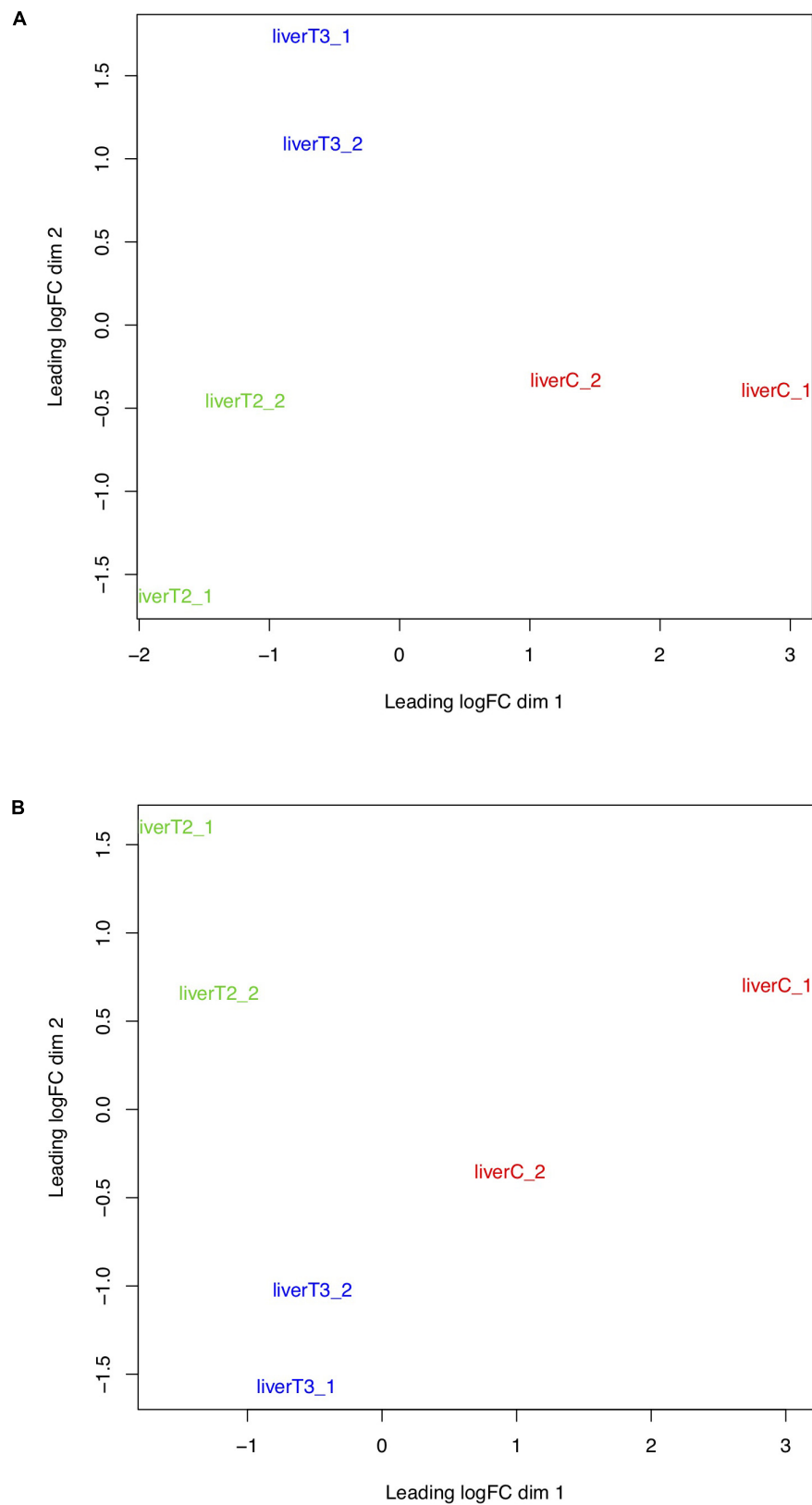
These plots depict all expressed genes but those with differential expression are represented in other color than

black. Basically, in all of them we can see the distribution of the gene expression according to a certain value. For example, in the expression plot (**Supplementary Figure S1**) the average expression values for each gene of the compared conditions are plotted and those highlighted in red are genes with a significant difference compared to the rest. In simple terms, the differentially expressed genes are those with outlier mean values.

In the MA-plot (**Supplementary Figure S2**), the  $\log_2$  fold change ( $\log_2\text{FC}$ ) expression and the normalized mean counts of each gene in the compared conditions are plotted. Features declared as differentially expressed are highlighted in different



**FIGURE 5 |** PCA plots. Groups of samples can be analyzed using Principal Component Analysis (PCA) plots where replicates of a certain conditions are clustered together. Clusters from different conditions are separated. **(A)** log2 normalized counts. **(B)** TMM normalized counts.



**FIGURE 6 |** MDS plots. Groups of samples can be analyzed using Multi-dimensional scaling (MDS) plots where the distance between samples and conditions reflect their similarity. **(A)** log2 normalized counts. **(B)** TMM normalized counts.

**TABLE 2** | Plots generated by each differential expression package.

Plot/Method	edgeR	limma	NOISeq	DESeq2
Expression	X	X	Yes	X
MA	X	X	X	Yes
MD	X	Yes	Yes	X
Smear	Yes	X	X	X
Volcano	Yes	X	X	X

colors according to the logFC threshold defined by the user and the expression directionality (UP or DOWN).

The mean-difference (MD) plot (**Supplementary Figure S3**) shows the average expression (mean: x-axis in limma or D for NOISeq) against logFC (difference: y-axis in limma or M for NOISeq). Again, values declared as differentially expressed are highlighted in red.

The smear plot allows to visualize the results of the analysis in a similar manner to the MA-plot, this plot shows the logFC against log-CPM, where genes declared as differentially expressed highlighted in red.

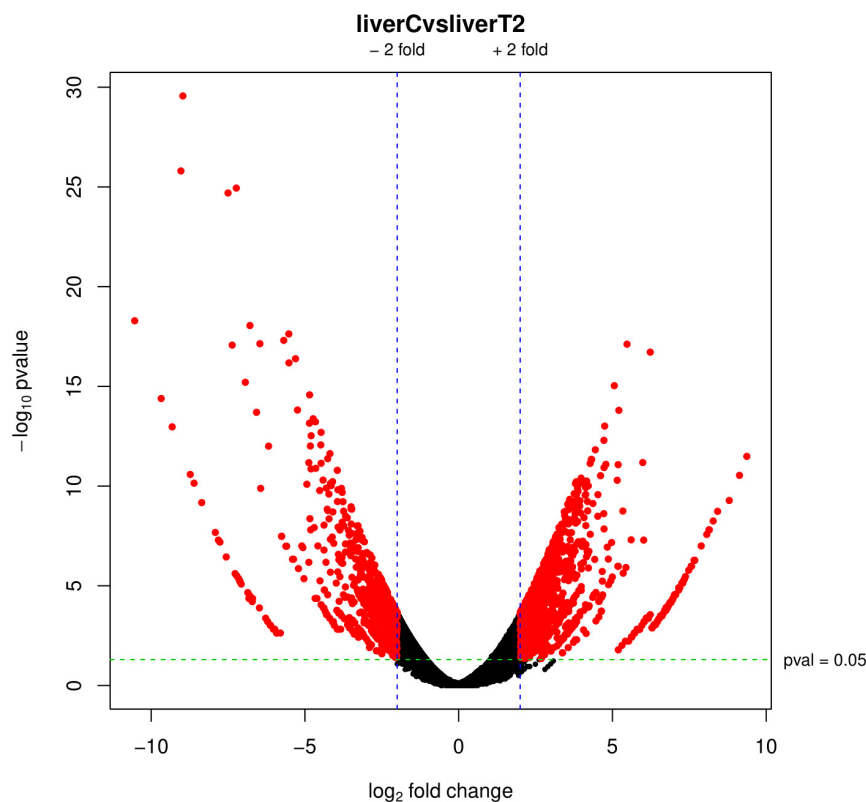
In summary, all these plots compare the expression rate or difference between conditions and the normalized values. The proportion of black and highlighted dots gives an idea of the expression change magnitude between the treatment and the control or untreated conditions.

## Volcano Plot

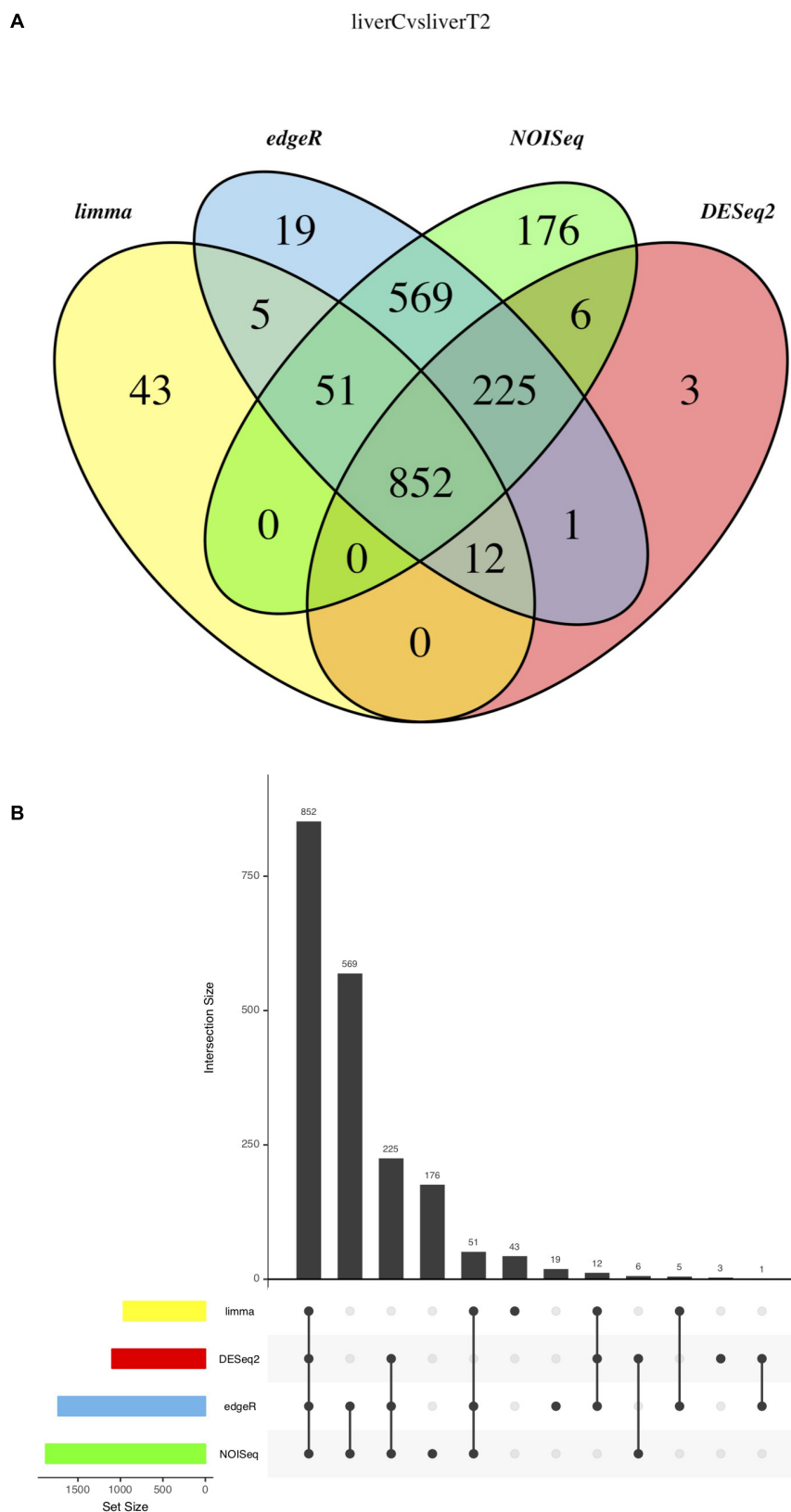
Arguably, the volcano plot (**Figure 7**) is the most popular and probably, the most informative graph since it summarizes both the expression rate (logFC) and the statistical significance ( $p$ -value). It is a scatter-plot of the negative log<sub>10</sub>-transformed  $p$ -values from the gene-specific test (on the y-axis) against the logFC (on the x-axis). The graph depicts datapoints with low  $p$ -values (highly significant) appearing toward the top of the plot. The logFC values are used to determine the change direction (up and down) appearing equidistant from the center. Features declared as differentially expressed are highlighted in red, according to the selected cut-off values.

## Results Integration

Finally, the “(3) Results Integration” section of the Analysis Results in the IDEAMEX web page contains several text files and graphs that integrates the results from all selected methods. In **Figure 8**, we present the results from the C vs. T2 comparison, using a Venn diagram (**Figure 8A**), upset bar (**Figure 8B**) and correlograms (**Supplementary Figure S5**) plots. For the analyzed data, the Venn diagram showed all method intersections and it is observed that 852 genes were validated as differentially expressed by all four methods, being NOIseq the main contributor as also observed in the upset bar plot. It is interesting that limma-Voom reported that 43 genes that no other method found as differentially expressed but agreed with the other methods in

**FIGURE 7** | Volcano plot. Red dots represent differentially expressed genes according to the  $p$ -adj and logFC cut-off values.





**FIGURE 8 |** Result Integration summary. **(A)** Venn diagram representing the result intersection for each selected method. **(B)** Upset plot representing the contribution of each selected method.

**TABLE 3 |** Example of intersect results table.

	edgeR	limma	NOISeq	DESeq2	Regulation
ENSONIT00000000023	1	1	1	1	DOWN_liverC_UP_liverT2
ENSONIT00000000075	1	1	1	1	DOWN_liverC_UP_liverT2
ENSONIT00000000081	1	1	1	1	DOWN_liverC_UP_liverT2
ENSONIT00000000102	1	1	1	1	UP_liverC_DOWN_liverT2
ENSONIT00000000120	1	1	1	1	DOWN_liverC_UP_liverT2
ENSONIT00000000129	1	1	1	1	DOWN_liverC_UP_liverT2
ENSONIT00000000206	1	1	1	1	DOWN_liverC_UP_liverT2

Snippet of the liver CvsT2 treatment. Original table is in simple text format. The Regulation column indicates the directionality of the gene expression.

920 genes (5 + 51 + 852 + 12). This gives the option to the user to work with either only the intersection or the union of all methods. However, working with all methods can be overwhelming for the user although using an enrichment analysis using the GO term or metabolic annotation from KEGG could help.

As mentioned, other generated plots are heatmaps (**Supplementary Figure S4**) and correlogram (**Supplementary Figure S5**) plots. Since each method has different normalization methods, fold change or statistical metrics (*p*-adj, FDR or Probability) to determine if a gene is differentially expressed, the correlograms can help the user to evaluate the correlation of these values among the different used methods. Also, heatmaps are created to observe samples clustered by their fold change, allowing the user to spot groups of genes with a similar expression change.

Among all the text file results that are explained in detail in the User Manual (**Supplementary Material S2**), the IntersectTopRegulation.txt file provides the list of all differentially expressed genes with a 0| 1 matrix that can be used select genes depending on how many and which methods reported them as differentially expressed. In the last column of the file, a description of the gene regulation can be found, where is indicated how and in which condition the genes was expressed. **Table 3** has a snippet of the liverCvsT2\_ IntersectTopRegulation.txt file where the Regulation structure results is as follows: UP\_conditionX\_DOWN\_conditionY or DOWN\_conditionX\_UP\_conditionY. Therefore, the user can select which genes were up or down regulated in a certain condition and be sure of the directionality of the expression without checking the fold change directionality.

DISCUSSION

The IDEAMEX web server is a useful resource for transcriptome experiments designed for differential expression analysis involving several condition comparisons. The methods for differential expression analysis in the workflow, were selected based on their performance in several benchmark analyses since the emergence of RNA-Seq data as a powerful alternative to microarrays (Anders et al., 2013; Soneson and Delorenzi, 2013; Seyednasrollah et al., 2015; Costa-Silva et al., 2017). In particular, our web server uses R packages that use different algorithms and normalization methods giving a broader view of

the results with a higher confidence based on their agreement, based on the idea that no statistical modeling can fully capture biological phenomena. In the case of limma and NOISeq, they use non-parametric methods that are statistical techniques for which we do not have to make any assumption of the gene expression; whereas DESeq2 and edgeR use parametric methods assuming a binomial distribution for the data and that no genes are differentially expressed.

Once the user had loaded the input data in the right format, our server allows the user to design which comparisons will be made and which cut-off values will be used, instead of running an all-vs.-all comparison and default parameters for each package. For parametric methods like edgeR and DESeq2, the FDR and *p*-adj values are the statistical parameters that define the probability of a gene to be differentially expressed in a multiple comparison and are used to define if a gene was differentially expressed or not from the statistical point of view. However, other parameters such as the CPM or logFC can have a biological meaning and also can be used a cut-off value. Is not straightforward how to select which cut-off values will be the best for a certain experiment but IDEAMEX allow users to try many combinations of them by running the comparisons several times and inspecting the different results.

Is very important that the user select which comparisons have a sense in terms of their experimental design. For example, in this work we used three conditions where one was used as a control to study the effect of two thyroid hormones treatments in tilapia liver (T2 and T3). The comparison between T2 and T3 has to be performed by comparing the results from comparing each one to the control or untreated condition. A direct comparison between T2 and T3 could miss several results since even if we can observe a gene with a certain expression change, the difference could not be statistically significant. Let's say that "gene A" has a differential expression of 10 times in T2 vs. C comparison and of 12 times in the T3 vs. C comparison. Roughly, the difference between T2 vs. T3 comparison for the same gene, will be 2 times which might not be statistically significant. For this reason, is very important to select the which comparisons make sense, instead of performing all possible comparison.

The results in the "Data Analysis" section, are several plots that allow the user to inspect the distribution of their data based on different metrics. This quality control check point is very important, since biological data tend to be very noisy. It is expected that the data from biological replicates within a certain

condition, will have the same distribution and a similar trend than those in other conditions. In particular, PCA and MDS plots allow the users to see if biological replicates of a certain condition are grouped together and if each condition forms a separate group. In this particular case, it was known that the samples didn't present any batch effect but as observed in **Figure 6**, there is some dispersion between samples. It is not trivial to determine if samples present a dispersion attributable to a batch effect. Therefore, it is important to obtain the information regarding the sample preparation to discriminate between high "biological" variability and "noise" from batch effect.

The distance or dispersion of the replicates and groups indicates how reproducible was the tested condition in different individuals or how variable were individuals despite the treatment. The more replicates available, the better statistical significance is observed. Having very disperse groups or samples from different conditions grouping together, should be considered as noisy or highly variable results that can skew the analysis and lead to misinterpretation of the experiments. However, NOIseq could be a good option when no biological replicates are available and as reported elsewhere, it delivers reliable results that have been confirmed by using quantitative PCR (qPCR) reactions.

The results from different methods are not mutually exclusive. From the statistical point of view, one of them, neither or all may be true. Therefore, working with the intersection or the union of all results is a decision that the user has to evaluate after exploring them based not only on the statistical significance but on the biological meaning that will depend on the gene annotation. The main problem with all statistics is the "fakeness" and misrepresentation of the results. However, if four different methods agreed with a certain result it could be assumed that those genes are differentially expressed, bearing in mind that an experimental orthogonal validation using a different technology like qPCR, should be necessary to confirm the result.

In the "Results Integration" section, there are several text lists and graphs that can guide the users to make sense out of the results from their experiments. As mentioned, the Venn diagram (**Figure 8A**) shows the intersection and union of the selected different methods. The user can choose one or more methods by evaluating the agreement between them since one method could generate either an overwhelming amount of results or very few of them. In the former case, the user can choose to work with the intersection of all methods or in the latter case, the union will provide the maximum amount of reported results.

In this work, we provide heatmaps and correlograms for different values obtained from each method. For example, heatmaps (**Supplementary Figure S4**) are useful to spot gene clusters with the same fold change pattern, suggesting that those genes could belong to a certain pathway or are regulated by the same mechanism. However, users have to be very careful when determining gene clusters since there is no straightforward method to do so. Defining the cluster size is not trivial and usually is a trial and error process. In terms of novelty, the most interesting plot could be the statistical

parameter correlogram (**Supplementary Figure S5**), where the threshold values such as p-adj (limma-Voom and DESeq2), FDR (edgeR) and Prob (NOIseq) values are correlated. To our knowledge, this correlation has not been reported in other studies. Surprisingly, methods usually correlate very well since the statistical threshold denotes the error probability of each result. In our experience, we have observed that NOIseq is the method with lower correlation regarding the error probability since this is calculated using a very different approach (Tarazona et al., 2011) compared to the rest of the methods. However, is somehow refreshing that all methods present a good correlation, suggesting that are consistent identifying differentially expressed genes and those with no significant change, despite using different statistics.

Finally, there are several other methods to continue the differential expression analysis, that can help users to put their results in a certain biological context. Probably the most popular methods are those based on Gene Ontology (GO) terms enrichment (Maere et al., 2005; Eden et al., 2009; Reimand et al., 2016) which will require of a well curated gene annotation. Other enrichment methods like Gene Set Enrichment Analysis (GSEA) determine whether a defined set of genes shows statistically significant based on molecular signatures (Subramanian et al., 2007; Liberzon et al., 2011) or metabolic pathway enrichment analysis (Luo et al., 2009; Liu et al., 2017; Ulgen et al., 2018) can provide a better picture of the biological meaning of the observed changes in gene expression for a given treatment or condition. These enrichment methods along with the heatmaps, can help the researcher to spot regulation networks or pathways which could be subject to further studies.

## CONCLUSION

We consider that the IDEAMEX web server can help other researchers with no previous bioinformatic knowledge, to perform their analyses in a simple manner. Also, more experienced users with some bioinformatics skills can use the results and perform a more detailed analysis and a different integration of them, since all the results are provided in simple text files which are very convenient to parse and handle using regular expression searches.

## DATA AVAILABILITY

The datasets analyzed for this study can be found in the NCBI SRA repository (<https://submit.ncbi.nlm.nih.gov/subs/sra/>), under the SRA identifiers: SRX2630485, SRX2630486, SRX2630487, SRX2630488, SRX2630489, and SRX2630490.

## AUTHOR CONTRIBUTIONS

VJ-J and LV-A developed the web deployment and scripts for the IDEAMEX server. VJ-J, LV-A, and AS-F conceived the web server workflow. AS-F wrote the manuscript. All authors read and authorized the publication of this manuscript.

## FUNDING

The computer hosting the IDEAMEX web server was provided and maintained by the Unidad Universitaria de Secuenciación Masiva y Bioinformática using its core budget and CONACyT #260481 grant from the “Laboratorios Nacionales” program.

## ACKNOWLEDGMENTS

We would like to thank “Laboratorio Nacional de Apoyo Tecnológico a las Ciencias Genómicas” CONACyT #260481 for infrastructural support hosting the server.

## REFERENCES

- Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., et al. (2016). The Ensembl gene annotation system. *Database* 2016:baw093. doi: 10.1093/database/baw093
- Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106
- Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., et al. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* 8, 1765–1786. doi: 10.1038/nprot.2013.099
- Brooks, A. N., Duff, M. O., and Yang, L. (2010). Conservation of an RNA regulatory map between drosophila and mammals. *Genome Res.* 21, 193–202. doi: 10.1101/gr.108662.110
- Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M., and Gu, J. (2017). AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics* 18:80. doi: 10.1186/s12859-017-1469-3
- Costa-Silva, J., Domingues, D., and Lopes, F. M. (2017). RNA-Seq differential expression analysis: an extended review and a software tool. *PLoS One* 12:e0190152. doi: 10.1371/journal.pone.0190152
- Cumby, J. S., Di, Y., and Kimbrel, J. A. (2011). Gene-Counter: a computational pipeline for the analysis of RNA-Seq data for gene expression differences. *PLoS One* 6:e25279. doi: 10.1371/journal.pone.0025279
- de Jong, A., van der Meulen, S., Kuipers, O. P., and Kok, J. (2015). T-REx: transcriptome analysis webserver for RNA-seq expression data. *BMC Genomics* 16:663. doi: 10.1186/s12864-015-1834-4
- Di, Y., Cumby, J. S., and Schafer, D. W. (2014). *Negative Binomial Model for RNA-Sequencing Data*. Available at: <https://cran.r-project.org/web/packages/NBPSeq/index.html>.
- Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48. doi: 10.1186/1471-2105-10-48
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883
- Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* 12, 115–121. doi: 10.1038/nmeth.3252
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36
- Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25. doi: 10.1186/gb-2009-10-3-r25
- Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324
- Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260
- Liu, L., Wei, J., and Ruan, J. (2017). Pathway enrichment analysis with networks. *Genes* 8:246. doi: 10.3390/genes8100246
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8
- Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* 10:161. doi: 10.1186/1471-2105-10-161
- Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* 21, 3448–3449. doi: 10.1093/bioinformatics/bti551
- McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4288–4297. doi: 10.1093/nar/gks042
- Monier, B., McDermaid, A., Zhao, J., Fennell, A., and Ma, Q. (2018). IRIS-EDA: an integrated RNA-Seq interpretation system for gene expression data analysis. *bioRxiv* [preprint]. doi: 10.1101/283341
- Olvera, A., Martyniuk, C. J., Buisine, N., Jiménez-Jacinto, V., Sanchez-Flores, A., Sachs, L. M., et al. (2017). Differential transcriptome regulation by 3,5-T2 and 3',5-T3 in brain and liver uncovers novel roles for thyroid hormones in tilapia. *Sci. Rep.* 7:15043. doi: 10.1038/s41598-017-14913-9
- Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033
- Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., et al. (2016). g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* 44, W83–W89. doi: 10.1093/nar/gkw199
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007
- Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* 12:R22. doi: 10.1186/gb-2011-12-3-r22

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.00279/full#supplementary-material>

**FIGURE S1** | Liver C vs. T3 MA plot.

**FIGURE S2** | Liver C vs. T3 MD plot.

**FIGURE S3** | Liver C vs. T3 smear plot.

**FIGURE S4** | Liver C vs. T3 heatmap.

**FIGURE S5** | Liver C vs. T3 statistic parameter correlogram.

**MATERIAL S1** | Raw count table.

**MATERIAL S2** | IDEAMEX user manual.

- Roser, L. G., Aguero, F., and Sánchez, D. (2018). FastqCleaner: an interactive Bioconductor application for quality-control, filtering and trimming of FASTQ files. *bioRxiv* [preprint]. doi: 10.1101/393140
- Seyednasrollah, F., Laiho, A., and Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.* 16, 59–70. doi: 10.1093/bib/bbt086
- Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14:91. doi: 10.1186/1471-2105-14-91
- Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J. P. (2007). GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics* 23, 3251–3253. doi: 10.1093/bioinformatics/btm369
- Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Res.* 21, 2213–2223. doi: 10.1101/gr.124321.111
- Ulgén, E., Ozisik, O., and Sezerman, O. U. (2018). pathfindR: an R package for pathway enrichment analysis utilizing active subnetworks. *bioRxiv* [preprint]. doi: 10.1101/272450
- Wu, T. D., Reeder, J., Lawrence, M., Becker, G., and Brauer, M. J. (2016). GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods Mol. Biol.* 1418, 283–334. doi: 10.1007/978-1-4939-3578-9\_15
- Zhang, C., Fan, C., Gan, J., Zhu, P., Kong, L., and Li, C. (2018). iSeq: web-based RNA-seq data analysis and visualization. *Methods Mol. Biol.* 1754, 167–181. doi: 10.1007/978-1-4939-7717-8\_10
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Copyright © 2019 Jiménez-Jacinto, Sanchez-Flores and Vega-Alvarado. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# DREAMSeq: An Improved Method for Analyzing Differentially Expressed Genes in RNA-seq Data

Zhihua Gao<sup>1,2</sup>, Zhiying Zhao<sup>1</sup> and Wenqiang Tang<sup>1\*</sup>

<sup>1</sup> Ministry of Education Key Laboratory of Molecular and Cellular Biology, Hebei Key Laboratory of Molecular and Cellular Biology, Hebei Collaboration Innovation Center for Cell Signaling, College of Life Sciences, Hebei Normal University, Shijiazhuang, China, <sup>2</sup> College of Biological Science and Engineering, Hebei University of Economics and Business, Shijiazhuang, China

## OPEN ACCESS

### Edited by:

Monica Bianchini,  
Università degli Studi di Siena, Italy

### Reviewed by:

Shihao Shen,  
University of California, Los Angeles,  
United States  
Taina Raiol,  
Fundação Oswaldo Cruz (Fiocruz),  
Brazil

### \*Correspondence:

Wenqiang Tang  
tangwq@mail.hebtu.edu.cn

### Specialty section:

This article was submitted to  
Bioinformatics and Computational  
Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 16 August 2018

**Accepted:** 15 November 2018

**Published:** 30 November 2018

### Citation:

Gao Z, Zhao Z and Tang W (2018)  
DREAMSeq: An Improved Method for  
Analyzing Differentially Expressed  
Genes in RNA-seq Data.  
Front. Genet. 9:588.  
doi: 10.3389/fgene.2018.00588

RNA sequencing (RNA-seq) has become a widely used technology for analyzing global gene-expression changes during certain biological processes. It is generally acknowledged that RNA-seq data displays equidispersion and overdispersion characteristics; therefore, most RNA-seq analysis methods were developed based on a negative binomial model capable of capturing both equidispersed and overdispersed data. In this study, we reported that in addition to equidispersion and overdispersion, RNA-seq data also displays underdispersion characteristics that cannot be adequately captured by general RNA-seq analysis methods. Based on a double Poisson model capable of capturing all data characteristics, we developed a new RNA-seq analysis method (DREAMSeq). Comparison of DREAMSeq with five other frequently used RNA-seq analysis methods using simulated datasets showed that its performance was comparable to or exceeded that of other methods in terms of type I error rate, statistical power, receiver operating characteristics (ROC) curve, area under the ROC curve, precision-recall curve, and the ability to detect the number of differentially expressed genes, especially in situations involving underdispersion. These results were validated by quantitative real-time polymerase chain reaction using a real Foxtail dataset. Our findings demonstrated DREAMSeq as a reliable, robust, and powerful new method for RNA-seq data mining. The DREAMSeq R package is available at <http://tanglab.hebtu.edu.cn/tanglab/Home/DREAMSeq>.

**Keywords:** RNA-seq, DREAMSeq, equidispersion, overdispersion, underdispersion, double Poisson model, negative binomial model

## INTRODUCTION

With the development of next-generation sequencing technology, RNA sequencing (RNA-seq) has become a routine and powerful method for evaluating global dynamic changes in gene expression during certain biological processes. Compared with microarray technologies, RNA-seq technologies have several advantages, including a wider measurable range of expression levels, higher throughput, less noise, more information for detecting allele-specific expression, and a higher capability to detect novel promoters and alternative gene-splicing isoforms (Marioni et al., 2008; Mortazavi et al., 2008; Sultan et al., 2008; Wang et al., 2009, 2010b; Oshlack et al., 2010). Therefore, developing powerful, reliable, and unbiased RNA-seq data-mining methods would facilitate the use of RNA-seq to explore basic biological questions in this era of big data.

Typically, RNA-seq experimental procedures can be divided into six steps: (1) sequencing the RNA samples to obtain raw reads, (2) filtering out low-quality reads, (3) mapping the high-quality reads to a reference genome or transcriptome, (4) summarizing the read counts for each gene, (5) detecting differentially expressed genes (DEGs), and (6) performing systems biology analysis [e.g., cluster analysis, principal components analysis (PCA), gene ontology (GO) analysis, and pathway enrichment analysis] (Oshlack et al., 2010). Of these steps, identifying DEGs across treatments/conditions is the key task and often the primary goal of RNA-seq data analysis. There are numerous statistical methods focusing directly on read-count data for DEG identification, with these classified into two categories: (1) parametric methods that rely on assumptions about discrete probability models and include methods based on a Poisson model, such as DESeq (Wang et al., 2010a) and TSPM (Auer and Doerge, 2011), methods based on a negative binomial (NB) model, such as edgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010), baySeq (Hardcastle and Kelly, 2010), NBPSeg (Di et al., 2011), EBSeq (Leng et al., 2013), ShrinkSeq (Van De Wiel et al., 2013), and DESeq2 (Love et al., 2014), methods based on a beta-binomial model, such as BBSeq (Zhou et al., 2011), methods based on a multivariate Poisson log-normal (LN) model, such as PLNseq (Zhang et al., 2015), and methods based on a generalized Poisson (GP) model, such as GPseq (Srivastava and Chen, 2010) and deGPS (Chu et al., 2015); and (2) non-parametric methods, such as NOISeq (Tarazona et al., 2011) and SAMseq (Li and Tibshirani, 2013), that do not assume any particular model.

Among count-based RNA-seq data-analysis methods, non-parametric methods were developed based on large-sample asymptotic theory and exhibit statistical power sufficient to detect DEGs only when the number of replicates per treatment condition is  $\geq 5$  (Tarazona et al., 2011; Seyednasrollah et al., 2013; Sonesson and Delorenzi, 2013). However, due to the high cost of RNA-seq, the general sample size in a typical RNA-seq experiment is  $< 5$  replicates, which limits the application of non-parametric methods in RNA-seq data mining. Therefore, the most popular RNA-seq data-analysis methods are parametric methods based on Poisson and NB models. In early RNA-seq studies where only technical replicates were used, the traditional Poisson model was highly capable of fitting read-count data characterized by equidispersion (i.e., the variance is equal to the mean) (Marioni et al., 2008; Bullard et al., 2010). However, when biological replicates are available, read-count data often exhibits more variability than the Poisson model expects, which limits the use of a Poisson model for analyzing RNA-seq data (Anders and Huber, 2010). Fortunately, the NB model, as a Gamma-Poisson mixture, can address the overdispersion issue (i.e., when the variance is larger than the mean), as well as capture equidispersion (Anders and Huber, 2010). Additionally, recent studies reported that some RNA-seq data demonstrates characteristics of underdispersion (i.e., the variance is smaller than the mean), which might be caused by RNA-seq coverage, as well as zero-inflation, cluster, or low expression level of the count data, and could lead to

underestimation of DEGs (Famoye, 1993; Srivastava and Chen, 2010; Rau et al., 2011; Mi et al., 2015; Choo-Wosoba et al., 2016; Low et al., 2017). However, neither a traditional Poisson model nor the NB model works well at mining underdispersed data.

The GP model is a generalization of the Poisson model with an additional parameter. This method can process data characterized by underdispersion and non-underdispersion (equidispersion and overdispersion) (LuValle, 1990), but can only capture certain levels of dispersion, because the model is truncated under certain conditions regarding its bounded dispersion parameter (Famoye, 1993). For example, the program deGPS employs the GP model to fit read-count data characterized by non-underdispersion (Chu et al., 2015), whereas GPseq uses this model to consider potential positional bias during DEG analysis and handle position-level counts instead of gene-level counts, which is different from other methods (Srivastava and Chen, 2010). Therefore, these methods derived from different discrete models can potentially perform poorly at fitting underdispersed count data due to the restrictions associated with the inherent properties in the models.

In this study, we described a mixed Poisson model called double Poisson (DP), which offers the advantage of flexibility in fitting a wide range of data exhibiting underdispersion and non-underdispersion using only two parameters (Efron, 1986). Based on this model, we developed a novel differential relative expression-analysis method for RNA-seq data mining (DREAMSeq). Because the results of differential gene-expression analysis are dependent upon the discrete model used to fit the RNA-seq data (Consortium, 2010), we also added NB-model functionality to the DREAMSeq pipeline in order to optimize the performance of our method. Therefore, depending on the model used in the pipeline, our method can be divided into three approaches: DREAMSeq.DP (based on the DP model), DREAMSeq.NB (based on the NB model), and DREAMSeq.Mix (based on the mixture of the DP and NB models, with the lower  $p$ -value between two  $p$ -values generated based on the DP and NB models chosen as the final  $p$ -value) in order to fit variable RNA-seq data. In order to evaluate the performance of DREAMSeq, we generated three simulated datasets using three real RNA-seq datasets. Because the DEGs can only be effectively identified when the sample size is  $\geq 3$  (Conesa et al., 2016; Lin et al., 2016), to assess DREAMSeq using the most common RNA-seq scenario, we focused on detecting DEGs under small sample sizes (three replicates per condition) and between two groups. Our results indicated that the performance of DREAMSeq at effectively detecting DEGs was comparable to other popular RNA-seq data-analysis methods, including edgeR, DESeq, DESeq2, NBPSeg, and TSPM, in non-underdispersion situations, but outperformed most of the other methods in underdispersion situations. This conclusion was validated by quantitative real-time polymerase chain reaction (qRT-PCR) using a real Foxtail dataset generated in our laboratory. Our findings demonstrated DREAMSeq as a reliable and robust DEG-detection method that provides an additional option in the RNA-seq data-analysis toolbox, especially for underdispersed-data mining.

## MATERIALS AND METHODS

### Models and Normalization

In this study, let  $Y$  represent the observed count and  $X$  the corresponding underlying gene expression (unknown) in an RNA-seq experiment. Let  $Y_{ijk}$  and  $X_{ijk}$  denote the read count and the true gene expression of gene  $i$  from sample  $j$  in treatment group  $k$ , where  $i = 1, \dots, I$  (the number of genes),  $j = 1, \dots, J$  (the number of replicates; here,  $J = 3$ ), and  $k = 1, \dots, K$  (the number of groups; here,  $K = 2$ ), respectively.

#### NB Model

We assume that  $Y$  follows an NB model with two parameters: the mean,  $\mu$ , and the dispersion,  $\phi$ . The probability mass function (PMF) of the NB model is given as:

$$P(Y = y | \mu, \phi) = \frac{\Gamma(y + \phi^{-1})}{y! \Gamma(\phi^{-1})} \left( \frac{1}{1 + \mu\phi} \right)^{\phi^{-1}} \left( \frac{\mu\phi}{1 + \mu\phi} \right)^y. \quad (1)$$

The expected value is estimated as:

$$E(Y) = \mu. \quad (2)$$

We parameterize the variance of the NB model according to a previous study (Robinson and Smyth, 2007):

$$\text{Var}(Y) = \sigma^2 = \mu + \mu^2 \phi, \quad (3)$$

where  $\phi \geq 0$  and determines the extra variability as compared with the Poisson model. When  $\phi > 0$ ,  $\sigma^2 > \mu$ ; and when  $\phi = 0$ ,  $\sigma^2 = \mu$ ; the NB model collapses to the Poisson model, which can be viewed as a special NB model with zero dispersion (Robinson and Smyth, 2007). Therefore, the NB model allows for both overdispersion and equidispersion.

#### DP Model

We assume that  $Y$  follows a DP model with two parameters: the mean,  $\mu$ , and the dispersion,  $\theta$ . The approximate PMF of the DP model is given as:

$$P(Y = y | \mu, \theta) = f_{\mu, \theta}(y) = (\theta^{\frac{1}{2}} e^{-\theta\mu}) \left( \frac{e^{-y} y^y}{y!} \right) \left( \frac{e\mu}{y} \right)^{\theta y}. \quad (4)$$

The exact DP density is:

$$P(Y = y | \mu, \theta) = \tilde{f}_{\mu, \theta}(y) = c(\mu, \theta) f_{\mu, \theta}(y), \quad (5)$$

where the factor  $c(\mu, \theta)$  can be calculated as:

$$\frac{1}{c(\mu, \theta)} = \sum_{y=0}^{\infty} \tilde{f}_{\mu, \theta}(y) \approx 1 + \frac{1 - \theta}{12\mu\theta} \left( 1 + \frac{1}{\mu\theta} \right) \quad (6)$$

with  $c(\mu, \theta)$  being the normalizing constant nearly equal to 1. The constant  $c(\mu, \theta)$  ensures that the density integrates to unity. The expected value and the variance of the DP model in reference to the exact density  $\tilde{f}_{\mu, \theta}(y)$  are estimated as follows:

$$E(Y) \approx \mu \quad (7)$$

and

$$\text{Var}(Y) = \sigma^2 = \frac{\mu}{\theta}, \quad (8)$$

respectively, where  $\theta > 0$  under RNA-seq data circumstances. The Poisson model is nested in the DP model for  $\theta = 1$ , indicating that the DP model can fit equidispersed read-count data when  $\theta = 1$ . Additionally, the DP model allows for both overdispersion ( $0 < \theta < 1$ ) and underdispersion ( $\theta > 1$ ) (Efron, 1986).

#### Normalization

Here, we assume that the expectation of  $Y_{ijk}$ ,  $\mu_{ijk}$ , is the product of  $X_{ijk}$  and  $s_{jk}$ :

$$\mu_{ijk} = X_{ijk} s_{jk}, \quad (9)$$

where  $s_{jk}$  is the size factor corresponding to sample  $j$  in treatment group  $k$ , which can be estimated using various existing normalization methods, such as total counts, upper quartile (Bullard et al., 2010), median (Dillies et al., 2012), quantile (Bolstad et al., 2003; Irizarry et al., 2003), trimmed mean of  $M$ -values (TMM) (Robinson and Oshlack, 2010), DESeq normalization (DESeq) (Anders and Huber, 2010), reads per kilobase per million (RPKM) (Mortazavi et al., 2008), to remove unwanted variation (Risso et al., 2014). Normalization is a process that makes unit-less data comparable among measurements by adjusting for sequencing depth and potentially other technical effects of different samples. Dillies et al. (2012) and Lin et al. (2016) found that TMM and DESeq normalization methods performed much better than the other methods described here. Therefore, the most widely used TMM method was chosen as the default data-normalization method in DREAMSeq and similar to previous studies (Robinson et al., 2010; Kadota et al., 2012; Soneson and Delorenzi, 2013; Sun et al., 2013).

### Dispersion Estimations

Estimating the dispersion parameter is a crucial step in DEG detection. Various dispersion-parameter estimation methods, including pseudo-likelihood (Smyth, 2003), quasi-likelihood (Nelder, 2000; Lund et al., 2012), conditional maximum likelihood (CML) (Smyth and Verbyla, 1996), quantile-adjusted CML (Robinson and Smyth, 2008), and shrinkage-estimation methods (Anders and Huber, 2010; Robinson et al., 2010), have been discussed previously. In particular, many Bayesian-based shrinkage-estimation methods, including baySeq, ShrinkSeq, DSS (Wu et al., 2013), and DESeq2, have been developed and are capable of obtaining accurate and robust estimates by sharing information across all genes when the sample size is small (Ji and Liu, 2010). Therefore, we also utilized an empirical Bayesian framework to shrink the dispersion parameter. Our strategy to estimate the dispersion parameter was divided into five steps described as follows.

#### Initial Dispersion Estimators

We first applied the method-of-moments (MoMs) described by Love et al. (2014) to estimate the initial value of dispersion for

each gene. According to previous studies (Anders and Huber, 2010; Robinson et al., 2010), we first use the normalized sample mean,  $\bar{X}_{ik}$ , to estimate the expectation for the  $i^{\text{th}}$  gene in group  $k$ :

$$\mu_{ik} = \frac{1}{J} \bar{X}_{ik} \sum_j s_{jk}. \quad (10)$$

We assume that the dispersions between two groups are the same under small sample sizes. Therefore, we denote  $n = KJ$  and substitute equation (10) with the following equation:

$$\mu_i = \frac{1}{n} \bar{X}_i \sum_n s_{jk}, \quad (11)$$

where  $\mu_i$  and  $\bar{X}_i$  are the expectation and sample mean, respectively, of the  $i^{\text{th}}$  gene. We then estimate the variance of the  $i^{\text{th}}$  gene,  $\sigma_i^2$ , by pooling count data from different groups using approaches previously described by Anders and Huber (2010) and Wu et al. (2013). For the NB model, the initial dispersion for the  $i^{\text{th}}$  gene can be estimated by:

$$\phi_i^{\text{init}} = \frac{\sigma_i^2 - \mu_i}{\mu_i^2}. \quad (12)$$

Note that  $\phi_i^{\text{init}}$  is often artificially assigned with an extremely low positive value (e.g.,  $1 \times 10^{-8}$  in DESeq) when  $\sigma_i^2 < \mu_i$ , because the NB model cannot fit underdispersed read-count data. A similar conservative strategy was also utilized for underdispersion in a previous study (Schissler et al., 2015). Under this scenario, the initial dispersion can be overestimated, which results in a conservative DEG test (Robinson and Smyth, 2008). By contrast, instead of the NB model, the DP model is capable of handling this kind of data. For the DP model, the initial dispersion for the  $i^{\text{th}}$  gene can be estimated by:

$$\theta_i^{\text{init}} = \frac{\mu_i}{\sigma_i^2}. \quad (13)$$

### Gene-Wise Dispersion Estimators

In RNA-seq experiments, there are typically tens of thousands of genes, but only a few replicates per treatment group, which describes the “large  $p$  and small  $n$ ” phenomenon. It is quite difficult to estimate a reliable gene-specific dispersion with the MoMs described in such a scenario. To address this problem, we used maximum likelihood estimate (MLE) methods based on the initial dispersion estimator,  $\phi_i^{\text{init}}$  (or  $\theta_i^{\text{init}}$ ), to estimate a gene-wise dispersion,  $\phi_i^{\text{genewise}}$  (or  $\theta_i^{\text{genewise}}$ ), for gene,  $i$ . The MLE of the dispersion parameters in the NB and DP models can be obtained by maximizing the log-likelihood summed over all reads between conditions for the  $i^{\text{th}}$  gene:

$$\phi_i^{\text{genewise}} = \operatorname{argmax}_{\phi} \left( \sum_n \log(f_{\text{NB}}(Y_{ijk}, \mu_{ik}, \phi)) \right) \quad (14)$$

and

$$\theta_i^{\text{genewise}} = \operatorname{argmax}_{\theta} \left( \sum_n \log(f_{\text{DP}}(Y_{ijk}, \mu_{ik}, \theta)) \right), \quad (15)$$

respectively, where  $\phi = \phi_i^{\text{init}}$ ,  $\theta = \theta_i^{\text{init}}$ , and  $f_{\text{NB}}(\cdot)$  and  $f_{\text{DP}}(\cdot)$  are the PMF of the NB and DP models, respectively.

### Common Dispersion Estimators

It is essential for reliable dispersion estimation that information is shared between genes, especially when few replicates are available (Robinson and Smyth, 2008). The simplest method of sharing information is to assume that the dispersion parameters are common for all genes and then to use the entire dataset to directly calculate a precise common dispersion. However, it is generally not true that each gene has the same dispersion in practice (Robinson and Smyth, 2007). Consequently, we should seek a more general common dispersion-estimation approach that compromises between entirely individual gene-wise dispersions and an entirely shared common dispersion. Here, we assumed that the dispersions are common across all genes having similar expression strengths, suggesting that if the means for some genes are similar, the dispersions (or variances) for these genes are also similar. We adopted a similar locally weighted regression as that for voom (Law et al., 2014) in order to obtain the common dispersion estimators ( $\phi_i^{\text{common}}$  for the NB model or  $\theta_i^{\text{common}}$  for the DP model) for the  $i^{\text{th}}$  gene by regressing the gene-wise dispersion estimators,  $\phi_i^{\text{genewise}}$  (or  $\theta_i^{\text{genewise}}$ ), onto the means,  $\mu_i$ , of the normalized read counts. This is similar to the data-driven parameter estimation used by DESeq through the smooth function by modeling the observed mean-variance (or mean-dispersion) relationship for the genes in the read-count data (Anders and Huber, 2010).

### Shrinkage-Dispersion Estimators

Shrinkage estimation can effectively improve statistical tests for differential gene expression in the case of a small number of samples (Cui et al., 2005). As mentioned previously, in order to obtain a more accurate and robust estimate, an empirical Bayes (EB) approach has been used to shrink gene-wise dispersions toward common dispersions, which could effectively allow the borrowing of information between genes (Robinson and Smyth, 2007; Robinson et al., 2010). The DSS and DESeq2 methods use an EB approach incorporating shrinkage with an NB model to squeeze the gene-wise dispersion estimates toward an LN prior, where the strength of shrinkage is dependent upon how reliably the individual gene-wise dispersions can be estimated (Wu et al., 2013; Love et al., 2014). Here, we assumed that the gene-wise dispersions,  $\alpha$ , followed an LN prior with two parameters: the mean,  $m_0$ , and the standard deviation (SD),  $\tau$ . The PMF of the LN model is given as:

$$P(\alpha | m_0, \tau) = \frac{1}{\alpha \sqrt{2\pi} \tau^2} e^{-\frac{(\log(\alpha) - m_0)^2}{2\tau^2}}, \quad (16)$$

where  $\alpha$  represents  $\phi_i^{\text{genewise}}$  and  $\theta_i^{\text{genewise}}$  for the NB and DP models, respectively. The two parameters of the LN model are estimated as follows:

$$m_0 = \operatorname{median}(\log(\beta)) \quad (17)$$



and

$$\tau = \text{mad}(\log(\alpha) - \log(\beta)), \quad (18)$$

respectively, where  $\text{mad}$  represents the median absolute deviation, and  $\beta$  represents  $\phi_i^{\text{common}}$  and  $\theta_i^{\text{common}}$  for the NB and DP models, respectively.

We adopted the same strategy as the DSS and DESeq2 methods to estimate the shrinkage dispersions for the  $i^{\text{th}}$  gene in the NB and DP models:

$$\phi_i^{\text{shrinkage}} = \text{argmax}_{\phi} \left( \sum_n \log(f_{\text{NB}}(Y_{ijk}, \mu_{ik}, \phi)) + f_{\text{LN}}(\phi, m_0, \tau) \right) \quad (19)$$

and

$$\theta_i^{\text{shrinkage}} = \text{argmax}_{\theta} \left( \sum_n \log(f_{\text{DP}}(Y_{ijk}, \mu_{ik}, \theta)) + f_{\text{LN}}(\theta, m_0, \tau) \right) \quad (20)$$

respectively, where  $\phi = \phi_i^{\text{genewise}}$ ,  $\theta = \theta_i^{\text{genewise}}$ , and  $f_{\text{NB}}(\cdot)$ ,  $f_{\text{DP}}(\cdot)$ , and  $f_{\text{LN}}(\cdot)$  are the PMF of the NB, DP, and LN models, respectively.

## Final Dispersion Estimators

Bias in dispersion estimation has serious effects on the expected false-positive rates (FPRs) in small-sample situations (Robinson and Smyth, 2008). To avoid bias, DESeq by default chooses the maximum value from the two dispersion estimators: the individual dispersion and the fitted dispersion as a final dispersion for the gene (Anders and Huber, 2010). However, DESeq is often overly conservative due to overestimation of the dispersion and results in conservation tests (Robinson and Smyth, 2008; Sonesson and Delorenzi, 2013). For this reason, we proposed a compromise approach called “window scan” to obtain the final dispersion estimators in five steps: (1) rank the genes from smallest to largest according to the means of samples across all conditions; (2) open a default 1-count window, where the mean is smallest; (3) based on the relationship between the shrinkage-dispersion estimator and the common-dispersion estimator, all genes in this window are divided into I-type genes (its shrinkage-dispersion estimator  $\geq$  its common dispersion estimator) and II-type gene (its shrinkage dispersion estimator  $<$  its common dispersion estimator); (4) estimate the final dispersion of each I-type gene (or II-type gene) by choosing the larger value between its shrinkage-dispersion estimator and the median of the shrinkage-dispersion estimators of all I-type genes (or II-type genes) for the NB model (or choosing the smaller value for the DP model); and (5) shift the window to the larger mean and repeat steps (3,4) until all of the genes are scanned.

## Test Statistic and Method Evaluation

### Test Statistic

For DEGs detected between two treatment groups, we tested the hypotheses of the form  $H_0: \mu_{i,1} = \mu_{i,2}$  for the gene  $i$ , where  $\mu_{i,1}$  and  $\mu_{i,2}$  are the expectations for the  $i^{\text{th}}$  gene in groups 1 and 2, respectively. The Wald test has been widely applied in many previous studies because of its simplicity and flexibility (Ng and

Tang, 2005; Chen et al., 2011; Yu et al., 2017). Similar to DSS and DESeq2, we constructed the Wald test statistic as:

$$W = \frac{|\mu_{i,1} - \mu_{i,2}|}{\sqrt{\sigma_{i,1}^2 + \sigma_{i,2}^2}}, \quad (21)$$

where  $\sigma_{i,1}^2$  and  $\sigma_{i,2}^2$  are the variances for the  $i^{\text{th}}$  gene in groups 1 and 2, respectively, and can be estimated using the final dispersion according to equation (3) in the NB model and equation (8) in the DP model.

## Method Evaluation

All methods analyzed will return nominal  $p$ -values. In order to obtain a more reliable list of DEGs, the  $p$ -values were adjusted by the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995). We evaluated the type I error rates (i.e., FPRs) and statistical powers (i.e., true-positive rates; TPRs) of different methods with a significance level of 0.05. Additionally, we used a receiver operating characteristic (ROC) curve, the area under the ROC curve (AUC), and a precision-recall curve (PRC) to compare the performances of eight methods in the simulated datasets. It is common for biologists to be interested in detecting genes with fold changes (FCs) estimated according to the ratios of the mean normalized counts between two treatment groups. Therefore, some methods use FC as an indicator of DE, such as DEGseq and AMAP.Seq (Si and Liu, 2013). Here, we defined the genes satisfying either  $\text{FC} < 0.67$  or  $\text{FC} > 1.5$ , and an adjusted  $p < 0.05$  as DEGs according to previous studies (Peart et al., 2005; Si and Liu, 2013). This quantitative filter combines the significance level with the FC threshold and might be considered more practical by biologists. Therefore, we also identified DEGs using this filter.

The performances of different methods were further validated by qRT-PCR analysis.

## Datasets

### Real Datasets

We chose three real datasets to represent different characteristics of RNA-seq data. The Pickrell dataset and the Hammer dataset were downloaded from the ReCount database (<http://bowtie-bio.sourceforge.net/recount>) (Frazee et al., 2011). The Pickrell dataset was obtained from lymphoblastoid cell lines derived from 69 unrelated Nigerian individuals as part of the International HapMap project (Pickrell et al., 2010) and contains 69 biological replicates. The Hammer dataset contains four biological replicates in each of two treatment groups: rat L4 dorsal-root-ganglion-treated groups in the presence or absence of induced chronic neuropathic pain (Hammer et al., 2010). The third real dataset was the Arab dataset provided as “arab” in the NBPSeq R package and that includes three biological replicates, where *Arabidopsis* leaves were inoculated with either a defense-eliciting  $\Delta hrcC$  mutant of *Pseudomonas syringae* pv. *tomato* DC3000 or 10 mM  $\text{MgCl}_2$  as a mock-treatment control (Di et al., 2011).



## Simulated Datasets

Simulation studies represent necessary processes for investigating the properties associated with certain statistical methods, given that the “true” DEGs are known in simulated data. An ideal simulation would generate data with similar characteristics to those produced in real RNA-seq experiments. Therefore, similar to Landau and Liu (2013), we generated three independent simulated datasets using a DP model based on three real datasets, respectively. The simulation processes were repeated 30 times to ensure reasonable precision in parameter estimation. Each simulated dataset contains 10,000 genes, including 2,000 DEGs and 8,000 non-DEGs, two treatment groups, and three replicates for each treatment group.

## Foxtail Dataset

Foxtail millet (*Setaria italica*) is an important cereal crop in northern China, and the whole-genome sequence of Foxtail millet (Yugu-1 cultivar) was published in 2012 (Bennetzen et al., 2012; Zhang et al., 2012). In this study, we used a Foxtail RNA-seq dataset obtained by our own laboratory to compare the performance of DREAMSeq with other methods. This Foxtail dataset includes three biological replicates, in which roots from 1-week-old Foxtail millet seedlings (Yugu-1 cultivar) were treated with or without 1  $\mu$ M epi-Brassinolide (eBL) for 2 h, followed by total RNA extraction using Trizol reagent (Invitrogen, Carlsbad, CA, United States). Extracted total RNA (2  $\mu$ g per sample) was sequenced on an Illumina HiSeq X-ten platform, and the remaining RNA was used for qRT-PCR validation. The paired-end reads were aligned to the Foxtail millet reference genome (JGIv2.0.34) (Bennetzen et al., 2012; Goodstein et al., 2012) using TopHat (version 2.0.12) (Trapnell et al., 2009; Kim et al., 2013), and gene read counts were obtained using the program htseq-count from the python package HTSeq (version 0.6.1) (Anders et al., 2015).

## qRT-PCR

First-strand cDNA was synthesized from 1  $\mu$ g total RNA using Reverse Transcriptase M-MLV (Takara Bio, Otsu, Japan) according to manufacturer instructions. qRT-PCR was performed according to the standard protocol using a Bio-Rad CFX Connect real-time PCR system (Bio-Rad Laboratories, Hercules, CA, United States). Primers used are listed in Table S1. The expression of target genes was normalized to Foxtail *Actin*, and the relative expression between treatment and control groups was averaged from three independent experiments, with the *p*-value calculated using a one-sample *t*-test. We defined genes satisfying relative expression  $>1.5$  or  $<0.67$  and  $p < 0.05$  as “true” DEGs.

## RESULTS

### The Mean–Variance Relationship in Real Datasets

When analyzing the Hammer, Arab, and Foxtail datasets, we found strong relationships between the variances and the means on the log-log scale for the read counts from different real datasets (Figure S1). For convenience of notation and

calculation, we used the unit line to represent a Poisson assumption-exhibited equidispersion. The data points on and above that line exhibit non-underdispersion, whereas the data points below that line exhibit underdispersion. Figure S1 shows that 2,606 of 18,635 genes (14.0%) in the Hammer dataset, 2,015 of 26,222 genes (7.7%) in the Arab dataset, and 4,412 of 35,158 genes (12.5%) in the Foxtail dataset were estimated as underdispersed genes. Therefore, there are a considerable proportion of underdispersed genes in the RNA-seq data. Furthermore, we noted that the underdispersed data points mostly distributed at low read-count regions (Figure S1). These results suggested that in addition to non-underdispersion, underdispersion also exists in RNA-seq data and should be properly handled during the RNA-seq data-mining process.

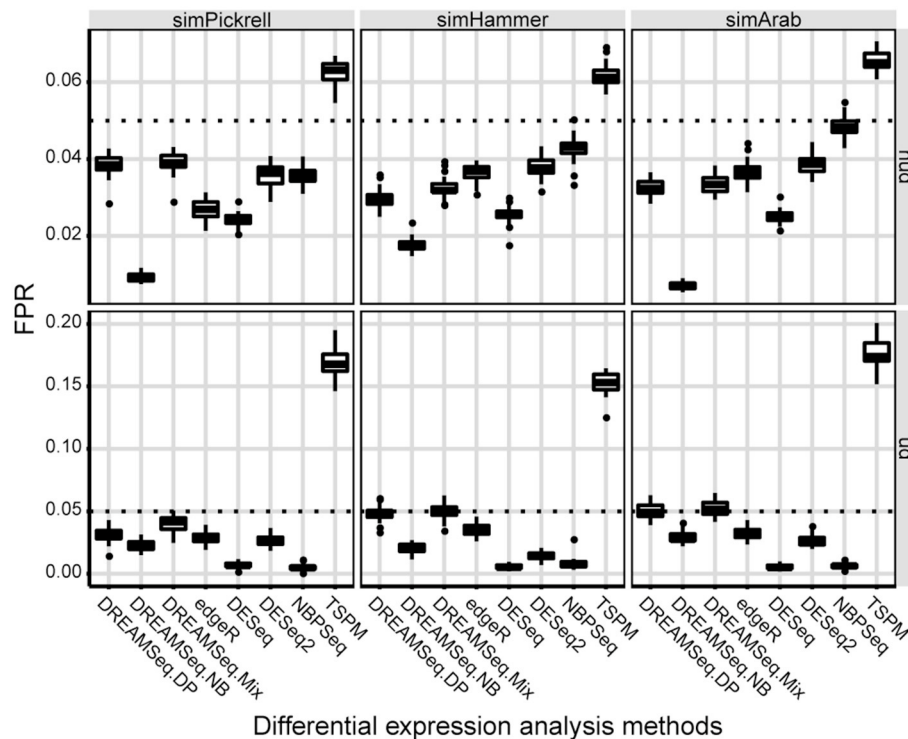
Most RNA-seq analysis methods were developed based on an NB model, which is able to capture both equidispersed and overdispersed data but not underdispersed data. In comparison, a DP model can capture all RNA-seq data (Efron, 1986). Using real Hammer, Arab, and Foxtail datasets, we found that both DP and NB models were able to fit read-count data very well (Figure S2). This suggested that the DP model can be used to mine RNA-seq data.

## Generation of Simulated Datasets

Wu et al. (2013) reported that using real data-driven simulations provided a better estimate for gene-wise dispersions and improved DEG detection, because the true DE status of each gene is known by controlling the settings (Wu et al., 2013). Therefore, we generated three simulated datasets with mean and dispersion parameters estimated from three real datasets based on a commonly used DP model and denoted these as simPickrell, simHammer, and simArab, respectively. The average number of underdispersed genes in simPickrell, simHammer, and simArab was 1299 (13%), 1935 (19%), and 1432 (14%), respectively. As shown in Figure S3, all simulated datasets were very similar to the corresponding real datasets in terms of distributions of the means and dispersions and relationships between means and dispersions. This indicated that our simulated data closely mimicked the real data.

## Type I Error Rate

Using the three simulated datasets, we first evaluated the type I error rates (i.e., FPRs) of the three DREAMSeq methods (DREAMSeq.DP, DREAMSeq.NB, and DREAMSeq.Mix) and five other widely used RNA-seq data-analysis methods (edgeR, DESeq, DESeq2, NBPSeq, and TSPM) under the null hypothesis. We found that except for TSPM, all other methods were able to control type I error rates well in both non-underdispersion and underdispersion situations (Figure 1). In comparison, DESeq was very conservative in term of type I error rate, whereas the abilities of FPR control by both DREAMSeq.NB and NBPSeq clearly varied between non-underdispersion and underdispersion situations. In contrast, the median FPRs of DREAMSeq.DP, DREAMSeq.Mix, edgeR, and DESeq2 were relatively stable and



**FIGURE 1 |** Comparison of type I error rates between different methods. Boxplots show the type I error rates (i.e., FPRs) of different methods, which were calculated over 30 simulations for the simPickrell, simHammer, and simArab datasets under the null hypothesis. The horizontal dotted lines indicate the nominal type I error rate of 0.05 in non-underdispersion and underdispersion scenarios. nud, non-underdispersion; ud, underdispersion.

consistently lower than or very close to the nominal type I error rate of 0.05 under all situations.

## Statistical Power, ROC, AUC, PRC, and Number of DEGs

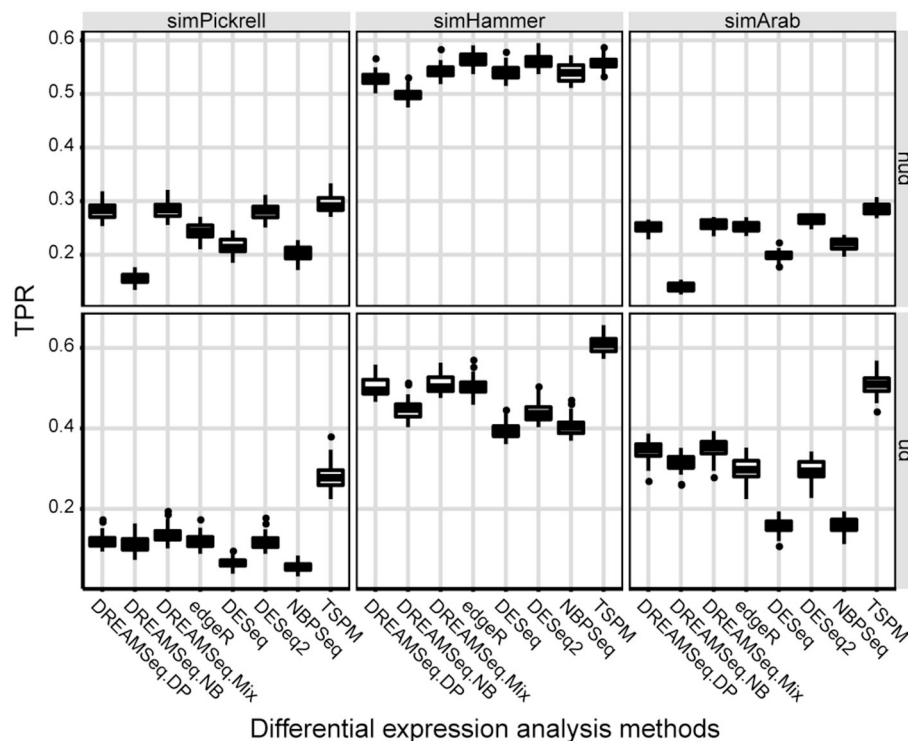
We then evaluated the statistical powers (i.e., TPRs) of different methods using the simulated datasets under the alternative hypothesis (**Figure 2**). The results showed that in underdispersion situations, the TPR of DREAMSeq.Mix was slightly higher than that of DREAMSeq.DP, although that of both methods was higher than those of DREAMSeq.NB, edgeR, DESeq, DESeq2, and NBPSeq (**Figure 2**). In non-underdispersion situations, the TPRs of DREAMSeq.Mix and DREAMSeq.DP were comparable with the other methods. Interestingly, TSPM consistently showed higher TPRs. Given that TSPM also showed higher FPRs in similar situations, it is likely that the TSPM method increased statistical power at the cost of poor FPR control.

The ROC curve was constructed using the TPR to FPR ratio for each method used for DE analysis. Theoretically, the method with the stronger statistical power at identifying DEGs should exhibit a ROC curve with a higher TPR relative to other methods at the same FPR level. **Figure S4** shows that NBPSeq and TSPM had lower TPRs when the FPR threshold was  $\sim 0.05$  in each scenario, whereas the ROC curves of the other methods were very similar. Additionally, we found that

ROC curves associated with the simHammer dataset were steeper than those for the simPickrell and simArab datasets, suggesting that the performance of DEG identification by different methods was strongly dependent upon innate data characteristics, such as heterogeneity.

AUC is a relative measure of the quality of a DEG test, where a higher AUC indicates relatively better performance. To quantify the performances of different methods in detecting DEGs, AUCs of different methods were calculated. The result showed that the AUCs of DREAMSeq.DP and DREAMSeq.Mix were higher than those of DREAMSeq.NB, edgeR, DESeq, DESeq2, and NBPSeq in most of the situations, except slightly lower than DESeq2 when analyzing simHammer and simArab underdispersed data (**Figure 3**). Together with the above FPR, TPR, and ROC results, these findings clearly demonstrated that both DREAMSeq.DP and DREAMSeq.Mix were able to control type I error rates well while maintaining a relatively higher statistical power in detecting DEGs.

PRC curve shows the precision for corresponding recall (TPR). Similar to the ROC curve, the PRC curve is also an important performance indicator used to evaluate different methods at identifying DEGs. **Figure S5** shows that all methods, except TSPM, had higher precision over the entire range of recall rates, regardless of dataset or dispersion. Additionally, we found that all methods exhibited their best predictive performance using the simHammer dataset, but did not predict very accurately



**FIGURE 2 |** Statistical power comparison between different methods. Boxplots show the statistical powers (i.e., TPRs) of different methods and calculated over 30 simulations for the simPickrell, simHammer, and simArab datasets under the alternative hypothesis in non-underdispersion and underdispersion scenarios. nu, non-underdispersion; u, underdispersion.

using the simPickrell dataset in an underdispersion situation, which might also be related to the dataset itself.

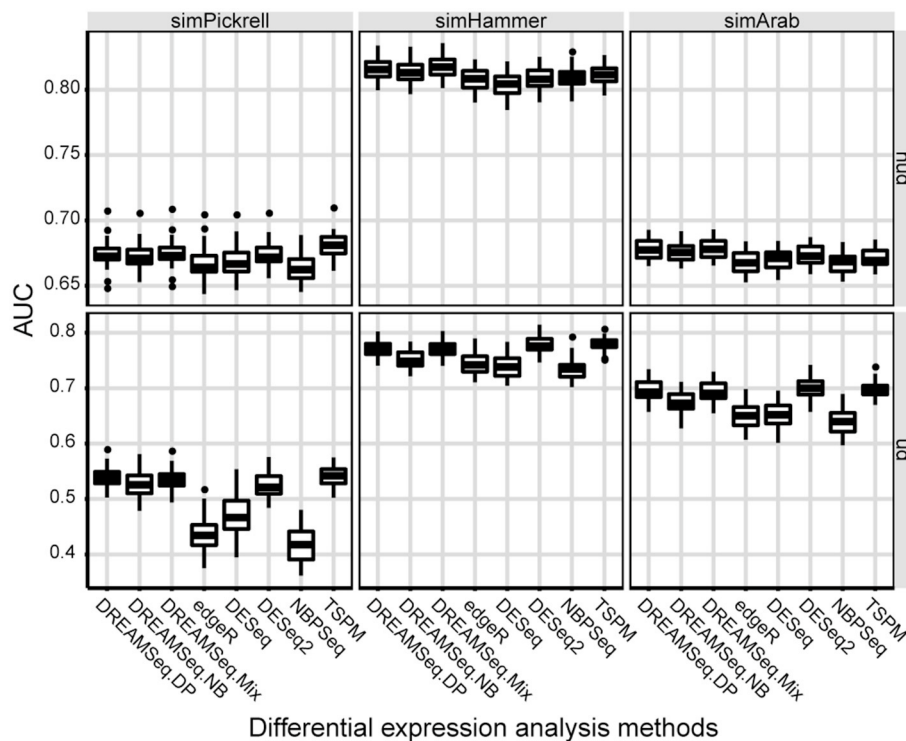
We also compared the identified DEG numbers of different methods, with the results showing that both DREAMSeq.DP and DREAMSeq.Mix generally detected a larger number of DEGs (except in the case of simHammer non-underdispersed data) than the other methods (except for TSPM, which displayed poor FDR control) when analyzing non-underdispersed or underdispersed data from three simulated datasets, respectively, (Figure 4).

## Analysis of the Foxtail Dataset

Our comprehensive evaluations showed that edgeR, DESeq, DESeq2, and DREAMSeq.Mix generally performed better as analyzing different simulated RNA-seq datasets; therefore, these methods were chosen to test their abilities to detect DEGs, especially underdispersed DEGs, using a real Foxtail dataset. A total of 128 non-underdispersed and 17 underdispersed DEGs were identified by at least one of the four methods (Figure 5 and Tables S2–S5). Overall, the number of DEGs identified by DREAMSeq.Mix was much higher than that by DESeq but lower than that by edgeR and DESeq2 (Figure 5A). However, DREAMSeq.Mix identified 15 underdispersed DEGs, whereas edgeR identified 12, and DESeq2 identified 9 underdispersed DEGs. We defined DEGs detected only by one method as unique DEGs. Notably, DREAMSeq.Mix detected the highest number

of unique DEGs in underdispersion scenarios, whereas DESeq did not identify any unique DEGs in either non-underdispersion or underdispersion scenarios (Figures 5B,C). Consistent with previous reports (Seyednasrollah et al., 2013; Tang et al., 2015), all of the DEGs found by DESeq were also found by edgeR (Figures 5B,C), possibly because these two methods use the same statistical model (i.e., the NB model) and hypothesis testing procedure (i.e., the Robinson and Smyth exact test) (Robinson and Smyth, 2008; Anders and Huber, 2010; Robinson et al., 2010). The presence of various unique DEGs also suggested the advantage of using more than one method to analyze the same RNA-seq data in order to allow maximum discovery of DEGs.

We then used qRT-PCR to validate whether the DEGs identified from the Foxtail dataset were “true” DEGs. Because DEGs identified by DESeq were also identified by edgeR, the unique DEGs identified by either edgeR, DESeq2, or DREAMSeq.Mix and the common DEGs identified simultaneously by any two methods were chosen for qRT-PCR analysis (Figure 6). The results showed that most of the DEGs chosen for validation exhibited similar upregulation or downregulation patterns as those shown from RNA-seq data analysis. For non-underdispersed DEGs, qRT-PCR results verified that 9 of 19 DEGs (47.4%) identified by DREAMSeq.Mix, 19 of 42 DEGs (45.2%) identified by edgeR, and 23 of 51 DEGs (45.1%) identified by DESeq2 were significantly upregulated or downregulated by eBL treatment by at least 1.5-fold. Notably,



**FIGURE 3 |** Comparison of AUCs between different methods. Boxplots show the AUCs of different methods and calculated over 30 simulations for the simPickrell, simHammer, and simArab datasets in non-underdispersion and underdispersion scenarios. nud, non-underdispersion; ud, underdispersion.

for underdispersed DEGs, 5 of 8 (62.5%) DEGs identified by DREAMSeq.Mix were validated as “true” DEGs. By contrast, only 2 of 5 (40.0%) DEGs identified by edgeR and no DEGs identified by DESeq2 were validated as “true” DEGs. These qRT-PCR results demonstrated that for non-underdispersed data, the number of DEGs identified by DREAMSeq.Mix was lower than those by edgeR and DESeq2, but the accuracy was slightly higher; however, for underdispersed data, DREAMSeq.Mix exhibited both a higher number of identified DEGs and better accuracy than the other two methods, demonstrating DREAMSeq.Mix as a powerful RNA-seq data-analysis method, especially for situations involving underdispersed data.

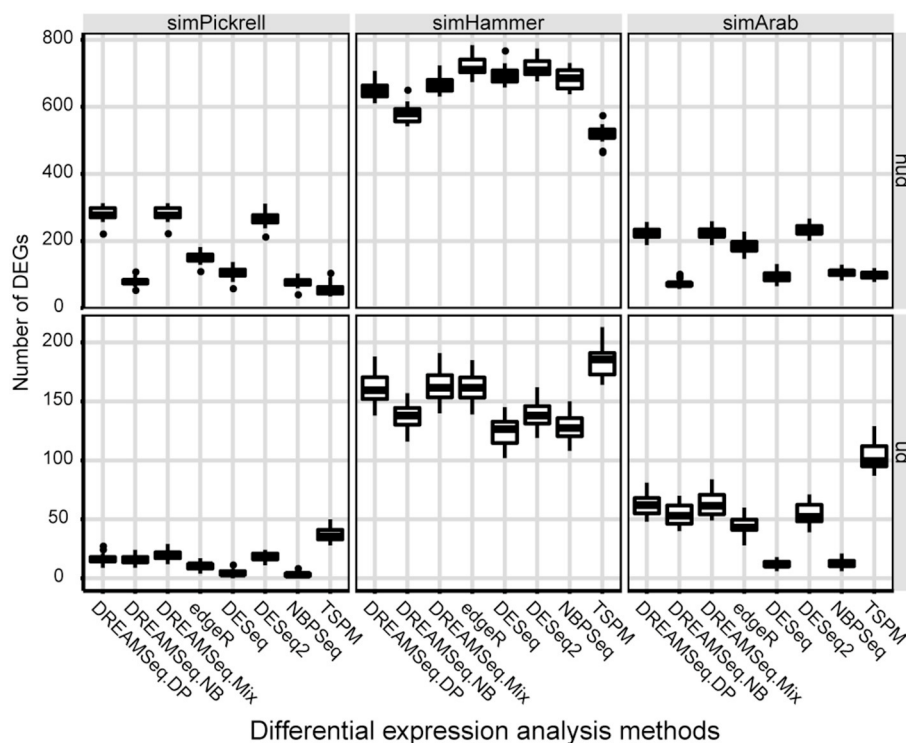
## DISCUSSION

RNA-seq is an increasingly popular method used to analyze global changes in gene expression during certain biological processes. Identifying DEGs is a key step in mining RNA-seq data and important for downstream biological analyses, such as cluster analysis, PCA analysis, GO analysis, and Kyoto Encyclopedia of Genes and Genomes enrichment analysis. When analyzing RNdA-seq data, most current methods focus on non-underdispersed data, with less attention given to underdispersed data. In this study, we observed that RNA-seq data also includes underdispersion characteristics. Additionally, Low et al. (2017) found that as the RNA-seq coverage increases, underdispersion becomes increasingly obvious. With the

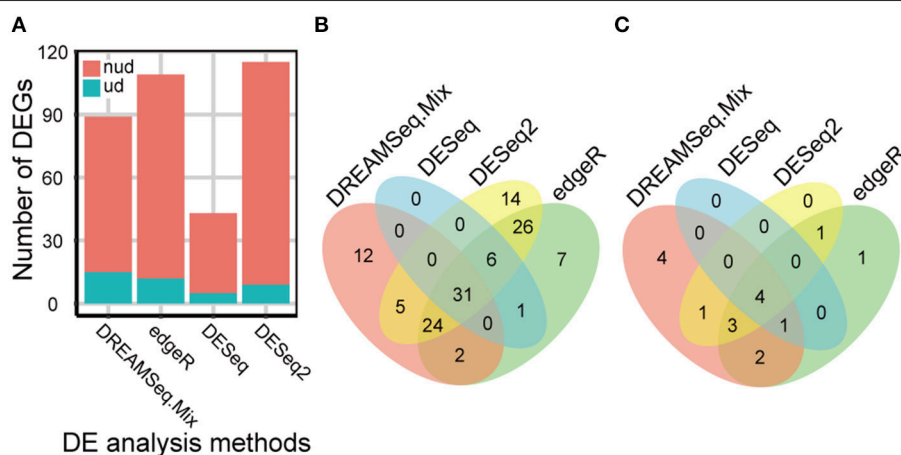
development of sequencing technology, the read length and RNA-seq coverage have increased significantly. Therefore, to take full advantage of RNA-seq data, it is important to explore both non-underdispersed and underdispersed data. However, most widely used DE-analysis methods, such as DESeq and edgeR, are based on the NB model. Due to the limitations of this model, underdispersed data are often overestimated, leading to conservative results in the determination of DEGs. In comparison, the DP model is capable of capturing not only non-underdispersion but also underdispersion. Considering the potential advantages of these two models, we developed a novel RNA-seq data-mining method (DREAMSeq.Mix) that combines the DP and NB models.

Using simulated datasets generated from three real RNA-seq experiments, we compared the performance of DREAMSeq.Mix at detecting DEGs with five other commonly used RNA-seq data-analysis methods. To provide a more comprehensive conclusion, we also added DREAMSeq.DP and DREAMSeq.NB methods, which were developed using only a DP model or an NB model, respectively, into the comparison. We found that DESeq, NBPSeq, and DREAMSeq.NB were often conservative, whereas TSPM, edgeR, and DESeq2 were more liberal in detecting DEGs. The poor performance of TSPM in our study might be due to the limited number of replicates in the RNA-seq datasets used (Auer and Doerge, 2011; Kvam et al., 2012; Sonesson and Delorenzi, 2013). In comparison, DREAMSeq.DP and DREAMSeq.Mix





**FIGURE 4 |** Comparison of the number of DEGs identified by different methods. Boxplots show the number of DEGs identified by different methods and calculated over 30 simulations for the simPickrell, simHammer, and simArab datasets in non-underdispersion and underdispersion scenarios. nud, non-underdispersion; ud, underdispersion.

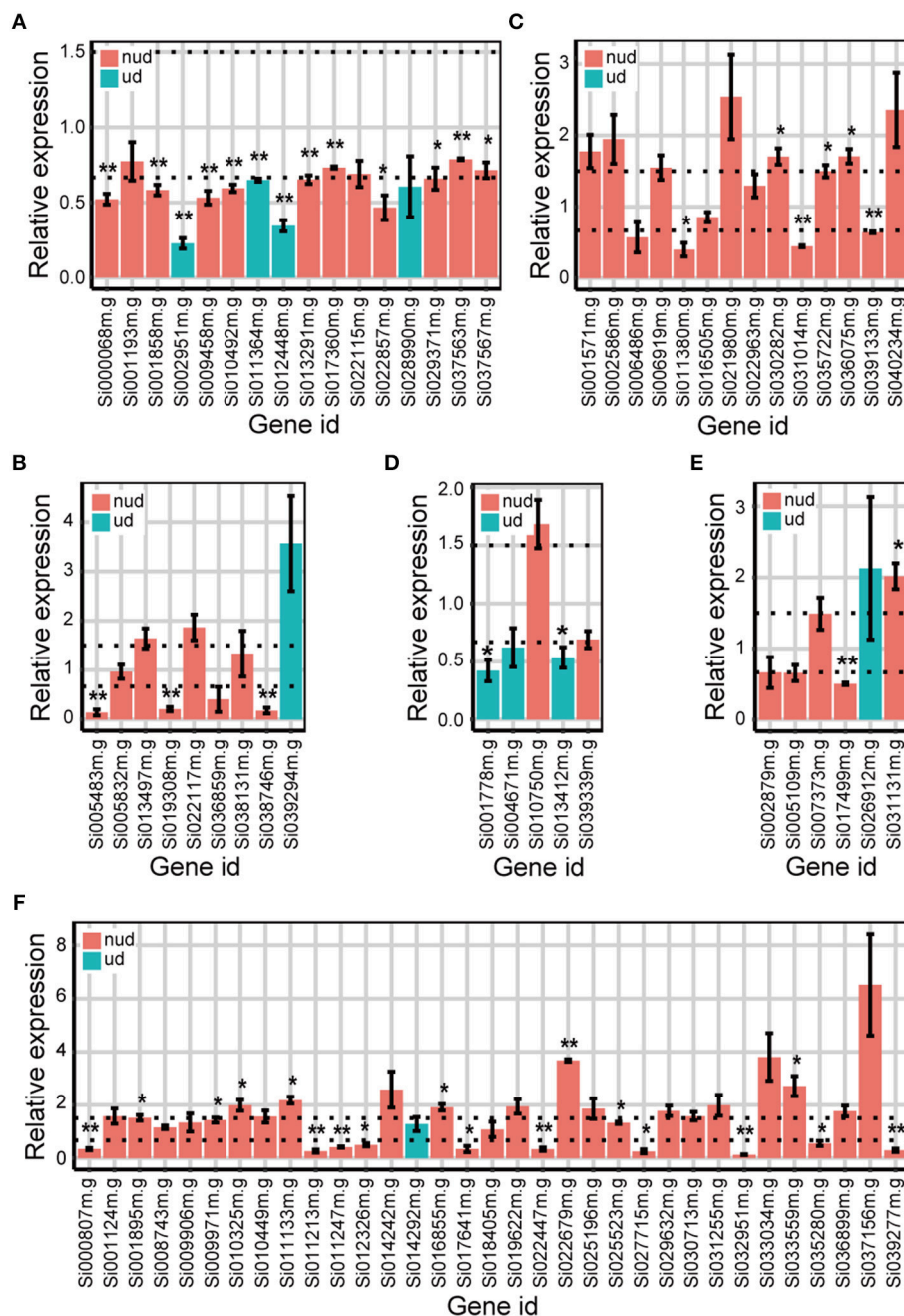


**FIGURE 5 |** eBL-regulated Foxtail millet-root DEGs identified by different methods. (A) Bar plot showing the number of eBL-regulated DEGs identified by DREAMSeq.Mix, edgeR, DESeq, and DESeq2. (B,C) Venn diagrams showing the overlap among the collections of eBL-regulated DEGs identified by DREAMSeq.Mix, edgeR, DESeq, and DESeq2 in non-underdispersion (B) and underdispersion (C) scenarios. nud, non-underdispersion; ud, underdispersion.

often outperformed the other methods in terms of TPR, AUC, and the number of DEGs detected (Figures 2–4). The following reasons suggest that DREAMSeq.Mix provided unique and important outcomes more advantageous than current RNA-seq data-mining methods.

First, DREAMSeq incorporates a more flexible DP model to fit highly complex and variable RNA-seq data. The dispersion parameter of the DP model is not subject to the same restrictions as the NB model when it is estimated in underdispersion situations. As a result, logarithmic dispersion estimated using





**FIGURE 6 |** qRT-PCR validation of the expression of eBL-regulated Foxtail DEGs detected by different methods. Bar plots show the relative expression of DEGs detected only by DREAMSeq.Mix (A), edgeR (B), and DESeq2 (C) or identified by DREAMSeq.Mix and edgeR (D), DREAMSeq.Mix and DESeq2 (E), or edgeR and DESeq2 (F), respectively, in eBL-treated Foxtail millet roots. The relative expression levels were normalized to the Foxtail millet *Actin* gene. Data represent the mean  $\pm$  SE of three independent experiments. *P*-values were calculated using a one-sample *t*-test. \**P* < 0.05; \*\**P* < 0.01. The horizontal dotted lines indicate relative expression of 1.5 or 0.67. nud, non-underdispersion; ud, underdispersion.

the DP model (Figure S3) showed a better normality than that acquired using the NB model (Figure 1 in Landau and Liu, 2013). This demonstrated that the DP model was able to accurately fit a widely range of read-count data without artificial intervention in RNA-seq data analysis. Therefore,

DREAMSeq.DP and DREAMSeq.Mix often outperformed the other methods, especially in underdispersion situations, in simulation studies. Moreover, in terms of identifying the “true” underdispersed DEGs, DREAMSeq.Mix outperformed edgeR, DESeq, and DESeq2 according to qRT-PCR validation.

Second, DREAMSeq incorporates strategies, such as MoMs, MLE, and EB, which are used in the edgeR, DESeq, DSS, and DESeq2 methods, to obtain reliable dispersion estimation. Importantly, to avoid bias, DREAMSeq used a “window scan” approach to estimate dispersion and enhance DREAMSeq’s robustness in analyzing a wider range of RNA-seq data. This enabled all DREAMSeq approaches maintain a higher AUC across different simulated datasets in either non-underdispersion or underdispersion scenarios.

Third, in multiple scenarios, DREAMSeq.Mix performed slightly better than DREAMSeq.DP, although the difference was small. This indicated that the efficiency and robustness of DREAMSeq.Mix was improved by taking full potential of the advantages of the DP and NB models to fit RNA-seq data.

Recently, single-cell RNA-seq (scRNA-seq) has rapidly become a powerful tool for analyzing gene-expression heterogeneity at the individual cell level and been widely applied to diverse fields of biological research, including stem cell differentiation, embryogenesis, and whole-tissue analysis (Saliba et al., 2014). However, scRNA-seq data displays typical features of bimodality (the NB model cannot capture bimodality) (Vu et al., 2016), making such data less efficient for mining using common RNA-seq data-analysis methods. Additionally, Choo-Wosoba et al. (2016) reported that genomic next-generation sequencing data also involves underdispersion. The increased accuracy and robustness displayed in finding “true” DEGs with higher confidence and its better performance at exploring underdispersed data make DREAMSeq a potentially valuable tool for mining sequencing data generated from many other high-throughput platforms, such as scRNA-seq and genomic sequencing.

During our analysis, we found that none of the eight tested methods consistently outperformed other methods under all situations, because different methods are capable of identifying specific groups of DEGs. Although some DEGs can be identified by all methods, the existence of unique DEGs suggested that different methods exhibited specific preferences during DEG detection. Additionally, our study showed that the same method sometimes displayed a wide range of performance variability when analyzing different datasets. It is likely that the intrinsic characteristics of the RNA-seq data determine the appropriateness of one method for data analysis over others. Therefore, to ensure maximum coverage of DEG identification, it is advantageous to use more than one method to analyze the same RNA-seq data. Based on our comparison studies, we recommend that using a combination of edgeR, DESeq2, and DREAMSeq.Mix for RNA-seq data analysis to potentially ensure the maximum retrieval of true DEGs in both non-underdispersion and underdispersion situations.

## CONCLUSIONS

Previous studies reported both equidispersion and overdispersion as important characteristics of RNA-seq

data. In this study, we showed that underdispersion also exists in RNA-seq data. The NB model widely used in RNA-seq data-mining methods can only capture non-underdispersion but not underdispersion. Here, we presented a DP model capable of capturing not only non-underdispersion but also underdispersion. Given the potential advantages of the two models, we developed a novel RNA-seq data-mining method (DREAMSeq) that combines both the DP and NB models to ensure its flexibility and robustness for RNA-seq data mining. Additionally, we used a “window scan” approach to estimate dispersion and enhance the reliability of DREAMSeq across a wider range of RNA-seq data. Using simulated datasets generated from three real RNA-seq datasets and an in-house-generated Foxtail dataset, we demonstrated the ability of DREAMSeq to reach a better balance between conservative and liberal tests as compared with other methods. Our findings demonstrated DREAMSeq as a reliable and robust RNA-seq data-analysis method that provides important improvements in the DE analysis of RNA-seq data, especially in underdispersion situations.

## DATA AVAILABILITY

DREAMSeq R package (version 1.0, Windows binary release) is available publicly (<http://tanglab.hebtu.edu.cn/tanglab/Home/DREAMSeq>). This package also contains a real Foxtail dataset obtained by our own laboratory.

## AUTHOR CONTRIBUTIONS

WT and ZG designed the research; ZG wrote the DREAMSeq R package and performed all data analyses; ZZ performed Foxtail RNA-seq and qRT-PCR experiments; and WT and ZG wrote the manuscript.

## FUNDING

This work was supported by grants from the National Natural Science Foundation of China (91417313, 2014CB943404, and 31670265) and the Science Foundation of Hebei University of Economics and Business (2013KYZ05).

## ACKNOWLEDGMENTS

We would like to thank Dr. Hong Zhang (School of Life Sciences, Fudan University) for valuable discussion and suggestion for this work.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2018.00588/full#supplementary-material>

## REFERENCES

- Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106
- Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi: 10.1093/bioinformatics/btu638
- Auer, P. L., and Doerge, R. W. (2011). A two-stage Poisson model for testing RNA-seq data. *Stat. Appl. Genet. Mol. Biol.* 10:26. doi: 10.2202/1544-6115.1627
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* 57, 289–300.
- Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., et al. (2012). Reference genome sequence of the model plant *Setaria*. *Nat. Biotechnol.* 30, 555–561. doi: 10.1038/nbt.2196
- Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193. doi: 10.1093/bioinformatics/bt19.2.185
- Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94. doi: 10.1186/1471-2105-11-94
- Chen, Z., Liu, J., Ng, H. K., Nadarajah, S., Kaufman, H. L., Yang, J. Y., et al. (2011). Statistical methods on detecting differentially expressed genes for RNA-seq data. *BMC Syst. Biol.* 5:S1. doi: 10.1186/1752-0509-5-S3-S1
- Choo-Wosoba, H., Levy, S. M., and Datta, S. (2016). Marginal regression models for clustered count data based on zero-inflated Conway–Maxwell–Poisson distribution with applications. *Biometrics* 72, 606–618. doi: 10.1111/biom.12436
- Chu, C., Fang, Z., Hua, X., Yang, Y., Chen, E., Cowley, A. W. Jr., et al. (2015). deGPS is a powerful tool for detecting differential expression in RNA-sequencing studies. *BMC Genomics* 16:455. doi: 10.1186/s12864-015-1676-0
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13. doi: 10.1186/s13059-016-0881-8
- Consortium, M. (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* 28, 827–838. doi: 10.1038/nbt.1665
- Cui, X., Hwang, J. T., Qiu, J., Blades, N. J., and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* 6, 59–75. doi: 10.1093/biostatistics/kxh018
- Di, Y., Schafer, D. W., Cumbie, J. S., and Chang, J. H. (2011). The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.* 10:24. doi: 10.2202/1544-6115.1637
- Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2012). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* 14, 671–683. doi: 10.1093/bib/bbs046
- Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* 81, 709–721. doi: 10.1080/01621459.1986.10478327
- Famoye, F. (1993). Restricted generalized Poisson regression model. *Comm. Statist. Theory Methods* 22, 1335–1354. doi: 10.1080/03610929308831089
- Frazee, A. C., Langmead, B., and Leek, J. T. (2011). ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics* 12:449. doi: 10.1186/1471-2105-12-449
- Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–1186. doi: 10.1093/nar/gkr944
- Hammer, P., Banck, M. S., Amberg, R., Wang, C., Petznick, G., Luo, S., et al. (2010). mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. *Genome Res.* 20, 847–860. doi: 10.1101/gr.101204.109
- Hardcastle, T. J., and Kelly, K. A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 11:422. doi: 10.1186/1471-2105-11-422
- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249
- Ji, H., and Liu, X. S. (2010). Analyzing 'omics data using hierarchical models. *Nat. Biotechnol.* 28, 337–340. doi: 10.1038/nbt.1619
- Kadota, K., Nishiyama, T., and Shimizu, K. (2012). A normalization strategy for comparing tag count data. *Algorithms Mol. Biol.* 7:5. doi: 10.1186/1748-7188-7-5
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36
- Kvam, V. M., Liu, P., and Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.* 99, 248–256. doi: 10.3732/ajb.1100340
- Landau, W. M., and Liu, P. (2013). Dispersion estimation and its effect on test performance in RNA-seq data analysis: a simulation-based comparison of methods. *PLoS ONE* 8:e81415. doi: 10.1371/journal.pone.0081415
- Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15:R29. doi: 10.1186/gb-2014-15-2-r29
- Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., et al. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* 29, 1035–1043. doi: 10.1093/bioinformatics/btt087
- Li, J., and Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* 22, 519–536. doi: 10.1177/0962280211428386
- Lin, Y., Golovnina, K., Chen, Z. X., Lee, H. N., Negron, Y. L., Sultana, H., et al. (2016). Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. *BMC Genomics* 17:28. doi: 10.1186/s12864-015-2353-z
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8
- Low, J. Z., Khang, T. F., and Tammi, M. T. (2017). CORNAS: coverage-dependent RNA-Seq analysis of gene expression data without biological replicates. *BMC Bioinformatics* 18:575. doi: 10.1186/s12859-017-1974-4
- Lund, S. P., Nettleton, D., McCarthy, D. J., and Smyth, G. K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat. Appl. in Genet. and Mol. Biol.* 11:8. doi: 10.1515/1544-6115.1826
- LuValle, M. J. (1990). Generalized Poisson distributions: properties and applications. *Technometrics* 32, 346–347. doi: 10.1080/00401706.1990.10484695
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18, 1509–1517. doi: 10.1101/gr.079558.108
- Mi, G., Di, Y., and Schafer, D. W. (2015). Goodness-of-fit tests and model diagnostics for negative binomial regression of RNA sequencing data. *PLoS ONE* 10:e0119254. doi: 10.1371/journal.pone.0119254
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226
- Nelder, J. (2000). Quasi-likelihood and pseudo-likelihood are not the same thing. *J. Appl. Statist.* 27, 1007–1011. doi: 10.1080/02664760050173328
- Ng, H. K., and Tang, M. L. (2005). Testing the equality of two Poisson means using the rate ratio. *Stat. Med.* 24, 955–965. doi: 10.1002/sim.1949
- Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biol.* 11:220. doi: 10.1186/gb-2010-11-12-220
- Pear, M. J., Smyth, G. K., van Laar, R. K., Bowtell, D. D., Richon, V. M., Marks, P. A., et al. (2005). Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 102, 3697–3702. doi: 10.1073/pnas.0500369102
- Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., et al. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–772. doi: 10.1038/nature08872

- Rau, A., Celeux, G., Martin-Magniette, M.-L., and Maugis-Rabusseau, C. (2011). *Clustering High-Throughput Sequencing Data With Poisson Mixture Models*. Research Report RR-7786, Inria Saclay, Ile-de-France.
- Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902. doi: 10.1038/nbt.2931
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616
- Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25. doi: 10.1186/gb-2010-11-3-r25
- Robinson, M. D., and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23, 2881–2887. doi: 10.1093/bioinformatics/btm453
- Robinson, M. D., and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9, 321–332. doi: 10.1093/biostatistics/kxm030
- Saliba, A. E., Westermann, A. J., Gorski, S. A., and Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* 42, 8845–8860. doi: 10.1093/nar/gku555
- Schissler, A. G., Gardeux, V., Li, Q., Achour, I., Li, H., Piegorsch, W. W., et al. (2015). Dynamic changes of RNA-sequencing expression for precision medicine: N-of-1-pathways Mahalanobis distance within pathways of single subjects predicts breast cancer survival. *Bioinformatics* 31, i293–302. doi: 10.1093/bioinformatics/btv253
- Seyednasrollah, F., Laiho, A., and Elo, L. L. (2013). Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.* 16, 59–70. doi: 10.1093/bib/bbt086
- Si, Y., and Liu, P. (2013). An optimal test with maximum average power while controlling FDR with application to RNA-seq data. *Biometrics* 69, 594–605. doi: 10.1111/biom.12036
- Smyth, G. K. (2003). Pearson's goodness of fit statistic as a score test statistic. *Lect. Notes Monogr. Ser.* 40, 115–126. doi: 10.1214/lnms/1215091138
- Smyth, G. K., and Verbyla, A. P. (1996). A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *J. R. Stat. Soc. Series B Methodol.* 58, 565–572.
- Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14:91. doi: 10.1186/1471-2105-14-91
- Srivastava, S., and Chen, L. (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.* 38:e170. doi: 10.1093/nar/gkq670
- Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321, 956–960. doi: 10.1126/science.1160342
- Sun, J., Nishiyama, T., Shimizu, K., and Kadota, K. (2013). TCC: an R package for comparing tag count data with robust normalization strategies. *BMC Bioinformatics* 14:219. doi: 10.1186/1471-2105-14-219
- Tang, M., Sun, J., Shimizu, K., and Kadota, K. (2015). Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. *BMC Bioinformatics* 16:361. doi: 10.1186/s12859-015-0794-7
- Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Res.* 21, 2213–2223. doi: 10.1101/gr.124321.111
- Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120
- Van De Wiel, M. A., Leday, G. G., Pardo, L., Rue, H., Van Der Vaart, A. W., and Van Wieringen, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* 14, 113–128. doi: 10.1093/biostatistics/kxs031
- Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen, M., et al. (2016). Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* 32, 2128–2135. doi: 10.1093/bioinformatics/btw202
- Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010a). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26, 136–138. doi: 10.1093/bioinformatics/btp612
- Wang, L., Li, P., and Brutnell, T. P. (2010b). Exploring plant transcriptomes using ultra high-throughput sequencing. *Brief. Funct. Genomics* 9, 118–128. doi: 10.1093/bfgp/elp057
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484
- Wu, H., Wang, C., and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* 14, 232–243. doi: 10.1093/biostatistics/kxs033
- Yu, L., Fernandez, S., and Brock, G. (2017). Power analysis for RNA-Seq differential expression studies. *BMC Bioinformatics* 18:234. doi: 10.1186/s12859-017-1648-2
- Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., et al. (2012). Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.* 30, 549–554. doi: 10.1038/nbt.2195
- Zhang, H., Xu, J., Jiang, N., Hu, X., and Luo, Z. (2015). PLNseq: a multivariate Poisson lognormal distribution for high-throughput matched RNA-sequencing read count data. *Stat. Med.* 34, 1577–1589. doi: 10.1002/sim.6449
- Zhou, Y. H., Xia, K., and Wright, F. A. (2011). A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics* 27, 2672–2678. doi: 10.1093/bioinformatics/btr449

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Gao, Zhao and Tang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# CircCode: A Powerful Tool for Identifying circRNA Coding Ability

Peisen Sun<sup>1,2</sup> and Guanglin Li<sup>1,2\*</sup>

<sup>1</sup> Key Laboratory of Ministry of Education for Medicinal Plant Resource and Natural Pharmaceutical Chemistry, Shaanxi Normal University, Xi'an, China, <sup>2</sup> College of Life Sciences, Shaanxi Normal University, Xi'an, China

## OPEN ACCESS

### Edited by:

Filippo Geraci,  
Italian National Research Council,  
(CNR) Italy

### Reviewed by:

Wojciech M. Karlowski,  
Adam Mickiewicz University in  
Poznań, Poland  
Cuncong Zhong,  
University of Kansas,  
United States

### \*Correspondence:

Guanglin Li  
glli@snnu.edu.cn

### Specialty section:

This article was submitted to  
Bioinformatics and  
Computational Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 21 June 2019

**Accepted:** 13 September 2019

**Published:** 10 October 2019

### Citation:

Sun P and Li G (2019) CircCode:  
A Powerful Tool for Identifying  
circRNA Coding Ability.  
Front. Genet. 10:981.  
doi: 10.3389/fgene.2019.00981

Circular RNAs (circRNAs), which play vital roles in many regulatory pathways, are widespread in many species. Although many circRNAs have been discovered in plants and animals, the functions of these RNAs have not been fully investigated. In addition to the function of circRNAs as microRNA (miRNA) decoys, the translation potential of circRNAs is important for the study of their functions; yet, few tools are available to identify their translation potential. With the development of high-throughput sequencing technology and the emergence of ribosome profiling technology, it is possible to identify the coding ability of circRNAs with high sensitivity. To evaluate the coding ability of circRNAs, we first developed the CircCode tool and then used CircCode to investigate the translation potential of circRNAs from humans and *Arabidopsis thaliana*. Based on the ribosome profile databases downloaded from NCBI, we found 3,610 and 1,569 translated circRNAs in humans and *A. thaliana*, respectively. Finally, we tested the performance of CircCode and found a low false discovery rate and high sensitivity for identifying circRNA coding ability. CircCode, a Python 3–based framework for identifying the coding ability of circRNAs, is also a simple and powerful command line-based tool. To investigate the translation potential of circRNAs, the user can simply fill in the given configuration file and run the Python 3 scripts. The tool is freely available at <https://github.com/PSSUN/CircCode>.

**Keywords:** bioinformatics, circular RNAs, ribosome profiling data, translation, coding potential, classification

## INTRODUCTION

Circular RNAs (circRNAs) are a special type of noncoding RNA molecule that has become a hot research topic in the field of RNA and is receiving a great deal of attention (Chen and Yang, 2015). Compared with traditional linear RNAs (containing 5' and 3' ends), circRNA molecules usually have a closed circular structure; rendering them more stable and less prone to degradation (Vicens and Westhof, 2014). Although the existence of circRNAs has been known for some time, these molecules were considered to be a by-product of RNA splicing. However, with the development of high-throughput sequencing and bioinformatics technologies, circRNAs have become widely recognized in animals and plants (Chen and Yang, 2015). Recent studies have also shown that a large number of circRNAs can be translated into small peptides in cells (Pamudurti et al., 2017) and have key roles despite their sometimes low level of expression (Hsu and Benfey, 2018; Yang et al., 2018). Although an increasing number of circRNAs are being identified, their functions in plants and animals generally remain to be studied. In addition to their functions as miRNA decoys, circRNAs have important translational potential, but no tools are available for specifically predicting the translational capabilities of these molecules (Jakobi and Dieterich, 2019).



Several tools do exist for the prediction and identification of circRNAs, such as CIRI (Gao et al., 2015), CIRCexplorer (Dong et al., 2019), CircPro (Meng et al., 2017), and circTools (Jakobi et al., 2018). Among them, CircPro can reveal translated circRNAs by calculating a translation potential score for circRNAs based on CPC (Kong et al., 2007), which is a tool for identifying the open reading frame (ORF) in a given sequence. However, because some circRNAs do not use the start codon during translation (Ingolia et al., 2011; Slavoff et al., 2013; Kearse and Wilusz, 2017; Spealman et al., 2018), employing CPC may filter out some truly translated circRNAs. In this study, we used BASiNET (Ito et al., 2018), which is an RNA classifier based on the machine learning methods (random forest and J48 model). It initially transforms the given coding RNAs (positive data) and noncoding RNAs (negative data) and represents them as complex networks; it then extracts the topological measures of these networks and constructs a feature vector to train the model that is used to classify the coding capacity of circRNAs. With this method, erroneous filtering of translated circRNAs that are not initiated by AUG is avoided. Additionally, Ribo-seq technology, which is based on high-throughput sequencing to monitor RPFs (ribosomal protected fragments) of transcripts (Guttman et al., 2013; Brar and Weissman, 2015), can be utilized to determine the locations of circRNAs that are being translated (Michel and Baranov, 2013). To identify the coding ability of circRNAs, we developed the tool CircCode, which involves a Python 3-based framework, and applied CircCode to investigate the translation potential of circRNAs from humans and *Arabidopsis thaliana*. Our work provides a rich resource for further study of the functions of circRNAs with coding capacity.

## METHODS

CircCode was written in the Python 3 programming language; it uses Trimmomatic (Bolger et al., 2014), bowtie (Langmead and Salzberg, 2012), and STAR (Dobin et al., 2013) to filter raw Ribo-seq reads and map these filtered reads to the genome. CircCode then identifies Ribo-seq read-mapped regions in circRNAs that contain junctions. After that, the candidate mapped sequences in the circRNAs are sorted based on classifiers (J48 model) into coding RNAs and noncoding RNAs by BASiNET. Finally, short peptides produced by translation are identified as potential coding regions of circRNAs. The entire process of CircCode consists of five steps (Figure 1).

### Filtering of Ribosomal Profiling Data

First, low-quality fragments and adapters in the Ribo-Seq reads are removed by Trimmomatic with the default parameters to obtain clean Ribo-seq reads. Second, these clean Ribo-seq reads are mapped to an rRNA library to remove reads derived from rRNA using bowtie. Because the read lengths of Ribo-seq are relatively short (generally less than 50 bp), it is possible for one read to match multiple regions. In this case, it is difficult to determine which region a particular read corresponds to. To avoid this, the clean Ribo-seq reads are mapped to the genome of a species of interest, and the reads that are not perfectly aligned to the genome are regarded as the final unique Ribo-seq reads.

## Assembling Virtual Genomes

CircRNAs usually appear as ring-shaped molecules in eukaryotes, and they can be identified based on their back-splicing junctions. However, the sequences of circRNAs in the fasta file are often in linear form. In theory, the result indicates that the junction is between the 5' terminal nucleotide and the 3' terminal nucleotide, although the junction and the sequence near the junction cannot be viewed directly, thus aligning Ribo-seq reads to circRNA sequences, including junctions, in a straightforward manner.

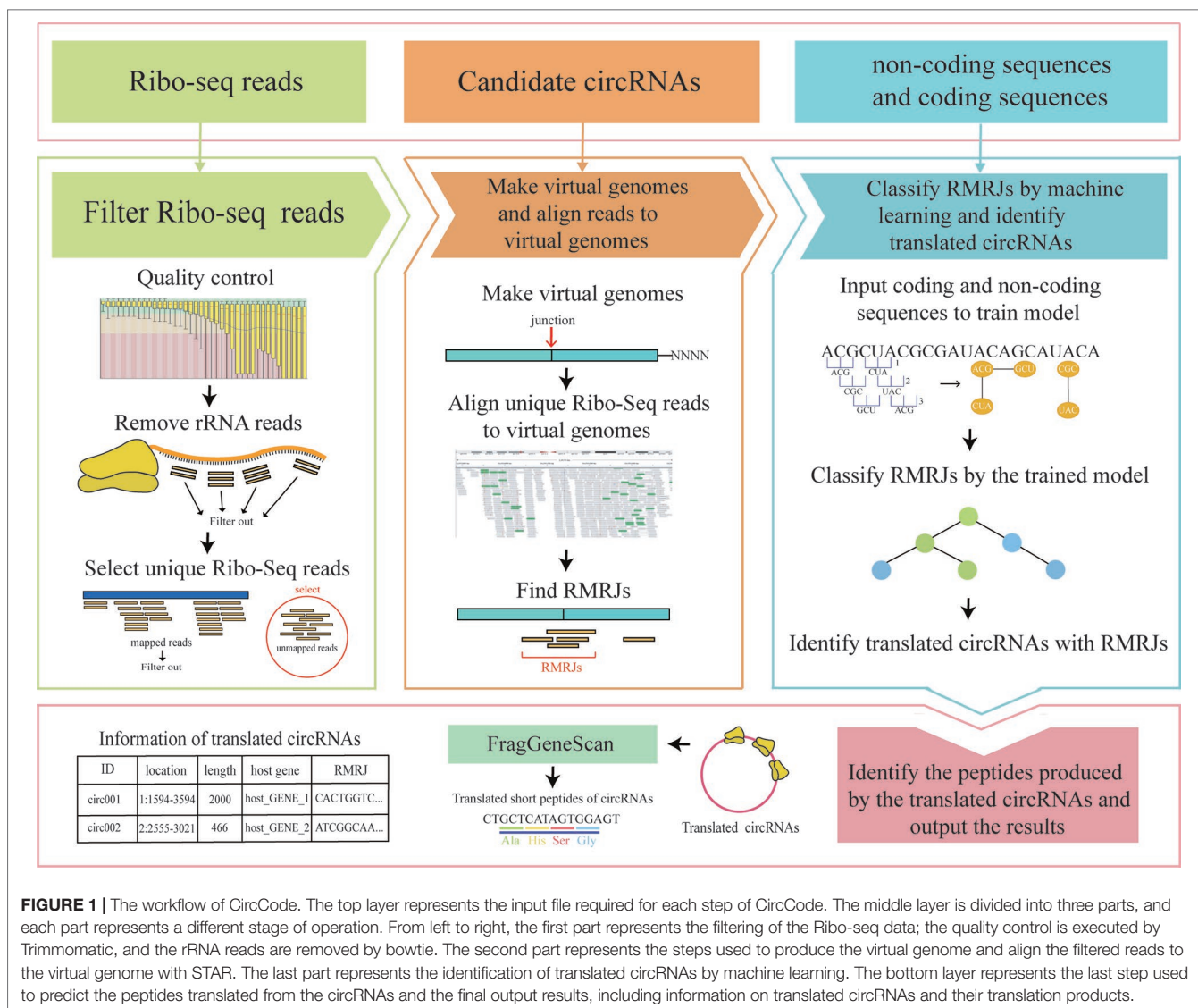
CircCode connects the sequence of each circRNA in tandem such that the junction for each is in the middle of the newly constructed sequence. We also separated each series unit by 100 N nucleotides to avoid confusion at the sequence alignment step (the length of each RPF is less than 50 bp). Finally, we obtained a virtual genome consisting only of candidate circRNAs in tandem separated by 100 Ns. Because CircCode focuses only on alignment between Ribo-seq reads and circRNA sequences, we can investigate the coding potential of circRNAs by mapping the Ribo-seq reads to this virtual genome, which can save a large amount of computational time (the virtual genome is much smaller than the whole genome) and increase the accuracy (by avoiding interference between upstream and downstream sequence comparisons of the circRNAs).

### Determination of the Ribo-seq Read-Mapped Region on a Junction (RMRJ) of circRNAs

The final unique Ribo-seq reads are mapped to a previously created virtual genome using STAR. Because each tandem circRNA unit was separated by 100 N bases before producing the virtual genome, the largest intron length was set to not exceed 10 bases with the parameter “-alignIntronMax 10.” This parameter eliminates any interaction between different circRNAs in the sequence alignment. In the second step of virtual genome production, CircCode stores positional junction information for each circRNA in the virtual genome. If the Ribo-seq read-mapped region in the virtual genome includes the junction of the circRNA, and the number of mapped Ribo-seq reads on junction (NMJ) is greater than 3, the Ribo-seq reads-mapped region on junction of the circRNAs can be regarded as an RMRJ, which reveals a roughly translated segment of circRNAs near the junction site.

### Training of the Model and Classification of RMRJs

Although RMRJs can constitute powerful proof of translation, there are still some shortcomings in this method. Because the length of the reads of the ribosomal map is short, a read may be compared to the wrong position. Therefore, it is not convincing to simply consider the region covered by the Ribo-seq reads as the translated region. To this end, the machine learning method is used to identify the coding ability of the RMRJ. First, CircCode extracts coding RNAs (positive data) and noncoding RNAs (negative data) from a species of interest and uses them for model training by means of the difference in feature vectors between coding and noncoding RNAs.



CircCode then uses the trained model to classify the RMRJs obtained in the previous step by BASiNET. If the RMRJ of a circRNA is recognized as coding RNA, then this circRNA can be identified as a translated circRNA.

## Prediction of Translated Peptides by RMRJs

As expression of circRNAs in organisms is low, Ribo-seq data do not show the exact 3-nt periodicity clearly in the case of fewer RPFs. Therefore, it is difficult to determine the exact translation start site of a translated circRNA. Due to the presence of a stop codon in some RMRJs and because the start codon is difficult to determine, the method of finding an ORF based on a start codon and a stop codon is not feasible.

To determine the true translation regions of these circRNAs and generate the final translation product, FragGeneScan (Rho et al., 2010), which can predict protein-coding regions

in fragmented genes and genes with frameshifts, is used to determine the translated peptides produced by circRNAs.

To avoid the cumbersome running process, all the models can be called by a shell script; the user can simply fill in the given configuration file and input it into script, and the entire process for predicting the translated circRNAs will then be run. In addition, CircCode can be run separately, step by step, such that the user can adjust the parameters in the middle of the procedure and view the results of each step as desired.

## RESULTS AND DISCUSSION

After testing on multiple computers, CircCode was found to run successfully with the required dependencies installed. To test the performance of CircCode, we used data for humans and *A. thaliana* to predict circRNAs with translation potential. The results were compared with circRNAs that have been verified

experimentally as confirmation. Thereafter, we tested the false discovery rate (FDR) value of CircCode further. We used GenRGenS (Ponty et al., 2006) to generate a data set for testing based on known translated circRNAs and confirmed that the FDR value was within an acceptable range and at a low level. Finally, we evaluated the effect of different sequencing depths of Ribo-seq data on CircCode predictions and compared CircCode with other software.

## Translated circRNAs in Humans and *A. thaliana*

To apply the CircCode tool to real data, we first downloaded the files including the human reference genome GRCh38, genome annotation, and human rRNA, from Ensembl. For *A. thaliana*, the reference genomes (TAIR10), genome annotation files, and corresponding rRNA sequences were all downloaded from Ensembl Plants. The Ribo-seq data for humans and *A. thaliana* were downloaded from RPFdb (accession numbers: GSE96643, GSE81295, GSE88794) (Hsu et al., 2016; Willems et al., 2017), and all the candidate circRNAs from human and *A. thaliana* were downloaded from CIRCpedia v2 (Dong et al., 2018) and PlantcircBase, respectively (Chu et al., 2017). Ultimately, we identified 3,610 translated circRNAs from human and 1,569 translated circRNAs from *A. thaliana* using CircCode (Supplementary Data 1).

## Functional Enrichment of Human and *A. thaliana* circRNAs With Coding Potential

Using the CircCode results for human and *A. thaliana*, the online tool KOBAS 3.0 (Wu et al., 2006) was employed to annotate these translated circRNAs based on their parent genes. Furthermore, we performed GO (Gene Ontology) functional analysis and KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment

analysis for these translated circRNAs using the R package clusterProfiler (Yu et al., 2012).

The KEGG results showed that the human circRNAs were enriched in protein processing in the endoplasmic reticulum pathway, carbon metabolism pathway, and RNA transport pathway. GO analysis indicated the participation of human translated circRNAs in the regulation of molecule binding, ATPase activity, and other RNA splicing-related biological processes. In addition, the translated circRNAs of *A. thaliana* are enriched in pathways related to stress resistance, suggesting that they play vital roles in this process (Supplementary Data 2).

## Accuracy Test for CircCode

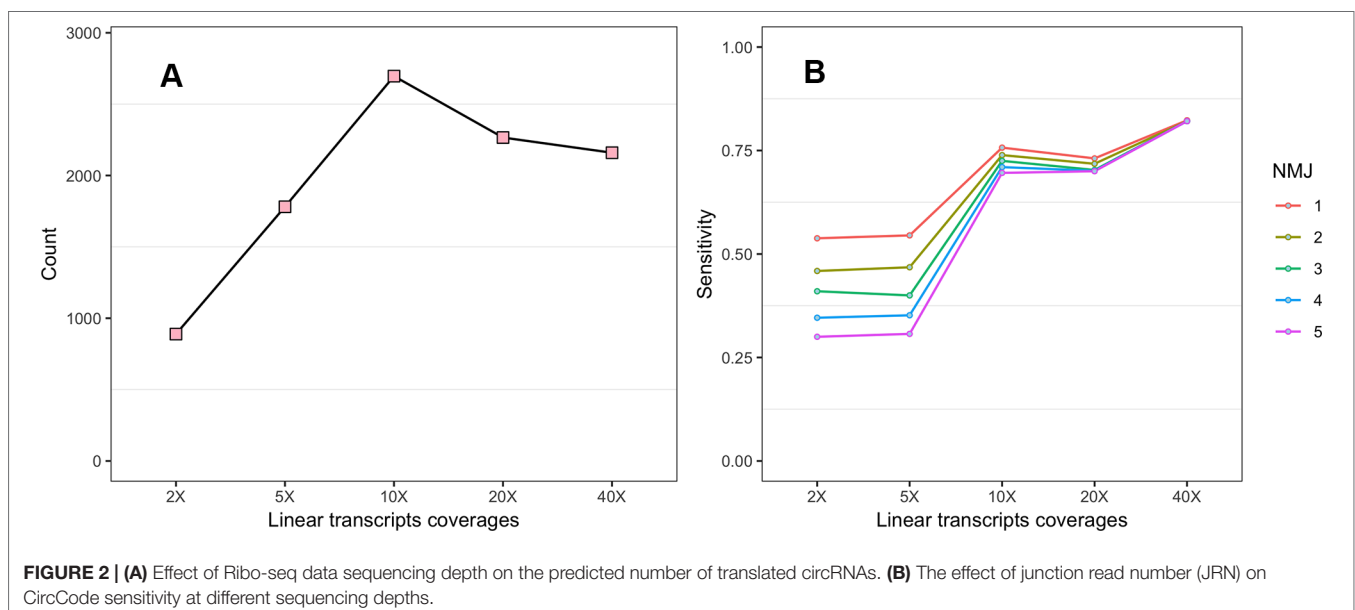
To investigate the accuracy of CircCode, test sequences generated by GenRGenS, which uses the hidden Markov model to produce sequences that have the same sequence characteristics (such as the frequencies of different nucleotides, different codons and different nucleotides at the start of the sequence), were used.

For this study, we used previously published human translated circRNAs (Yang et al., 2017) as the input for GenRGenS and generated 10,000 sequences to test CircCode. We repeated the test 10 times, and on average, 27 translated circRNAs were predicted each time. The FDR value was calculated to be 0.0027, which is much less than 0.05, indicating that the predicted results are credible.

In addition, we compared the translated circRNAs from humans as identified by CircCode with verified polysome-associated circRNA data (Yang et al., 2017). Among them, 60% of the circRNAs were identified by CircCode (Supplementary Data 3).

## Influence of the Ribo-seq Data Sequencing Depth

To investigate the impact of the sequencing depth of Ribo-seq data on the CircCode identification results, we first tested the effect of



sequencing depth on the number of translated circRNAs (**Figure 2A**). When the sequencing depth was low, the predicted number of translated circRNAs was low, and the number of translated circRNAs increased with increasing sequencing depth. The number of translated circRNAs became stable when the sequencing depth reached no less than 10× linear transcript coverage.

Second, the influence of NMJ on sensitivity at different sequencing depths was also assessed (**Figure 2B**). The results showed that NMJ had less impact on sensitivity as the sequencing depth increased. CircCode also had higher sensitivity when using Ribo-seq data with higher sequencing depth.

## Comparison of CircCode With Other Tools

To compare CircCode with other tools, such as CircPro, the same set of Ribo-seq data (SRR3495999) from *A. thaliana* was used to identify translated circRNAs using six processors, with 16 gigabytes of RAM. CircPro identified 44 translated circRNAs in 13 min, whereas CircCode identified 76 translated circRNAs in 20 min. Thus, CircCode is more sensitive than CircPro at the same computer hardware level, but it takes more time. CircPro is concise and less time consuming than CircCode, but CircCode can identify more circRNAs with coding ability than CircPro.

## CONCLUSIONS

CircRNAs play an important role in biology, and it is crucial to accurately identify circRNAs with coding ability for subsequent research. Based on Python 3, we developed CircCode, an easy-to-use command line tool that has high sensitivity for identifying translated circRNAs from Ribo-Seq reads with high accuracy. CircCode exhibits good performance in both plants and animals. Future work will add the downstream character analysis to CircCode by visualizing each step in the process and optimizing the accuracy of the prediction.

## AVAILABILITY AND REQUIREMENTS

CircCode is available at <https://github.com/PSSUN/CircCode>; operating system(s): Linux, programming languages: Python 3 and R; other requirements: bedtools (version 2.20.0 or later), bowtie, STAR, Python 3 packages (Biopython, Pandas, rpy2), R-packages (BASiNET, Biostrings). The installation packages for all of the

required software are available on the CircCode homepage. Users do not need to download them individually. The CircCode home page also provides detailed user manuals for reference. The tool is freely available. There are no restrictions on use by nonacademics.

## DATA AVAILABILITY STATEMENT

All relevant data are within the manuscript and its Supporting Information files.

## AUTHOR CONTRIBUTIONS

Conceptualization: PS, GL. Data Curation: PS, GL. Formal Analysis: PS, GL. Writing – Original Draft: PS, GL. Writing – Review and Editing: PS, GL.

## FUNDING

This work was supported by grants from the National Natural Science Foundation of China (grant nos. 31770333, 31370329, and 11631012), the Program for New Century Excellent Talents in University (NCET-12-0896), and the Fundamental Research Funds for the Central Universities (no. GK201403004). The funding agencies had no role in the study, its design, the data collection and analysis, the decision to publish, or the preparation of the manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.00981/full#supplementary-material>

**SUPPLEMENTARY DATA 1** | The sequence of the predicted translated circRNA and short peptide.

**SUPPLEMENTARY DATA 2** | GO enrichment and KEGG enrichment results for humans and *Arabidopsis thaliana*.

**SUPPLEMENTARY DATA 3** | Comparison of predicted translated circRNAs with validated translated circRNAs.

## REFERENCES

- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170
- Brar, G. A., and Weissman, J. S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* 16, 651–664. doi: 10.1038/nrm4069
- Chen, L.-L., and Yang, L. (2015). Regulation of circRNA biogenesis. *RNA Biol.* 12, 381–388. doi: 10.1080/15476286.2015.1020271
- Chu, Q., Zhang, X., Zhu, X., Liu, C., Mao, L., Ye, C., et al. (2017). PlantCircBase: a database for plant circular RNAs. *Mol. Plant* 10, 1126–1128. doi: 10.1016/j.molp.2017.03.003
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635
- Dong, R., Ma, X.-K., Chen, L.-L., and Yang, L. (2019). “Genome-wide annotation of circRNAs and their alternative back-splicing/splicing with CIRCexplorer Pipeline,” in *Epitranscriptomics*. Eds. N. Wajapeyee and R. Gupta (New York, NY: Springer New York), 137–149. doi: 10.1007/978-1-4939-8808-2\_10
- Dong, R., Ma, X.-K., Li, G.-W., and Yang, L. (2018). CIRCpedia v2: an updated database for comprehensive circular RNA annotation and expression comparison. *Genomics Proteomics Bioinf.* 16, 226–233. doi: 10.1016/j.gpb.2018.08.001
- Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an efficient and unbiased algorithm for *de novo* circular RNA identification. *Genome Biol.* 16, 4. doi: 10.1186/s13059-014-0571-3



- Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., and Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154, 240–251. doi: 10.1016/j.cell.2013.06.009
- Hsu, P. Y., and Benfey, P. N. (2018). Small but mighty: functional peptides encoded by small ORFs in plants. *PROTEOMICS* 18, 1700038. doi: 10.1002/pmic.201700038
- Hsu, P. Y., Calviello, L., Wu, H.-Y. L., Li, F.-W., Rothfels, C. J., Ohler, U., et al. (2016). Super-resolution ribosome profiling reveals unannotated translation events in *Arabidopsis*. *Proc. Natl. Acad. Sci.* 113, E7126–E7135. doi: 10.1073/pnas.1614788113
- Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789–802. doi: 10.1016/j.cell.2011.10.002
- Ito, E. A., Katahira, I., Vicente, F. F., da, R., Pereira, L. F. P., and Lopes, F. M. (2018). BASiNET—Biological Sequences NETWORK: a case study on coding and non-coding RNAs identification. *Nucleic Acids Res.* 46, e96–e96. doi: 10.1093/nar/gky462
- Jakobi, T., and Dieterich, C. (2019). Computational approaches for circular RNA analysis. *Wiley Interdiscip. Rev. RNA* 10 (3), e1528. doi: 10.1002/wrna.1528
- Jakobi, T., Uvarovskii, A., and Dieterich, C. (2018). circTools—a one-stop software solution for circular RNA research. *Bioinformatics* 35 (13), 2326–2328. doi: 10.1093/bioinformatics/bty948
- Kearse, M. G., and Wilusz, J. E. (2017). Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev.* 31, 1717–1731. doi: 10.1101/gad.305250.117
- Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., et al. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, W345–W349. doi: 10.1093/nar/gkm391
- Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923
- Meng, X., Chen, Q., Zhang, P., and Chen, M. (2017). CircPro: an integrated tool for the identification of circRNAs with protein-coding potential. *Bioinformatics* 33, 3314–3316. doi: 10.1093/bioinformatics/btx446
- Michel, A. M., and Baranov, P. V. (2013). Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale: ribosome profiling. *Wiley Interdiscip. Rev. RNA* 4, 473–490. doi: 10.1002/wrna.1172
- Pamudurti, N. R., Bartok, O., Jens, M., Ashwal-Fluss, R., Stottmeister, C., Ruhe, L., et al. (2017). Translation of CircRNAs. *Mol. Cell* 66, 9–21.e7. doi: 10.1016/j.molcel.2017.02.021
- Ponty, Y., Termier, M., and Denise, A. (2006). GenRGenS: software for generating random genomic sequences and structures. *Bioinformatics* 22, 1534–1535. doi: 10.1093/bioinformatics/btl113
- Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38, e191–e191. doi: 10.1093/nar/gkq747
- Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., et al. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* 9, 59–64. doi: 10.1038/nchembio.1120
- Spealman, P., Naik, A. W., May, G. E., Kuersten, S., Freeberg, L., Murphy, R. F., et al. (2018). Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. *Genome Res.* 28, 214–222. doi: 10.1101/gr.221507.117
- Vicens, Q., and Westhof, E. (2014). Biogenesis of circular RNAs. *Cell* 159, 13–14. doi: 10.1016/j.cell.2014.09.005
- Willems, P., Ndah, E., Jonckheere, V., Stael, S., Sticker, A., Martens, L., et al. (2017). N-terminal proteomics assisted profiling of the unexplored translation initiation landscape in *Arabidopsis thaliana*. *Mol. Cell. Proteomics* 16, 1064–1080. doi: 10.1074/mcp.M116.066662
- Wu, J., Mao, X., Cai, T., Luo, J., and Wei, L. (2006). KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.* 34, W720–W724. doi: 10.1093/nar/gkl167
- Yang, L., Fu, J., and Zhou, Y. (2018). Circular RNAs and Their Emerging Roles in Immune Regulation. *Front. Immunol.* 9, 2977. doi: 10.3389/fimmu.2018.02977
- Yang, Y., Fan, X., Mao, M., Song, X., Wu, P., Zhang, Y., et al. (2017). Extensive translation of circular RNAs driven by N6-methyladenosine. *Cell Res.* 27, 626–641. doi: 10.1038/cr.2017.31
- Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS J. Integr. Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Sun and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Single-Cell RNA-Seq Technologies and Related Computational Data Analysis

Geng Chen<sup>1\*</sup>, Baitang Ning<sup>2</sup> and Tielu Shi<sup>1\*</sup>

<sup>1</sup> Center for Bioinformatics and Computational Biology, and Shanghai Key Laboratory of Regulatory Biology, Institute of Biomedical Sciences, School of Life Sciences, East China Normal University, Shanghai, China, <sup>2</sup> National Center for Toxicological Research, United States Food and Drug Administration, Jefferson, AR, United States

## OPEN ACCESS

### Edited by:

Filippo Geraci,  
Italian National Research Council  
(CNR), Italy

### Reviewed by:

Vsevolod Jurievich Makeev,  
Vavilov Institute of General Genetics  
(RAS), Russia  
Iros Barozzi,  
Imperial College London,  
United Kingdom

### \*Correspondence:

Geng Chen  
gchen@bio.ecnu.edu.cn  
Tielu Shi  
tieliushi@yahoo.com

### Specialty section:

This article was submitted to  
Bioinformatics and Computational  
Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 05 December 2018

**Accepted:** 21 March 2019

**Published:** 05 April 2019

### Citation:

Chen G, Ning B and Shi T (2019)  
Single-Cell RNA-Seq Technologies  
and Related Computational Data  
Analysis. *Front. Genet.* 10:317.  
doi: 10.3389/fgene.2019.00317

Single-cell RNA sequencing (scRNA-seq) technologies allow the dissection of gene expression at single-cell resolution, which greatly revolutionizes transcriptomic studies. A number of scRNA-seq protocols have been developed, and these methods possess their unique features with distinct advantages and disadvantages. Due to technical limitations and biological factors, scRNA-seq data are noisier and more complex than bulk RNA-seq data. The high variability of scRNA-seq data raises computational challenges in data analysis. Although an increasing number of bioinformatics methods are proposed for analyzing and interpreting scRNA-seq data, novel algorithms are required to ensure the accuracy and reproducibility of results. In this review, we provide an overview of currently available single-cell isolation protocols and scRNA-seq technologies, and discuss the methods for diverse scRNA-seq data analyses including quality control, read mapping, gene expression quantification, batch effect correction, normalization, imputation, dimensionality reduction, feature selection, cell clustering, trajectory inference, differential expression calling, alternative splicing, allelic expression, and gene regulatory network reconstruction. Further, we outline the prospective development and applications of scRNA-seq technologies.

**Keywords:** single-cell RNA-seq, cell clustering, cell trajectory, alternative splicing, allelic expression

## INTRODUCTION

Bulk RNA-seq technologies have been widely used to study gene expression patterns at population level in the past decade. The advent of single-cell RNA sequencing (scRNA-seq) provides unprecedented opportunities for exploring gene expression profile at the single-cell level. Currently, scRNA-seq has become a favorable choice for studying the key biological questions of cell heterogeneity and the development of early embryos (only include a few number of cells), since bulk RNA-seq mainly reflects the averaged gene expression across thousands of cells. In recent years, scRNA-seq has been applied to various species, especially to diverse human tissues (including normal and cancer), and these studies revealed meaningful cell-to-cell gene expression variability (Jaitin et al., 2014; Grun et al., 2015; Chen et al., 2016a; Cao et al., 2017; Rosenberg et al., 2018). With the innovation of sequencing technologies, some different scRNA-seq protocols have been proposed in the past few years, which largely facilitated the understanding of dynamic gene

expression at single-cell resolution (Kolodziejczyk et al., 2015; Haque et al., 2017; Picelli, 2017; Chen et al., 2018). One of them is the highly efficient strategy LCM-seq (Nichterwitz et al., 2016) which combines laser capture microscopy (LCM) and Smart-seq2 (Picelli et al., 2013) for single-cell transcriptomics without tissue dissociation. Currently available scRNA-seq protocols can be mainly split into two categories based on the captured transcript coverage: (i) full-length transcript sequencing approaches [such as Smart-seq2 (Picelli et al., 2013), MATQ-seq (Sheng et al., 2017) and SUPeR-seq (Fan X. et al., 2015)]; and (ii) 3'-end [e.g., Drop-seq (Macosko et al., 2015), Seq-Well (Gierahn et al., 2017), Chromium (Zheng et al., 2017), and DroNC-seq (Habib et al., 2017)] or 5'-end [such as STRT-seq (Islam et al., 2011, 2012)] transcript sequencing technologies. Each scRNA-seq protocol has its own benefits and drawbacks, resulting in that different scRNA-seq approaches have distinct features and disparate performances (Ziegenhain et al., 2017). In conducting single-cell transcriptomic study, specific scRNA-seq technology may need to be employed in consideration of the balance between research goal and sequencing cost.

Owing to the low amount of starting material, scRNA-seq has limitations of low capture efficiency and high dropouts (Haque et al., 2017). Compared to bulk RNA-seq, scRNA-seq produces noisier and more variable data. The technical noise and biological variation (e.g., stochastic transcription) raise substantial challenges for computational analysis of scRNA-seq data. A variety of tools have been designed to conducting diverse bulk RNA-seq data analyses, but many of those methods cannot be directly applied to scRNA-seq data (Stegle et al., 2015). Except short-read mapping, almost all data analyses (such as differential expression, cell clustering, and gene regulatory network inference) have certain disparities in methods between scRNA-seq and bulk RNA-seq. Due to the high technical noise, quality control (QC) is crucial for identifying and removing the low-quality scRNA-seq data to get robust and reproducible results. Furthermore, some analyses including alternative splicing (AS) detection, allelic expression exploration and RNA-editing identification are not suitable for the 3' or 5'-tag sequencing protocols of scRNA-seq, but these analyses could be applicable to the data generated by whole-transcript scRNA-seq. On the other hand, an increasing number of tools are specially proposed for analyzing scRNA-seq data, and each method has its own pros and cons (Stegle et al., 2015; Bacher and Kendzierski, 2016). Therefore, to effectively handle the high variability of scRNA-seq data, attention should be paid to choosing appropriately analytical approaches.

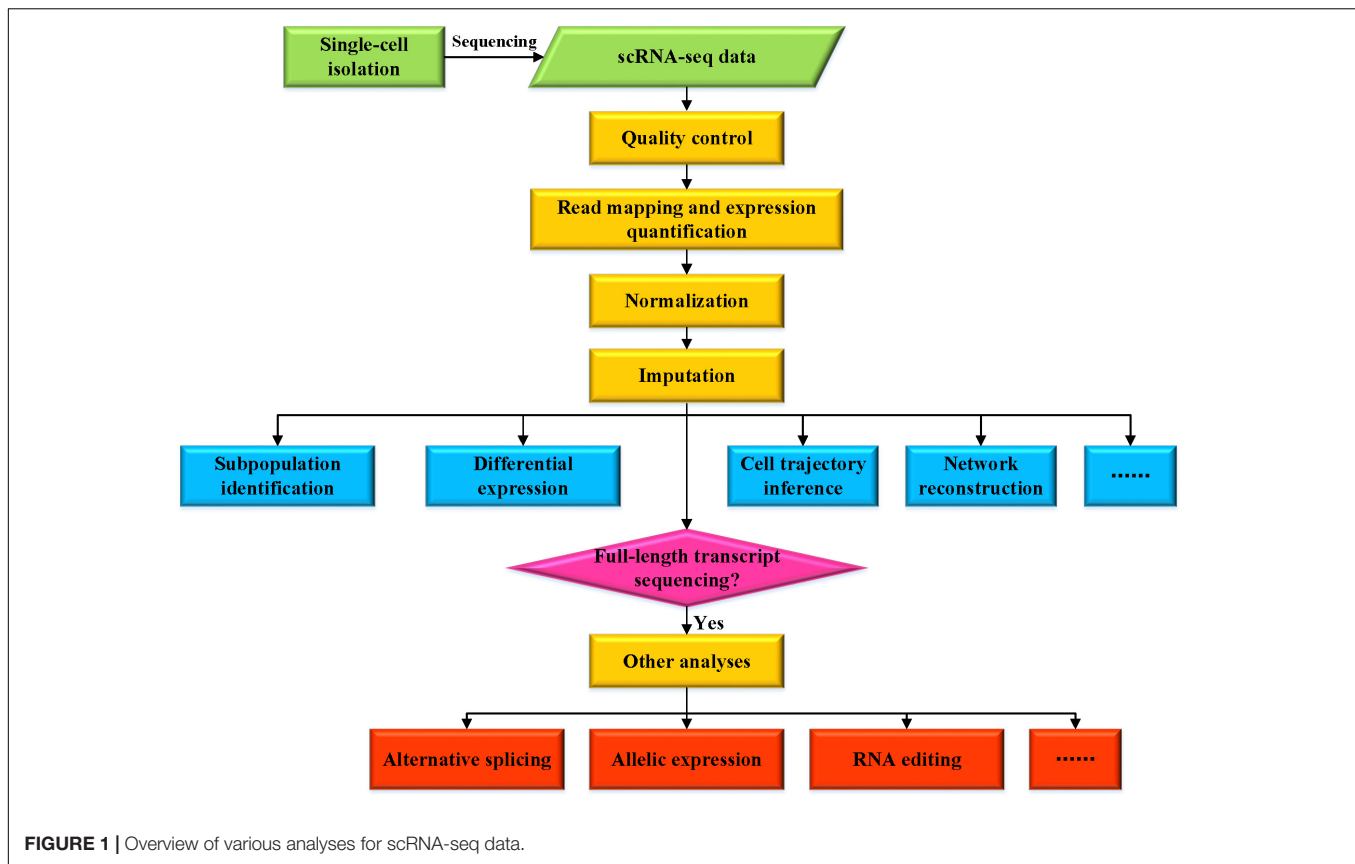
This Review aims to summarize and discuss currently available scRNA-seq technologies and various data analysis methods. We first introduce distinct single-cell isolation protocols and various scRNA-seq technologies developed in recent years. Then we focus on the analyses of scRNA-seq data and highlight the analytical differences between bulk RNA-seq and scRNA-seq data. Considering the high technical noise and complexity of scRNA-seq data, we also provide recommendations on the selection of suitable tools to analyze scRNA-seq data and ensure the reproducibility of results.

## ISOLATION OF SINGLE CELLS

The first step of scRNA-seq is isolation of individual cells (**Figure 1**), although the capture efficiency is a big challenge for scRNA-seq. Currently, several different approaches are available for isolating single cells, including limiting dilution, micromanipulation, flow-activated cell sorting (FACS), laser capture microdissection (LCM), and microfluidics (Gross et al., 2015; Kolodziejczyk et al., 2015; Hwang et al., 2018). Limiting dilution technique uses pipettes to isolate cells by dilution, the main limitation of this method is inefficient. Micromanipulation is a classical approach used to retrieve cells from samples with a small number of cells, such as early embryos or uncultivated microorganisms, while this technique is time-consuming and low throughput. FACS has been widely used for isolating single cells, which requires large starting volumes (>10,000 cells) in suspension. LCM is an advanced strategy used for isolating individual cells from solid tissues by using a laser system aided by computer. Microfluidics is increasingly popular due to its property of low sample consumption, precise fluid control and low analysis cost. These single-cell isolation protocols have their own advantages and show distinct performances in terms of capture efficiency and purity of the target cells (Gross et al., 2015; Hu et al., 2016).

## CURRENTLY AVAILABLE SCRNA-SEQ TECHNOLOGIES

To date, a number of scRNA-seq technologies have been proposed for single-cell transcriptomic studies (**Table 1**). The first scRNA-seq method was published by Tang et al. (2009), and then many other scRNA-seq approaches were subsequently developed. Those scRNA-seq technologies differ in at least one of the following aspects: (i) cell isolation; (ii) cell lysis; (iii) reverse transcription; (iv) amplification; (v) transcript coverage; (vi) strand specificity; and (vii) UMI (unique molecular identifiers, molecular tags that can be applied to detect and quantify the unique transcripts) availability. One conspicuous difference among these scRNA-seq methods is that some of them can produce full-length (or nearly full-length) transcript sequencing data (e.g., Smart-seq2, SUPeR-seq, and MATQ-seq), whereas others only capture and sequence the 3'-end [such as Drop-seq, Seq-Well and DroNC-seq, SPLiT-seq (Rosenberg et al., 2018)] or 5'-end (e.g., STRT-seq) of the transcripts (**Table 1**). Distinct scRNA-seq protocols may possess disparate strengths and weaknesses, and several published reviews have compared a portion of them in detail (Kolodziejczyk et al., 2015; Haque et al., 2017; Picelli, 2017; Ziegenhain et al., 2017). A previous study demonstrated that Smart-seq2 can detect a bigger number of expressed genes than other scRNA-seq technologies including CEL-seq2 (Hashimshony et al., 2016), MARS-seq (Jaitin et al., 2014), Smart-seq (Ramskold et al., 2012), and Drop-seq protocols (Ziegenhain et al., 2017). Recently, Sheng et al. (2017) showed that another full-length transcript sequencing approach MATQ-seq could outperform Smart-seq2 in detecting low-abundance genes.



Compared to 3'-end or 5'-end counting protocols, full-length scRNA-seq methods have incomparable advantages in isoform usage analysis, allelic expression detection, and RNA editing identification owing to their superiority of transcript coverage. Moreover, for detecting certain lowly expressed genes/transcripts, full-length scRNA-seq approaches could be better than 3' sequencing methods (Ziegenhain et al., 2017). Notably, droplet-based technologies [e.g., Drop-seq (Macosko et al., 2015), InDrop (Klein et al., 2015), and Chromium (Zheng et al., 2017)] can generally provide a larger throughput of cells and a lower sequencing cost per cell compared to whole-transcript scRNA-seq. Thus, droplet-based protocols are suitable for generating huge amounts of cells to identify the cell subpopulations of complex tissues or tumor samples.

Strikingly, several scRNA-seq technologies can capture both polyA+ and polyA- RNAs, such as SUPeR-seq (Fan X. et al., 2015) and MATQ-seq (Sheng et al., 2017). These protocols are extremely useful for sequencing long noncoding RNAs (lncRNAs) and circular RNAs (circRNAs). Lots of studies have demonstrated that lncRNAs and circRNAs play important roles in diverse biological processes of cells and may serve as crucial biomarkers for cancers (Barrett and Salzman, 2016; Chen et al., 2016b; Quinn and Chang, 2016; Kristensen et al., 2018); therefore, such scRNA-seq methods can provide unprecedented opportunities to comprehensively explore the expression dynamics of both protein-coding and noncoding RNAs at the single-cell level.

Compared to traditional bulk RNA-seq technologies, scRNA-seq protocols suffer higher technical variations. In order to estimate the technical variances among different cells, spike-ins [such as External RNA Control Consortium (ERCC) controls (External, 2005)] and UMIs have been widely used in corresponding scRNA-seq methods. The RNA spike-ins are RNA transcripts (with known sequences and quantity) that are applied to calibrate the measurements of RNA hybridization assays, such as RNA-Seq, and UMIs can theoretically enable the estimation of absolute molecular counts. It is worth noting that ERCC and UMIs are not applicable to all scRNA-seq technologies due to the inherent protocol differences. Spike-ins are used in approaches like Smart-seq2 and SUPeR-seq but are not compatible with droplet-based methods, whereas UMIs are typically applied to 3'-end sequencing technologies [such as Drop-seq (Macosko et al., 2015), InDrop (Klein et al., 2015), and MARS-seq (Jaitin et al., 2014)]. Consequently, users can select the suitable scRNA-seq method according to the technical properties and advantages, number of cells to be sequenced and cost considerations.

## READ ALIGNMENT AND EXPRESSION QUANTIFICATION OF SCRNA-SEQ DATA

The mapping ratio of reads is an important indicator of the overall quality of scRNA-seq data. Since both scRNA-seq and

**TABLE 1** | Summary of widely used scRNA-seq technologies.

Methods	Transcript coverage	UMI possibility	Strand specific	References
Tang method	Nearly full-length	No	No	Tang et al., 2009
Quartz-Seq	Full-length	No	No	Sasagawa et al., 2013
SUPeR-seq	Full-length	No	No	Fan X. et al., 2015
Smart-seq	Full-length	No	No	Ramskold et al., 2012
Smart-seq2	Full-length	No	No	Picelli et al., 2013
MATQ-seq	Full-length	Yes	Yes	Sheng et al., 2017
STRT-seq and STRT/C1	5'-only	Yes	Yes	Islam et al., 2011, 2012
CEL-seq	3'-only	Yes	Yes	Hashimshony et al., 2012
CEL-seq2	3'-only	Yes	Yes	Hashimshony et al., 2016
MARS-seq	3'-only	Yes	Yes	Jaitin et al., 2014
CytoSeq	3'-only	Yes	Yes	Fan H.C. et al., 2015
Drop-seq	3'-only	Yes	Yes	Macosko et al., 2015
InDrop	3'-only	Yes	Yes	Klein et al., 2015
Chromium	3'-only	Yes	Yes	Zheng et al., 2017
SPLIT-seq	3'-only	Yes	Yes	Rosenberg et al., 2018
sci-RNA-seq	3'-only	Yes	Yes	Cao et al., 2017
Seq-Well	3'-only	Yes	Yes	Gierahn et al., 2017
DroNC-seq	3'-only	Yes	Yes	Habib et al., 2017
Quartz-Seq2	3'-only	Yes	Yes	Sasagawa et al., 2018

bulk RNA-seq technologies generally sequence transcripts into reads to generate the raw data in fastq format, no differences exist between these two types of RNA-seq data in read alignment. The mapping tools originally developed for bulk RNA-seq are also applicable to scRNA-seq data. Numerous spliced alignment programs have been designed for mapping RNA-seq data, which was extensively discussed previously (Li and Homer, 2010; Chen et al., 2011). Generally, the read mapping algorithms mainly fall into two categories: spaced-seed indexing based and Burrows-Wheeler transform (BWT) based (Li and Homer, 2010). Currently popular aligners like TopHat2 (Kim et al., 2013), STAR (Dobin and Gingeras, 2015), and HISAT (Kim et al., 2015) perform well in mapping speed and accuracy, and they can efficiently map billions of reads to the reference genome or transcriptome (Table 2). STAR is a suffix-array based method and is faster than TopHat2, but it requires a huge memory size (28 gigabytes for human genome) for read mapping (Dobin and Gingeras, 2015). Engstrom et al. systematically evaluated 26 read alignment protocols (did not include HISAT) and found that different mapping tools exhibit distinct strengths and weakness, where some programs are with a faster mapping speed but a lower accuracy in splice junction detection (Engstrom et al., 2013). HISAT is developed based on BWT and Ferragina-Manzini (FM) methods. Kim et al. (2015) showed that HISAT is currently the fastest tool that can achieve equal or better accuracy than other available aligners.

For gene/transcript expression quantification, distinct approaches are needed, based on the range of transcript sequence captured by scRNA-seq. The data generated by whole-transcript scRNA-seq methods (such as Smart-seq2 and MATQ-seq) can

**TABLE 2** | Tools for read mapping and expression quantification of scRNA-seq data.

Tools	Category	URL	References
TopHat2	Read mapping	<a href="https://ccb.jhu.edu/software/tophat/index.shtml">https://ccb.jhu.edu/software/tophat/index.shtml</a>	Kim et al., 2013
STAR	Read mapping	<a href="https://github.com/alexdobin/STAR">https://github.com/alexdobin/STAR</a>	Dobin and Gingeras, 2015
HISAT2	Read mapping	<a href="https://ccb.jhu.edu/software/hisat2/index.shtml">https://ccb.jhu.edu/software/hisat2/index.shtml</a>	Kim et al., 2015
Cufflinks	Expression quantification	<a href="https://github.com/cole-trapnell-lab/cufflinks">https://github.com/cole-trapnell-lab/cufflinks</a>	Trapnell et al., 2010
RSEM	Expression quantification	<a href="https://github.com/deweylab/RSEM">https://github.com/deweylab/RSEM</a>	Li and Dewey, 2011
StringTie	Expression quantification	<a href="https://github.com/gpertea/stringtie">https://github.com/gpertea/stringtie</a>	Pertea et al., 2015

be analyzed with the software developed for bulk RNA-seq to quantify gene/transcript expression. Two main approaches are available for transcriptome reconstruction: *de novo* assembly (does not need a reference genome) and reference-based or genome-guided assembly (Chen et al., 2017b). *De novo* transcriptome assembly methods are primarily applied to the organisms that lack a reference genome, and are generally with a lower accuracy than that of genome-guided assembly (Garber et al., 2011). The popular genome-guided assembly tools including Cufflinks (Trapnell et al., 2010), RSEM (Li and Dewey, 2011), and StringTie (Pertea et al., 2015) have been broadly used in many scRNA-seq studies to get relative gene/transcript expression estimation in reads or fragments per kilobase per million mapped reads (RPKM or FPKM) or transcripts per million mapped reads (TPM) (Table 2). Pertea et al. (2015) stated that StringTie outperforms other genome-guided approaches in gene/transcript reconstruction and expression quantification. On the other hand, for the 3'-end scRNA-seq protocols (e.g., CEL-seq2, MARS-seq, Drop-seq, and InDrop), specific algorithms are required to calculate gene/transcript expression based on UMIs. SAVER (single-cell analysis via expression recovery) is an efficient UMI-based tool recently proposed for accurately estimating gene expression of single cells (Huang et al., 2018). In theory, UMI-based scRNA-seq can largely reduce the technical noise, which remarkably benefits the estimation of absolute transcript counts (Islam et al., 2014).

## QUALITY CONTROL OF SCRNA-SEQ DATA

The limitations in scRNA-seq including bias of transcript coverage, low capture efficiency, and sequencing coverage result in that scRNA-seq data are with a higher level of technical noise than bulk RNA-seq data (Kolodziejczyk et al., 2015). Even for the most sensitive scRNA-seq protocol, it is a frequent phenomenon that some specific transcripts cannot be detected (termed dropout events) (Haque et al., 2017). Generally, scRNA-seq experiments



can generate a portion of low-quality data from the cells that are broken or dead or mixed with multiple cells (Ilicic et al., 2016). These low-quality cells will hinder the downstream analysis and may lead to misinterpretation of the data. Accordingly, QC of scRNA-seq data is crucial to identify and remove the low-quality cells.

To exclude the low-quality cells from scRNA-seq, close attention should be paid to avoid multi-cells or dead cells in the cell capture step. After sequencing, a series of QC analyses are required to eliminate the data from low-quality cells. Those samples contain only a few number of reads should be discarded first since insufficient sequencing depth may lead to the loss of a large portion of lowly and moderately expressed genes. Then tools initially developed for QC of bulk RNA-seq data, such as FastQC<sup>1</sup>, can be employed to check the sequencing quality of scRNA-seq data. Moreover, after read alignment, samples with very low mapping ratio should be eliminated because they contain massively unmappable reads that might be resulted from RNA degradation. If extrinsic spike-ins (such ERCC) were used in scRNA-seq, technical noise could be estimated. The cells with an extremely high portion of reads mapped to the spike-ins indicate that they were probably broken during cell capture process and should be removed (Ilicic et al., 2016). Cytoplasmic RNAs are usually lost but mitochondrial RNAs are retained for broken cells, thus the ratio of reads mapped to mitochondrial genome is also informative for identifying low-quality cells (Bacher and Kendzierski, 2016). Additionally, the number of expressed genes/transcripts can be detected in each cell is also suggestive. If only a small number of genes can be detected in a cell, this cell is probably damaged or dead or suffered from RNA degradation. Considering the high noise of scRNA-seq data, a threshold of 1 FPKM/RPKM was usually applied to define the expressed genes. Some QC methods for scRNA-seq have been proposed (Stegle et al., 2015; Ilicic et al., 2016), including SinQC (Jiang et al., 2016) and Scater (McCarthy et al., 2017), these tools are useful for QC of scRNA-seq data.

## BATCH EFFECT CORRECTION

Batch effect is a common source of technical variation in high-throughput sequencing experiments. The innovation and decreasing cost of scRNA-seq enable many studies to profile the transcriptomes of a huge amount of cells. The large scale scRNA-seq data sets might be separately generated with distinct operators at different times, and could also be produced in multiple laboratories using disparate cell dissociation protocols, library preparation approaches and/or sequencing platforms. These factors would introduce systematic error and confound the technical and biological variability, leading to that the gene expression profile in one batch systematically differs from that in another (Leek et al., 2010; Hicks et al., 2018). Therefore, batch effect is a major challenge in scRNA-seq data analysis, which may mask the underlying biology and cause spurious results. To avoid incorrect data integration and interpretation, batch effects must

be corrected before the downstream analysis. Because of the data feature differences between scRNA-seq and bulk RNA-seq, batch-correction approaches specially proposed for bulk RNA-seq [e.g., RUVseq (Risso et al., 2014) and svaseq (Leek, 2014)] may not be suitable for scRNA-seq. Several methods have been recently designed to mitigate the batch effects in scRNA-seq data, such as MNN (mutual nearest neighbor) (Haghverdi et al., 2018) and kBET (k-nearest neighbor batch effect test) (Buttner et al., 2019). MNN corrects the batch effects using the data from the most similar cells in different batches. kBET is a  $\chi^2$ -based method for quantifying batch effects in scRNA-seq data. These specific batch-correction approaches for scRNA-seq data can perform better than the methods developed for bulk RNA-seq (Haghverdi et al., 2018; Buttner et al., 2019).

## NORMALIZATION OF SCRNA-SEQ DATA

To correctly interpret the results from scRNA-seq data, normalization is an essential step to get the signal of interest by adjusting unwanted biases resulted from capture efficiency, sequencing depth, dropouts, and other technical effects. Technical noise of scRNA-seq is an obvious problem due to the low starting material and challenging experimental protocols. Normalization of scRNA-seq data will benefit the downstream analyses including cell subpopulation identification and differential expression calling. In general, normalization can be divided into two different types: within-sample normalization and between-sample normalization (Vallejos et al., 2017). Within-sample normalization aims to remove the gene-specific biases (e.g., GC content and gene length), which makes gene expression comparable within one sample (such as RPKM/FPKM and TPM). In contrast, between-sample normalization is to adjust sample-specific differences (e.g., sequencing depth and capture efficiency) to enable the comparison of gene expression between samples. Generally, those simple normalization strategies are based on sequencing depth or upper quartile. If spike-ins or UMIs are used in scRNA-seq protocol, normalization can be refined based on the performance of spike-ins/UMIs (Bacher and Kendzierski, 2016).

A number of approaches have been developed for between-sample normalization of bulk RNA-seq data, such as DESeq2 (Love et al., 2014) and trimmed mean of M values (TMM) (Robinson and Oshlack, 2010). DESeq2 calculates scaling factor based on the read counts across different samples, while TMM removes the extreme log fold changes (Vallejos et al., 2017). However, bulk-based normalization approaches may be not suitable for the data of single-cell transcriptomics. Because scRNA-seq generates abundant zero-expression values and has a higher level of technical variation than bulk RNA-seq, using bulk RNA-seq normalization approaches may cause overcorrection in scRNA-seq for lowly expressed genes (Vallejos et al., 2017). Several normalization methods have been proposed for scRNA-seq data, such as SCnorm (Bacher et al., 2017), SAMstr (Katayama et al., 2013) and a recently introduced deconvolution approach that uses the summed expression values across pools of cells to conduct normalization (Lun et al., 2016). SCnorm is based on quantile regression, while

<sup>1</sup><https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>



SAMstrr relies on spike-ins. Bacher et al. (2017) believed that traditional normalization methods developed for bulk RNA-seq may introduce artifacts for normalizing scRNA-seq data, while SCnorm can effectively normalize scRNA-seq data and improve principal component analysis (PCA) and the identification of differentially expressed genes.

## IMPUTATION OF SCRNA-SEQ DATA

Single-cell RNA sequencing data generally contain many missing values or dropouts that were caused by failed amplification of the original RNAs. The frequency of dropout events for scRNA-seq is protocol-dependent, and is closely associated with the number of sequencing reads generated for each cell (Svensson et al., 2017). The dropout events increase the cell-to-cell variability, leading to signal influence on every gene, and obscuration of gene-gene relationship detection. Therefore, dropouts can largely affect the downstream analyses since a significant portion of truly expressed transcripts may not be detectable in scRNA-seq. Imputation is a useful strategy to replace the missing data (dropouts) with substituted values. Although some methods have been proposed for imputation of bulk RNA-seq data, they are not directly applicable to scRNA-seq data (Zhang and Zhang, 2018). Several imputation methods have been recently developed for scRNA-seq, including SAVER (Huang et al., 2018), MAGIC (van Dijk et al., 2018), ScImpute (Li and Li, 2018), DrImpute (Gong et al., 2018), and AutoImpute (Talwar et al., 2018). SAVER is a Bayesian-based model designed for UMI-based scRNA-seq data to recover the true expression level of all genes. MAGIC imputes the gene expression by building Markov affinity-based graph. The developers of ScImpute suggested that SAVER and MAGIC may lead to expression changes of the genes unaffected by dropouts, while ScImpute can impute the dropout values without introducing new biases through using the information from the same genes unlikely affected by dropouts in other similar cells. DrImpute is a clustering-based approach and can effectively separate the dropout zeros from true zeros. AutoImpute is an autoencoder-based method that learns the inherent distribution of scRNA-seq data to impute the missing values. Recently, Zhang et al. evaluated different imputation methods and found that the performances of these approaches are correlated with their model hypothesis and scalability (Zhang and Zhang, 2018).

## DIMENSIONALITY REDUCTION AND FEATURE SELECTION

Single-cell RNA sequencing data are with a high dimensionality, which may involve thousands of genes and a large number of cells. Dimensionality reduction and feature selection are two main strategies for dealing with high dimensional data (Andrews and Hemberg, 2018a). Dimensionality reduction methods generally project the data into a lower dimensional space by optimally preserving some key properties of the original data. PCA is a linear dimensional reduction algorithm, which assumes

that the data is approximately normally distributed. T-distributed stochastic neighbor embedding (t-SNE) is a non-linear approach mainly designed for visualizing high dimensional data (van der Maaten and Hinton, 2008). Both PCA and t-SNE have been broadly used in diverse scRNA-seq studies to reduce the data dimension and visualize the cells discriminated into distinct subpopulations (Chen et al., 2016a; Rosenberg et al., 2018). It is worth noting that PCA cannot effectively represent the complex structure of scRNA-seq data and t-SNE has limitations of slow computation time and different embeddings for processing the same dataset multiple times. Recently, UMAP (uniform manifold approximation and projection) (Becht et al., 2018), and scvis (Ding et al., 2018) were specially developed for reducing the dimensions of scRNA-seq data. Becht et al. showed that UMAP provides the fastest run times, the highest reproducibility and the most meaningful organization of cell clusters than other dimensionality reduction approaches (Becht et al., 2018).

Feature selection removes the uninformative genes and identifies the most relevant features to reduce the number of dimensions used in downstream analysis. Reducing the number of genes by performing feature selection can largely speed up the calculations of large-scale scRNA-seq data (Andrews and Hemberg, 2018b). Differential expression is a widely used method for feature selection in bulk RNA-seq experiments, but it is hard to apply to scRNA-seq data since the information of predetermined and/or homogeneous subpopulations needed for differential expression calling of scRNA-seq data [e.g., SCDE (Kharchenko et al., 2014)] is often unavailable. Unsupervised feature selection algorithms specially designed for scRNA-seq data can be divided into the following groups: (i) highly variable genes (HVG) based; (ii) spike-in based; and (iii) dropout-based (Andrews and Hemberg, 2018a). HVG methods rely on the assumption that the genes with highly variable expression across cells are resulted from biological effects rather than technical noise. The HVG approaches include algorithms proposed by Brennecke et al. (2013), and FindVariableGenes (FVG) implemented in Seurat (Satija et al., 2015). Spike-in based approaches identify the genes showing significant higher variance than those of spike-ins with similar expression levels [e.g., scLVM (Buettner et al., 2015) and BASiCS (Vallejos et al., 2015)], which shares similar idea of HVG. Dropout based methods take advantage of the dropout distribution of scRNA-seq data to perform feature selection, like M3Drop (Andrews and Hemberg, 2018b). Andrews and Hemberg showed that their M3Drop tool outperforms existing variance-based feature selection approaches.

## CELL SUBPOPULATION IDENTIFICATION

A key goal of scRNA-seq data analysis is to identify cell subpopulations (different populations are often distinct cell types) within a certain condition or tissue to unravel the heterogeneity of cells. Notably, cell subpopulation identification should be carried out after QC and normalization of scRNA-seq data, otherwise artifacts could be introduced. Approaches for

clustering cells can be mainly grouped into two categories based on whether prior information is used. If a set of known markers was used in clustering, the methods are prior information based. Alternatively, unsupervised clustering methods can be used for *de novo* identification of cell populations with scRNA-seq data. The algorithms for unsupervised clustering can be primarily divided into the following types: (i) k-means; (ii) hierarchical clustering; (iii) density-based clustering; and (iv) graph-based clustering (Andrews and Hemberg, 2018a). K-means is a fast approach that assigns cells to the nearest cluster center, and it requires the predetermined number of clusters. Hierarchical clustering can determine the relationships between clusters, but it generally works slower than k-means. Density-based clustering methods need a large number of samples to accurately calculate densities and usually assume that all clusters have equal density. Graph-based clustering can be considered as an extension of density-based clustering, and it can be applied to millions of cells. Some clustering methods have been specially designed for scRNA-seq data, such as single-cell consensus clustering (SC3) (Kiselev et al., 2017) and the clustering approach implemented in Seurat (Satija et al., 2015), which can facilitate the identification of cell subpopulations (Table 3). SC3 is an unsupervised approach that combines multiple clustering approaches, which has a high accuracy and robustness in single-cell clustering. Seurat identifies the cell clusters mainly based on a shared nearest neighbor (SNN) clustering algorithm. Once the subpopulations are determined, the markers that can best discriminate distinct subpopulations are usually identified through differential expression calling or analysis of variance (ANOVA).

## DIFFERENTIAL EXPRESSION ANALYSIS OF SCRNA-SEQ DATA

Differential expression analysis is very useful to find the significantly differentially expressed genes (DEGs) between distinct subpopulations or groups of cells. The DEGs are crucial for interpreting the biological difference between two compared

conditions. The technical variability, high noise (e.g., dropouts) and massive sample size of scRNA-seq data raise challenges in differential expression calling (McDavid et al., 2013). Moreover, multiple possible cell states can exist within a population of cells, leading to the multimodality of gene expression in cells (Vallejos et al., 2016). The tools originally developed for bulk RNA-seq data have been used in many single-cell studies to identify the DEGs, but the applicability of these methods for scRNA-seq data is still unclear. In recent years, some specific methods have been proposed for conducting differential expression calling based on scRNA-seq data, such as MAST (Finak et al., 2015), SCDE (Kharchenko et al., 2014), DEsingle (Miao et al., 2018), Census (Qiu et al., 2017), and BCseq (Chen and Zheng, 2018) (Table 4). MAST is based on linear model fitting and likelihood ratio testing. SCDE is a Bayesian approach using a low-magnitude Poisson process to account for dropouts. DEsingle employs Zero-Inflated Negative Binomial model to estimate the dropouts and real zeros. BCseq mitigates the technical noise in a data-adaptive manner. Soneson and Robinson recently assessed 36 differential expression methods (including the tools designed for scRNA-seq and bulk RNA-seq data) and revealed significant differences among these approaches in the characteristics and number of DEGs (Soneson and Robinson, 2018). An increasing number of tools for differential expression analysis of scRNA-seq data will be developed, and users are encouraged to choose the tools specially

**TABLE 4 |** Differential expression analysis tools for RNA-seq data.

Methods	Category	URL	References
ROTS	Single cell	<a href="https://bioconductor.org/packages/release/bioc/html/ROTS.html">https://bioconductor.org/packages/release/bioc/html/ROTS.html</a>	Seyednasrollah et al., 2016
MAST	Single cell	<a href="https://github.com/RGLab/MAST">https://github.com/RGLab/MAST</a>	Finak et al., 2015
BCseq	Single cell	<a href="https://bioconductor.org/packages/devel/bioc/html/bcSeq.html">https://bioconductor.org/packages/devel/bioc/html/bcSeq.html</a>	Chen and Zheng, 2018
SCDE	Single cell	<a href="http://hms-dbmi.github.io/scde/">http://hms-dbmi.github.io/scde/</a>	Kharchenko et al., 2014
DEsingle	Single cell	<a href="https://bioconductor.org/packages/DEsingle">https://bioconductor.org/packages/DEsingle</a>	Miao et al., 2018
Census	Single cell	<a href="http://cole-trapnell-lab.github.io/monocle-release/">http://cole-trapnell-lab.github.io/monocle-release/</a>	Qiu et al., 2017
D3E	Single cell	<a href="https://github.com/hemberg-lab/D3E">https://github.com/hemberg-lab/D3E</a>	Delmans and Hemberg, 2016
BPSC	Single cell	<a href="https://github.com/nghiavtr/BPSC">https://github.com/nghiavtr/BPSC</a>	Vu et al., 2016
DESeq2	Bulk	<a href="https://bioconductor.org/packages/release/bioc/html/DESeq2.html">https://bioconductor.org/packages/release/bioc/html/DESeq2.html</a>	Love et al., 2014
edgeR	Bulk	<a href="https://bioconductor.org/packages/release/bioc/html/edgeR.html">https://bioconductor.org/packages/release/bioc/html/edgeR.html</a>	Robinson et al., 2010
Limma	Bulk	<a href="https://bioconductor.org/packages/release/bioc/html/limma.html">https://bioconductor.org/packages/release/bioc/html/limma.html</a>	Ritchie et al., 2015
Ballgown	Bulk	<a href="http://www.bioconductor.org/packages/release/bioc/html/ballgown.html">http://www.bioconductor.org/packages/release/bioc/html/ballgown.html</a>	Frazee et al., 2015

**TABLE 3 |** Subpopulation identification methods for scRNA-seq data.

Methods	URL	References
SC3	<a href="http://bioconductor.org/packages/SC3">http://bioconductor.org/packages/SC3</a>	Kiselev et al., 2017
ZIFA	<a href="https://github.com/epierson9/ZIFA">https://github.com/epierson9/ZIFA</a>	Pierson and Yau, 2015
Destiny	<a href="https://github.com/theislab/destiny">https://github.com/theislab/destiny</a>	Angerer et al., 2016
SNN-Cliq	<a href="http://bioinfo.uncc.edu/SNNCliq/">http://bioinfo.uncc.edu/SNNCliq/</a>	Xu and Su, 2015
RaceID	<a href="https://github.com/dgrun/RaceID">https://github.com/dgrun/RaceID</a>	Grun et al., 2015
SCUBA	<a href="https://github.com/gcyuan/SCUBA">https://github.com/gcyuan/SCUBA</a>	Marco et al., 2014
BackSPIN	<a href="https://github.com/linnarsson-lab/BackSPIN">https://github.com/linnarsson-lab/BackSPIN</a>	Zeisel et al., 2015
PAGODA	<a href="http://hms-dbmi.github.io/scde/">http://hms-dbmi.github.io/scde/</a>	Fan et al., 2016
CIDR	<a href="https://github.com/VCCRI/CIDR">https://github.com/VCCRI/CIDR</a>	Lin et al., 2017
pcaReduce	<a href="https://github.com/JustinaZ/pcaReduce">https://github.com/JustinaZ/pcaReduce</a>	Zurauskiene and Yau, 2016
Seurat	<a href="https://github.com/satijalab/seurat">https://github.com/satijalab/seurat</a>	Satija et al., 2015
TSCAN	<a href="https://github.com/zij90/TSCAN">https://github.com/zij90/TSCAN</a>	Ji and Ji, 2016

designed for scRNA-seq to identify DEGs in consideration of the complex features of scRNA-seq data.

## CELL LINEAGE AND PSEUDOTIME RECONSTRUCTION

The cells in many biological systems exhibit a continuous spectrum of states and involve transitions between different cellular states. Such dynamic processes within a portion of cells can be computationally modeled by reconstructing the cell trajectory and pseudotime based on scRNA-seq data. Pseudotime is an ordering of cells along the trajectory of a continuously developmental process in a system, which allows the identification of the cell types at the beginning, intermediate, and end states of the trajectory (Griffiths et al., 2018). Besides revealing the gene expression dynamics across cells, single-cell trajectory inference can also benefit the identification of the factors triggering state transitions. A number of tools have been proposed for trajectory inference, e.g., Monocle (Trapnell et al., 2014), Waterfall (Shin et al., 2015), Wishbone (Setty et al., 2016), TSCAN (Ji and Ji, 2016), Monocle2 (Qiu et al., 2017), Slingshot (Street et al., 2018), and CellRouter (Lummertz da Rocha et al., 2018) (Table 5). The resulting trajectory topology can be linear, bifurcating, or a tree/graph structure. Monocle builds a minimum spanning tree (MST) for cells to search for the longest backbone based on independent component analysis (ICA). Monocle2 uses a distinct approach that incorporates unsupervised data-driven methods with reversed graph embedding (RGE), which is more robust and much faster than Monocle. Slingshot is a cluster-based approach for identifying multiple trajectories with varying levels of supervision. CellRouter utilizes flow networks to identify cell-state transition trajectories. Recently, Saelens et al. (2018) evaluated a number of single-cell trajectory inference approaches (did not include CellRouter), and found that Slingshot, TSCAN and Monocle2 outperform other methods.

## ALTERNATIVE SPLICING AND RNA EDITING ANALYSIS OF SCRNA-SEQ DATA

Most of published single-cell studies mainly explored the transcriptome variation between individual cells at gene level. In eukaryotic genome, AS allows multi-exon genes to generate different isoforms, which can largely increase the diversity of both protein-coding and noncoding RNAs. Five basic modes are generally recognized for AS, including exon-skipping (cassette exon), mutually exclusive exons, alternative donor site, alternative acceptor site, and intron retention. Lots of studies have shown that AS is very common in mammals and over 90% of human genes could undergo AS based on bulk RNA-seq data (Wang et al., 2008; Chen et al., 2017a). Moreover, AS play crucial roles in a variety of biological processes and abnormal AS may be correlated with cancers (Sveen et al., 2016). The findings revealed by bulk RNA-seq data can only reflect the averaged AS patterns of numerous cells at population level.

**TABLE 5 |** Methods for single-cell trajectory inference.

Tools	Dimensionality reduction	URL	References
Monocle	ICA	<a href="http://cole-trapnell-lab.github.io/monocle-release/">http://cole-trapnell-lab.github.io/monocle-release/</a>	Trapnell et al., 2014
Waterfall	PCA	<a href="https://www.cell.com/cms/10.1016/j.stem.2015.07.013/attachment/3e966901-034f-418a-a439-996c50292a11/mmc9.zip">https://www.cell.com/cms/10.1016/j.stem.2015.07.013/attachment/3e966901-034f-418a-a439-996c50292a11/mmc9.zip</a>	Shin et al., 2015
Wishbone	Diffusion maps	<a href="https://github.com/ManuSetty/wishbone">https://github.com/ManuSetty/wishbone</a>	Setty et al., 2016
GrandPrix	Gaussian Process Latent Variable Model	<a href="https://github.com/ManchesterBioinference/GrandPrix">https://github.com/ManchesterBioinference/GrandPrix</a>	Ahmed et al., 2019
SCUBA	t-SNE	<a href="https://github.com/gcyuan/SCUBA">https://github.com/gcyuan/SCUBA</a>	Marco et al., 2014
DPT	Diffusion maps	<a href="https://media.nature.com/original/nature-assets/nmeth/journal/v13/n10/extref/nmeth.3971-S3.zip">https://media.nature.com/original/nature-assets/nmeth/journal/v13/n10/extref/nmeth.3971-S3.zip</a>	Haghighverdi et al., 2016
TSCAN	PCA	<a href="https://github.com/zji90/TSCAN">https://github.com/zji90/TSCAN</a>	Ji and Ji, 2016
Monocle2	RGE	<a href="http://cole-trapnell-lab.github.io/monocle-release/">http://cole-trapnell-lab.github.io/monocle-release/</a>	Qiu et al., 2017
Slingshot	Any	<a href="https://github.com/kstreet13/slingshot">https://github.com/kstreet13/slingshot</a>	Street et al., 2018
CellRouter	Any	<a href="https://github.com/edroaldo/cellrouter">https://github.com/edroaldo/cellrouter</a>	Lummertz da Rocha et al., 2018

Due to the high noise (e.g., dropouts and uneven transcript coverage) and low sequencing coverage of scRNA-seq data, the splicing quantification methods initially developed for bulk RNA-seq data are not suitable for scRNA-seq data. Since expression dynamics is a key aspect of cell populations, it is promising to study AS at single-cell resolution to gain insights into cell-level isoform usage. To date, only a few number of AS detection approaches were devised for scRNA-seq data, such as SingleSplice (Welch et al., 2016), Census (Qiu et al., 2017), BRIE (Huang and Sanguinetti, 2017), and Expedition (Song et al., 2017) (Table 6). SingleSplice uses a statistical model to detect the genes with a significant isoform usage without estimating the expression levels of full-length transcripts. Census models the isoform counts of each gene with a linear model as a Dirichlet-multinomial distribution. BRIE is a Bayesian hierarchical model for differential isoform quantification. Expedition contains a suite of algorithms for identifying AS, assigning splicing modalities and visualize modality changes. The AS detection approaches specially designed for scRNA-seq data are just emerging, thus the innovation and improvement of such methods will largely facilitate AS exploration at the single-cell level.

On the other hand, RNA-editing is an important post-transcriptional processing event that leads to sequence changes on RNA molecules (Gott and Emeson, 2000). Similarly, RNA-editing is mainly studied using bulk RNA-seq technologies but rarely explored at the single-cell level. Currently, the limitations of scRNA-seq largely prevented the application of RNA-editing

**TABLE 6 |** Alternative splicing detection tools for scRNA-seq data.

Tools	URL	References
SingleSplice	<a href="https://github.com/jw156605/SingleSplice">https://github.com/jw156605/SingleSplice</a>	Welch et al., 2016
Expedition	<a href="https://github.com/YeoLab/Expedition">https://github.com/YeoLab/Expedition</a>	Song et al., 2017
BRIE	<a href="https://github.com/huangyh09/brie">https://github.com/huangyh09/brie</a>	Huang and Sanguinetti, 2017
Census	<a href="http://cole-trapnell-lab.github.io/monocle-release/">http://cole-trapnell-lab.github.io/monocle-release/</a>	Qiu et al., 2017

detection to individual cells. Accordingly, with the development of both scRNA-seq technologies and single-cell editing detection algorithms, exploration of RNA-editing dynamics among single cells will be feasible. Notably, both AS and RNA-editing are mainly suitable for the data generated by scRNA-seq protocols that can sequence full-length transcripts such as Smart-seq2 and MATQ-seq rather than 3'-end scRNA-seq approaches.

## ALLELIC EXPRESSION EXPLORATION WITH SCRNA-SEQ DATA

Diploid species contain two sets of chromosomes that are separately obtained from their parents. Allelic expression analysis can reveal whether genes are equally expressed between parental and maternal genomes. For autosomes, the parental and maternal expression are generally expressed equally, and aberrant expression of parental or maternal genome may cause certain diseases (McKean et al., 2016). Up to now, few methods were developed to detect the genome-wide allelic expression profile of genes based on scRNA-seq data. One main caution of allelic expression calling is that the high dropouts of scRNA-seq data may introduce many false positives. Deng et al. (2014) used a series of stringent criteria to filter the potentially false allelic calls resulted from the technical variability of scRNA-seq in studying allelic expression profile of mouse preimplantation embryos. The robustness of this strategy was further demonstrated in analyzing the dynamics of X chromosome inactivation along developmental progression using mouse embryonic stem cells (Chen et al., 2016a). SCALE was recently proposed to classify the gene expression into silent, monoallelic and biallelic, states by adopting an empirical Bayes approach (Jiang et al., 2017). We believe that allelic expression analysis at single-cell level can largely facilitate the understanding of the underlying mechanisms of dosage compensation and related diseases. It is worth noting that allelic expression investigation at single-cell level also needs the whole-transcript scRNA-seq and is mainly applicable to the organism that has available paternal and maternal single nucleotide polymorphism (SNP) information.

## GENE REGULATORY NETWORK RECONSTRUCTION

Gene regulatory network inference has been widely conducted in numerous bulk RNA-seq studies, while scRNA-seq also

provides great potential for such analysis. For bulk RNA-seq data, networks are usually constructed from a number of samples using the tools like weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008; Chen et al., 2017a). A basic assumption is that the genes highly correlated in expression could be co-regulated. Because such an analysis is unable to determine the regulatory relationship, the resulting networks are typically undirected. Theoretically, the cells of scRNA-seq can be treated as the samples of bulk RNA-seq, then similar approaches are applicable to scRNA-seq data for constructing gene regulatory network.

Network inference of scRNA-seq data may reveal meaningful gene correlations and provide biologically important insights that could not be uncovered by population-level data of bulk RNA-seq. However, due to the technical noise of scRNA-seq and different subpopulations or states of cells, attention should be paid to network reconstruction. To reduce spurious results, network inference should be carried out on each subpopulation or the cells with the same stage. Recently, Aibar et al. (2017) developed SCENIC method to reconstruct the gene regulatory network from scRNA-seq data and they showed that SCENIC can robustly predict the interactions between transcription factors and target genes. PIDC is another software designed to infer gene regulatory network from single-cell data using multivariate information theory (Chan et al., 2017). Such network inference tools facilitate the identification of expression regulatory network from single-cell transcriptomic data and provide critically biological insights into the regulatory relationships between genes.

## CONCLUSION

In the past 10 years, a great advancement has been achieved in scRNA-seq and a variety of scRNA-seq protocols have been developed. The development and innovation of scRNA-seq largely facilitated single-cell transcriptomic studies, leading to insightful findings in cell expression variability and dynamics. Moreover, the throughput of scRNA-seq has significantly increased with the exciting progress in cellular barcoding and microfluidics. Meanwhile, scRNA-seq methods that can be used for fixation and frozen samples have also been proposed recently, which will greatly benefit the study of highly heterogeneous clinical samples. However, currently available scRNA-seq approaches still have a high dropout problem, in which weakly expressed genes would be missed. The improvement of RNA capture efficiency and transcript coverage will definitely reduce the technical noise of scRNA-seq. Moreover, since most of current scRNA-seq methods mainly capture polyA<sup>+</sup> RNAs, the development of protocols that can capture both polyA<sup>+</sup> and polyA<sup>−</sup> RNAs (such as MATQ-seq) will enable comprehensive investigation of both protein-coding and non-coding gene expression dynamics at single-cell resolution.

Since the noise of scRNA-seq data is high, it is crucial to use appropriate methods to overcome the problem in analyzing scRNA-seq data. QC is necessary to exclude those low-quality cells to avoid involving artifacts in data interpretation. Furthermore, batch effect correction (if need), between sample



normalization and imputation are also important and should be conducted before cell subpopulation identification, differential expression calling, and other downstream analyses. Additionally, factors such as cell size and cell cycle state could play important roles in cell variability for certain types of cells, such biases are also need to be considered. Although an increasing number of methods have been specially designed to interpret scRNA-seq data, advances of novel methods that can effectively handle the technical noise and expression variability of cells are still required. Specifically, the approaches that can accurately analyze AS and RNA-editing with scRNA-seq data are highly useful to unravel post-transcriptional mechanisms in individual cells. Overall, bioinformatics analysis of scRNA-seq data is still challenging, special attention should be paid in data interpretation, and more efficient tools are in urgent need.

Collectively, scRNA-seq and its related computational methods largely promote the development of single-cell

transcriptomics. The continuous innovation of scRNA-seq technologies and concomitant advances in bioinformatics approaches will greatly facilitate biological and clinical researches, and provide deep insights into the gene expression heterogeneity and dynamics of cells.

## AUTHOR CONTRIBUTIONS

GC and TS designed the study and wrote the manuscript. BN edited the manuscript and provided constructive comments.

## FUNDING

This work was supported by the National Science Foundation of China (31771460, 91629103 and 31671377), National Key Research and Development Program of China (2016YFC0902100).

## REFERENCES

- Ahmed, S., Rattray, M., and Boukouvelas, A. (2019). GrandPrix: scaling up the Bayesian GPLVM for single-cell data. *Bioinformatics* 35, 47–54. doi: 10.1093/bioinformatics/bty533
- Aibar, S., Gonzalez-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. doi: 10.1038/nmeth.4463
- Andrews, T. S., and Hemberg, M. (2018a). Identifying cell populations with scRNASeq. *Mol. Aspects Med.* 59, 114–122. doi: 10.1016/j.mam.2017.07.002
- Andrews, T. S., and Hemberg, M. (2018b). M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics* doi: 10.1093/bioinformatics/bty1044 [Epub ahead of print].
- Angerer, P., Haghighi, L., Buttner, M., Theis, F. J., Marr, C., and Buettner, F. (2016). destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* 32, 1241–1243. doi: 10.1093/bioinformatics/btv715
- Bacher, R., Chu, L. F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., et al. (2017). SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* 14, 584–586. doi: 10.1038/nmeth.4263
- Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* 17:63. doi: 10.1186/s13059-016-0927-y
- Barrett, S. P., and Salzman, J. (2016). Circular RNAs: analysis, expression and potential functions. *Development* 143, 1838–1847. doi: 10.1242/dev.128074
- Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., et al. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. doi: 10.1038/nbt.4314
- Brennecke, P., Anders, S., Kim, J. K., Kolodziejczyk, A. A., Zhang, X., Proserpio, V., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* 10, 1093–1095. doi: 10.1038/nmeth.2645
- Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., et al. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* 33, 155–160. doi: 10.1038/nbt.3102
- Buttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., and Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* 16, 43–49. doi: 10.1038/s41592-018-0254-1
- Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661–667. doi: 10.1126/science.aam8940
- Chan, T. E., Stumpf, M. P. H., and Babbie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* 5, 251–267.e3. doi: 10.1016/j.cels.2017.08.014
- Chen, G., Chen, J., Yang, J., Chen, L., Qu, X., Shi, C., et al. (2017a). Significant variations in alternative splicing patterns and expression profiles between human-mouse orthologs in early embryos. *Sci. China Life Sci.* 60, 178–188. doi: 10.1007/s11427-015-0348-5
- Chen, G., Shi, T. L., and Shi, L. M. (2017b). Characterizing and annotating the genome using RNA-seq data. *Sci. China Life Sci.* 60, 116–125. doi: 10.1007/s11427-015-0349-4
- Chen, G., Schell, J. P., Benitez, J. A., Petropoulos, S., Yilmaz, M., Reinius, B., et al. (2016a). Single-cell analyses of X Chromosome inactivation dynamics and pluripotency during differentiation. *Genome Res.* 26, 1342–1354. doi: 10.1101/gr.201954.115
- Chen, G., Yang, J., Chen, J., Song, Y., Cao, R., Shi, T., et al. (2016b). Identifying and annotating human bifunctional RNAs reveals their versatile functions. *Sci. China Life Sci.* 59, 981–992. doi: 10.1007/s11427-016-0054-1
- Chen, G., Wang, C., and Shi, T. (2011). Overview of available methods for diverse RNA-Seq data analyses. *Sci. China Life Sci.* 54, 1121–1128. doi: 10.1007/s11427-011-4255-x
- Chen, L., and Zheng, S. (2018). BCseq: accurate single cell RNA-seq quantification with bias correction. *Nucleic Acids Res.* 46:e82. doi: 10.1093/nar/gky308
- Chen, X., Teichmann, S. A., and Meyer, K. B. (2018). From tissues to cell types and back: single-cell gene expression analysis of tissue architecture. *Annu. Rev. Biomed. Data Sci.* 1, 29–51. doi: 10.1146/annurev-biodatasci-080917-013452
- Delmans, M., and Hemberg, M. (2016). Discrete distributional differential expression (D3E)—a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics* 17:110. doi: 10.1186/s12859-016-0944-6
- Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343, 193–196. doi: 10.1126/science.1245316
- Ding, J., Condon, A., and Shah, S. P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* 9:2002. doi: 10.1038/s41467-018-04368-5
- Dobin, A., and Gingeras, T. R. (2015). Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinformatics* 51, 11.14.1–11.14.19. doi: 10.1002/0471250953.bi114551
- Engstrom, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Ratsch, G., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* 10, 1185–1191. doi: 10.1038/nmeth.2722
- External, R. N. A. C. C. (2005). Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics* 6:150. doi: 10.1186/1471-2164-6-150
- Fan, H. C., Fu, G. K., and Fodor, S. P. (2015). Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science* 347:1258367. doi: 10.1126/science.1258367
- Fan, X., Zhang, X., Wu, X., Guo, H., Hu, Y., Tang, F., et al. (2015). Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse



- preimplantation embryos. *Genome Biol.* 16:148. doi: 10.1186/s13059-015-0706-1
- Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., et al. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* 13, 241–244. doi: 10.1038/nmeth.3734
- Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16:278. doi: 10.1186/s13059-015-0844-5
- Frazee, A. C., Pertea, G., Jaffe, A. E., Langmead, B., Salzberg, S. L., and Leek, J. T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat. Biotechnol.* 33, 243–246. doi: 10.1038/nbt.3172
- Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* 8, 469–477. doi: 10.1038/nmeth.1613
- Gierahn, T. M., Wadsworth, M. H. II, Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., et al. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* 14, 395–398. doi: 10.1038/nmeth.4179
- Gong, W., Kwak, I. Y., Pota, P., Koyano-Nakagawa, N., and Garry, D. J. (2018). DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* 19:220. doi: 10.1186/s12859-018-2226-y
- Gott, J. M., and Emeson, R. B. (2000). Functions and mechanisms of RNA editing. *Annu. Rev. Genet.* 34, 499–531. doi: 10.1146/annurev.genet.34.1.499
- Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* 14:e8046. doi: 10.1525/msb.20178046
- Gross, A., Schoendube, J., Zimmermann, S., Steeb, M., Zengerle, R., and Koltay, P. (2015). Technologies for single-cell isolation. *Int. J. Mol. Sci.* 16, 16897–16919. doi: 10.3390/ijms160816897
- Grun, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525, 251–255. doi: 10.1038/nature14966
- Habib, N., Avraham-David, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., et al. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* 14, 955–958. doi: 10.1038/nmeth.4407
- Haghverdi, L., Buttner, M., Wolf, F. A., Büttner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 13, 845–848. doi: 10.1038/nmeth.3971
- Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36, 421–427. doi: 10.1038/nbt.4091
- Haue, A., Engel, J., Teichmann, S. A., and Lonnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* 9:75. doi: 10.1186/s13073-017-0467-4
- Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., et al. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* 17:77. doi: 10.1186/s13059-016-0938-8
- Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* 2, 666–673. doi: 10.1016/j.celrep.2012.08.003
- Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19, 562–578. doi: 10.1093/biostatistics/kxx053
- Hu, P., Zhang, W., Xin, H., and Deng, G. (2016). Single cell isolation and analysis. *Front. Cell. Dev. Biol.* 4:116. doi: 10.3389/fcell.2016.00116
- Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., et al. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* 15, 539–542. doi: 10.1038/s41592-018-0033-z
- Huang, Y., and Sanguinetti, G. (2017). BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biol.* 18:123. doi: 10.1186/s13059-017-1248-5
- Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* 50:96. doi: 10.1038/s12276-018-0071-8
- Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., et al. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* 17:29. doi: 10.1186/s13059-016-0888-1
- Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J. B., Lonnerberg, P., et al. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21, 1160–1167. doi: 10.1101/gr.110882.110
- Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J. B., Lonnerberg, P., et al. (2012). Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nat. Protoc.* 7, 813–828. doi: 10.1038/nprot.2012.022
- Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., et al. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* 11, 163–166. doi: 10.1038/nmeth.2772
- Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343, 776–779. doi: 10.1126/science.1247651
- Ji, Z., and Ji, H. (2016). TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* 44:e117. doi: 10.1093/nar/gkw430
- Jiang, P., Thomson, J. A., and Stewart, R. (2016). Quality control of single-cell RNA-seq by SinQC. *Bioinformatics* 32, 2514–2516. doi: 10.1093/bioinformatics/btw176
- Jiang, Y., Zhang, N. R., and Li, M. (2017). SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol.* 18:74. doi: 10.1186/s13059-017-1200-8
- Katayama, S., Tohonen, V., Linnarsson, S., and Kere, J. (2013). SAMstrat: statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics* 29, 2943–2945. doi: 10.1093/bioinformatics/btt511
- Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11, 740–742. doi: 10.1038/nmeth.2967
- Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36
- Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14, 483–486. doi: 10.1038/nmeth.4236
- Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201. doi: 10.1016/j.cell.2015.04.044
- Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620. doi: 10.1016/j.molcel.2015.04.005
- Kristensen, L. S., Hansen, T. B., Veno, M. T., and Kjems, J. (2018). Circular RNAs in cancer: opportunities and challenges in the field. *Oncogene* 37, 555–565. doi: 10.1038/onc.2017.361
- Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559
- Leek, J. T. (2014). svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* 42:e161. doi: 10.1093/nar/gku864
- Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733–739. doi: 10.1038/nrg2825
- Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323
- Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* 11, 473–483. doi: 10.1093/bib/bbq015
- Li, W. V., and Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* 9:997. doi: 10.1038/s41467-018-03405-7
- Lin, P., Troup, M., and Ho, J. W. (2017). CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* 18:59. doi: 10.1186/s13059-017-1188-0

- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8
- Lummertz da Rocha, E., Rowe, R. G., Lundin, V., Malleshaiah, M., Jha, D. K., Rambo, C. R., et al. (2018). Reconstruction of complex single-cell trajectories using CellRouter. *Nat. Commun.* 9:892. doi: 10.1038/s41467-018-03214-y
- Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17:75. doi: 10.1186/s13059-016-0947-7
- Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002
- Marco, E., Karp, R. L., Guo, G., Robson, P., Hart, A. H., Trippa, L., et al. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. U.S.A.* 111, E5643–E5650. doi: 10.1073/pnas.1408931111
- McCarthy, D. J., Campbell, K. R., Lun, A. T., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33, 1179–1186. doi: 10.1093/bioinformatics/btw777
- McDavid, A., Finak, G., Chattopadhyay, P. K., Dominguez, M., Lamoreaux, L., Ma, S. S., et al. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* 29, 461–467. doi: 10.1093/bioinformatics/bts714
- McKean, D. M., Homsy, J., Wakimoto, H., Patel, N., Gorham, J., DePalma, S. R., et al. (2016). Loss of RNA expression and allele-specific expression associated with congenital heart disease. *Nat. Commun.* 7:12824. doi: 10.1038/ncomms12824
- Miao, Z., Deng, K., Wang, X., and Zhang, X. (2018). DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* 34, 3223–3224. doi: 10.1093/bioinformatics/bty332
- Nichterwitz, S., Chen, G., Aguila Benitez, J., Yilmaz, M., Storvall, H., Cao, M., et al. (2016). Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. *Nat. Commun.* 7:12139. doi: 10.1038/ncomms12139
- Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122
- Picelli, S. (2017). Single-cell RNA-sequencing: the future of genome biology is now. *RNA Biol.* 14, 637–650. doi: 10.1080/15476286.2016.1201618
- Picelli, S., Bjorklund, A. K., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10, 1096–1098. doi: 10.1038/nmeth.2639
- Pierson, E., and Yau, C. (2015). ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* 16:241. doi: 10.1186/s13059-015-0805-z
- Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y. A., and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* 14, 309–315. doi: 10.1038/nmeth.4150
- Quinn, J. J., and Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* 17, 47–62. doi: 10.1038/nrg.2015.10
- Ramskold, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., Faridani, O. R., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30, 777–782. doi: 10.1038/nbt.2282
- Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902. doi: 10.1038/nbt.2931
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616
- Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25. doi: 10.1186/gb-2010-11-3-r25
- Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360, 176–182. doi: 10.1126/science.aam8999
- Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2018). A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv* [Preprint]. doi: 10.1101/276907
- Sasagawa, Y., Danno, H., Takada, H., Ebisawa, M., Tanaka, K., Hayashi, T., et al. (2018). Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* 19:29. doi: 10.1186/s13059-018-1407-3
- Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K. D., Imai, T., et al. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* 14:R31. doi: 10.1186/gb-2013-14-4-r31
- Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33, 495–502. doi: 10.1038/nbt.3192
- Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., et al. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* 34, 637–645. doi: 10.1038/nbt.3569
- Seyednasrollah, F., Rantanen, K., Jaakkola, P., and Elo, L. L. (2016). ROTS: reproducible RNA-seq biomarker detector-prognostic markers for clear cell renal cell cancer. *Nucleic Acids Res.* 44:e1. doi: 10.1093/nar/gkv806
- Sheng, K., Cao, W., Niu, Y., Deng, Q., and Zong, C. (2017). Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods* 14, 267–270. doi: 10.1038/nmeth.4145
- Shin, J., Berg, D. A., Zhu, Y., Shin, J. Y., Song, J., Bonaguidi, M. A., et al. (2015). Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* 17, 360–372. doi: 10.1016/j.stem.2015.07.013
- Soneson, C., and Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* 15, 255–261. doi: 10.1038/nmeth.4612
- Song, Y., Botvinnik, O. B., Lovci, M. T., Kakaradov, B., Liu, P., Xu, J. L., et al. (2017). Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Mol. Cell* 67, 148–161.e5. doi: 10.1016/j.molcel.2017.06.003
- Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16, 133–145. doi: 10.1038/nrg3833
- Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., et al. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19:477. doi: 10.1186/s12864-018-4772-0
- Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R. A., and Skotheim, R. I. (2016). Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene* 35, 2413–2427. doi: 10.1038/nc.2015.318
- Svensson, V., Natarajan, K. N., Ly, L. H., Miragaia, R. J., Labalette, C., Macaulay, I. C., et al. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14, 381–387. doi: 10.1038/nmeth.4220
- Talwar, D., Mongia, A., Sengupta, D., and Majumdar, A. (2018). AutoImpute: autoencoder based imputation of single-cell RNA-seq data. *Sci. Rep.* 8:16329. doi: 10.1038/s41598-018-34688-x
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi: 10.1038/nmeth.1315
- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi: 10.1038/nbt.2859
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621
- Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). BASICS: bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* 11:e1004333. doi: 10.1371/journal.pcbi.1004333

- Vallejos, C. A., Richardson, S., and Marioni, J. C. (2016). Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biol.* 17:70. doi: 10.1186/s13059-016-0930-3
- Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* 14, 565–571. doi: 10.1038/nmeth.4292
- van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.
- van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell* 174, 716–729.e27. doi: 10.1016/j.cell.2018.05.061
- Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen, M., et al. (2016). Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* 32, 2128–2135. doi: 10.1093/bioinformatics/btw202
- Wang, E. T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476. doi: 10.1038/nature07509
- Welch, J. D., Hu, Y., and Prins, J. F. (2016). Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Res.* 44:e73. doi: 10.1093/nar/gkv1525
- Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980. doi: 10.1093/bioinformatics/btv088
- Zeisel, A., Munoz-Manchado, A. B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142. doi: 10.1126/science.aaa1934
- Zhang, L., and Zhang, S. (2018). Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2018.2848633 [Epub ahead of print].
- Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049. doi: 10.1038/ncomms14049
- Ziegenhain, C., Vieth, B., Parekh, S., Reinus, B., Guillaumet-Adkins, A., Smets, M., et al. (2017). Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* 65, 631–643.e4. doi: 10.1016/j.molcel.2017.01.023
- Zurauskiene, J., and Yau, C. (2016). pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* 17:140. doi: 10.1186/s12859-016-0984-y

**Disclaimer:** The information in these materials is not a formal dissemination of the United States Food and Drug Administration.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Chen, Ning and Shi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Benchmark and Parameter Sensitivity Analysis of Single-Cell RNA Sequencing Clustering Methods

Monika Krzak<sup>1\*</sup>, Yordan Raykov<sup>2</sup>, Alexis Boukouvalas<sup>3</sup>, Luisa Cuttillo<sup>4</sup> and Claudia Angelini<sup>1</sup>

<sup>1</sup> Institute for Applied Mathematics "Mauro Picone", Naples, Italy, <sup>2</sup> Department of Mathematics, Aston University, Birmingham, United Kingdom, <sup>3</sup> Machine Learning Engineer Team, Prowler.io, Cambridge, United Kingdom, <sup>4</sup> School of Mathematics, University of Leeds, Leeds, United Kingdom

## OPEN ACCESS

### Edited by:

Filippo Geraci,  
Italian National Research Council,  
Italy

### Reviewed by:

Giovanna Rosone,  
University of Pisa, Italy  
Antonio Federico,  
Tampere University, Finland

### \*Correspondence:

Monika Krzak  
monika.sonia.krzak@gmail.com

### Specialty section:

This article was submitted to  
Bioinformatics and  
Computational Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 19 July 2019

**Accepted:** 13 November 2019

**Published:** 11 December 2019

### Citation:

Krzak M, Raykov Y, Boukouvalas A,  
Cuttillo L and Angelini C (2019)  
Benchmark and Parameter Sensitivity  
Analysis of Single-Cell RNA  
Sequencing Clustering Methods.  
Front. Genet. 10:1253.  
doi: 10.3389/fgene.2019.01253

Single-cell RNA-seq (scRNAseq) is a powerful tool to study heterogeneity of cells. Recently, several clustering based methods have been proposed to identify distinct cell populations. These methods are based on different statistical models and usually require to perform several additional steps, such as preprocessing or dimension reduction, before applying the clustering algorithm. Individual steps are often controlled by method-specific parameters, permitting the method to be used in different modes on the same datasets, depending on the user choices. The large number of possibilities that these methods provide can intimidate non-expert users, since the available choices are not always clearly documented. In addition, to date, no large studies have investigated the role and the impact that these choices can have in different experimental contexts. This work aims to provide new insights into the advantages and drawbacks of scRNAseq clustering methods and describe the ranges of possibilities that are offered to users. In particular, we provide an extensive evaluation of several methods with respect to different modes of usage and parameter settings by applying them to real and simulated datasets that vary in terms of dimensionality, number of cell populations or levels of noise. Remarkably, the results presented here show that great variability in the performance of the models is strongly attributed to the choice of the user-specific parameter settings. We describe several tendencies in the performance attributed to their modes of usage and different types of datasets, and identify which methods are strongly affected by data dimensionality in terms of computational time. Finally, we highlight some open challenges in scRNAseq data clustering, such as those related to the identification of the number of clusters.

**Keywords:** single-cell RNA-seq, clustering methods, benchmark, parameter sensitivity analysis, high-dimensional data analysis

## INTRODUCTION

Single-cell RNA sequencing (scRNAseq) has emerged as an important technology that allows profiling gene expression at single-cell resolution, giving new insights into cellular development (Biase et al., 2014; Goolam et al., 2016), dynamics (Vuong et al., 2018; Farbehi et al., 2019), and cell composition (Darmanis et al., 2015; Zeisel et al., 2015; Segerstolpe et al., 2016). Although the scRNAseq analysis inherits many features from bulk RNA-seq approaches, the algorithms require



constant adaptation due to the several types of challenges present in scRNAseq data (Kiselev et al., 2019). For example, current droplet-based technologies allow measuring hundreds of thousands of cells which greatly exceeds the number of samples typically handled by bulk RNA-seq protocols. The low amount of measured RNA transcripts per cell and stochastic nature of the genes expression can also introduce missing information about gene profiles (dropouts). The scRNAseq data specific noise and the increasing number of scRNAseq protocols differing in accuracy and scalability (Svensson et al., 2017; Svensson et al., 2018) make the systematic data analysis even more challenging.

Over the last few years, a number of computational algorithms have been proposed to analyze scRNAseq data, focusing on different aspects (Chen et al., 2019). In particular, a growing class of computational methods is being developed for identifying distinct cell populations (Andrews and Hemberg, 2018). These methods are based on various types of clustering techniques, which aim to divide cells into groups that share similar gene expression patterns. In this way, each group can be associated with a specific cell type or subtype on the basis of well-known markers, or novel cell subtypes can be identified. However, before applying the clustering algorithm, such methods often require to perform a series of mandatory or optional steps that include preprocessing, filtering or dimension reduction (Luecken and Theis, 2019). In several cases, such steps can be adapted by the user by choosing an appropriate set of parameters. Thus, methods turn to be very heterogeneous in the way they model data and perform the individual steps. Differences arise at each stage of the analysis and are not yet fully understood. For example, some algorithms work with raw count dataset (Zurauskiene and Yau, 2016; Lin et al., 2017; Sun et al., 2018), others require normalized gene expression values (Macosko et al., 2015; Ji and Ji, 2016; Senabouth et al., 2019) or can handle both formats (Yip et al., 2017; Qiu et al., 2017; Kiselev et al., 2017; Wang et al., 2017). Some of the tools do incorporate an additional method-specific preprocessing step in terms of filtering or normalization (Senabouth et al., 2019; Yip et al., 2017), to remove noise present in the data, other require such step to be done externally before the execution of the method (Julia et al., 2015). In addition to preprocessing, many methods often utilize dimension reduction techniques, such as Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (tSNE), in order to reduce the high-dimensional space (expression of tens of thousands of genes) prior to clustering (Julia et al., 2015; Herman and Grün, 2018; Ren et al., 2019).

Another great difference is given by the specific clustering techniques implemented in each method. Some of the methods use partitioning algorithms (Kiselev et al., 2017; Wang et al., 2017) in order to infer distinct cell populations, others are based on hierarchical clustering (Senabouth et al., 2019; Lin et al., 2017), graph theory (Macosko et al., 2015) or density based-approach (Ester et al., 1996). There is also a growing class of model-based algorithms (Fraley and Raftery, 2002; Ji and Ji, 2016; Sun et al., 2018) which utilize probabilistic properties of a given model to account for distinct challenges present in the data. Moreover, some methods require the number of cell populations to be known in advance (Zurauskiene and Yau, 2016; Sun et al., 2018), while others estimate the optimal value with an external procedure or as part of the clustering inference (Macosko et al., 2015; Senabouth

et al., 2019; Ren et al., 2019). The available methods also vary in terms of the programming language they have been implemented in (i.e. R, Matlab, Python), computational cost and other system requirements.

All of the mentioned variations across clustering pipelines affect the performance of the methods. Currently, there is a limited amount of studies that infer clustering performance and robustness under various data-driven scenarios (Freytag et al., 2018; Duò et al., 2018; Tian et al., 2019). The main purpose of existing studies is to investigate the performance of the methods limited to a selected parameter setting. Such limitation leads to a narrow view on the performance of the methods making it difficult to explore their full potential and identify the open challenges. For example, some algorithms provide multiple possibilities in the choice of parameters (Julia et al., 2015; Qiu et al., 2017; Herman and Grün, 2018; Ren et al., 2019) that can allow the user to adapt/modify the main method in each step. At the same time, the selection of parameter settings can be crucial in various data-driven conditions. The performance of the algorithms can also depend on the presence or absence of any preprocessing steps, either external or method-specific, carried out prior to clustering. Since both, parameter settings and data preprocessing can greatly affect the clustering result, we decided to investigate both aspects on the performance of the methods by carrying a comprehensive benchmark of the existing clustering methods and performing parameter sensitivity analysis.

For that purpose, we first described different modes of usage and parameter settings of 13 among the most widely used scRNAseq clustering methods implemented in R, then we applied them on a large set of real scRNAseq and simulated datasets. In order to fully understand the potential of each method, we tested them varying a wide range of available parameter settings which greatly expands the number of possible results. Through the analysis pipeline, we evaluated the performance of the methods with respect to several factors. First, we divided the real datasets into two groups, those that were expressed in the raw counts and those expressed on normalized fragments per kilobase of transcript per million mapped reads (FPKM) or reads per kilobase of transcript per million mapped reads (RPKM) counts. On the first group, we evaluated the performance of the methods on three data basic preprocessing types (not preprocessed counts, filtered counts, filtered and normalized counts). On the second group, we evaluated the performance of the methods depending on a various number of dimensions supplied to dimension reduction techniques prior to clustering. Synthetic datasets were used to prove the capacity of each method in handling varying dataset dimensions that can additionally be diverse in the number of simulated cell groups and the type of group balance. In the simulation, we also accessed the accuracy of the methods in recovering cell population structure in the presence of noise. The type of noise that we simulated were dropouts and overlapping cell populations which are key features of scRNAseq datasets. In all cases, we evaluated the performance of the methods in terms of i) Adjusted Rand Index (ARI) index, ii) accuracy of methods in estimating the correct number of clusters, iii) running time.

Overall, this work aims to provide new insights into the advantages and drawbacks of several scRNAseq clustering methods,



by describing the ranges of possibilities that are offered to users and the impact that these choices can have on the final results. We also tried to identify some open challenges for future research that still need to be faced when doing the population inference.

## MATERIALS AND METHODS

### Real Datasets

In order to evaluate the performance of the clustering methods considered in this study we used 17 real scRNAseq datasets popular in the literature and listed in **Table 1**. To prepare the

gene expression matrix for clusterization, we followed the main instructions for data import and processing from the online repository <https://hemberg-lab.github.io/scRNA.seq.datasets/>.

The selected scRNAseq datasets vary in terms of organisms, tissues under study and experimental protocols. As illustrated in **Table 1**, some datasets were profiled using 3' or 5' tag and droplet-based approaches (such as inDrop), others using full-length plate-based approaches, such as Smart-Seq protocols. Moreover, depending on the used platform, each study investigates a different number of cells and data are subjected to a different proportion of dropouts. Depending on the protocol, count matrices were of different types (see **Table 2**) including Raw unique molecular

**TABLE 1** | List of the real datasets used to perform the clustering evaluation.

Single cell dataset	Organism	Cells under study	Protocol	Accession
Baron2016_m	Mouse	Pancreas	inDrop	GSE84133
Klein2015	Mouse	Embryonic stem cells	inDrop	GSE65525
Zeisel2015	Mouse	Cerebral cortex	STRT/C1 UMI	GSE60361
Darmanis2015	Human	Brain	SMARTer	GSE67835
Deng2014_raw	Mouse	Preimplantation embryos	Smart-Seq	GSE45719
Goolam2016	Mouse	Early embryos	Smart-Seq2	E-MTAB-3321
Kolodziejczyk2015	Mouse	Stem cells	SMARTer	E-MTAB-2600
Li2017	Human	Colorectal tumors	SMARTer	GSE81861
Romanov2016	Mouse	Hypothalamus	Fluidigm C1	GSE74672
Tasic2016_raw	Mouse	Brain	SMARTer	GSE71585
Deng2014_rpkm	Mouse	Preimplantation embryos	Smart-Seq	GSE45719
Segerstolpe2016	Human	Pancreas	Smart-Seq2	E-MTAB-5061
Tasic2016_rpkm	Mouse	Brain	SMARTer	GSE71585
Xin2016	Human	Pancreas	SMARTer	GSE81608
Yan2013	Human	Preimplantation embryos	Tang	GSE36552
Biase2014	Mouse	Embryos	SMARTer	GSE57249
Treutlein2014	Mouse	Lung epithelial cells	SMARTer	GSE52583

*Datasets (named by the author and date of publication) contain gene expression of cells from various organisms and tissues that have been processed by different experimental protocols. Protocols include 3' or 5' tag and droplet-based approaches (inDrop and STRT/C1 UMI), or full-length plate-based approaches, such as Smart-Seq, Smart-Seq2, SMARTer or Fluidigm C1. Tang protocol corresponds to mRNA-Seq assay described in (Tang et al., 2009). For more information about protocols see (Svensson et al., 2018).*

**TABLE 2** | Brief description of the main features of each real dataset considered in this study.

Single cell dataset	Data type	Nr cells	Nr cell populations	Publication
Baron2016_m	Raw UMI counts	1886	13	Baron et al. (2016)
Klein2015	Raw UMI counts	2717	4	Klein et al. (2015)
Zeisel2015	Raw UMI counts	3005	9	Zeisel et al. (2015)
Darmanis2015	Raw read counts	466	9	Darmanis et al. (2015)
Deng2014_raw	Raw read counts	268	6	Deng et al. (2014)
Goolam2016	Raw read counts	124	4	Goolam et al. (2016)
Kolodziejczyk2015	Raw read counts	704	3	Kolodziejczyk et al. (2015)
Li2017	Raw read counts	561	9	Li et al. (2017)
Romanov2016	Raw read counts	2881	7	Romanov et al. (2016)
Tasic2016_raw	Raw read counts	1679	18	Tasic et al. (2016)
Deng2014_rpkm	RPKM	268	6	Deng et al. (2014)
Segerstolpe2016	RPKM	3514	15	Segerstolpe et al. (2016)
Tasic2016_rpkm	RPKM	1679	18	Tasic et al. (2016)
Xin2016	RPKM	1600	8	Xin et al. (2016)
Yan2013	RPKM	90	6	Yan et al. (2013)
Biase2014	FPKM	56	4	Biase et al. (2014)
Treutlein2014	FPKM	80	5	Treutlein et al. (2014)

*Datasets can contain counts of 3 different types: Raw UMI counts, Raw read counts, and normalized FPKM/RPKM counts. Raw counts stands for the non-normalized counts that differ in terms of gene expression quantification method. FPKM/RPKM counts mean library size and gene length normalized counts. The number of reported cell populations is obtained from the annotation as described in the corresponding datasets publications.*

identifier (UMI) counts (3 datasets), Raw read counts (7 datasets) and FPKM/RPKM counts (7 datasets). The raw counts (either UMI or read counts) consist of datasets with gene expression quantified in terms of the number of mapped reads (counts) and that have not been further processed, while FPKM or RPKM data are library size and gene length adjusted counts. Note that two datasets in **Table 2**, Deng2014 and Tasic2016, were of both types (raw read counts and FPKM/RPKM counts). Overall, the datasets cover various ranges of experimental complexity in terms of the number of sequenced cells (from tens to several thousands) and number of cell populations in the sample (with minimum of 3 and maximum of 18 number of cell populations). The cell populations (hidden groups to detect) can represent distinct cell types or cells at various time points of differentiation. Within this study, we will consider the cell population annotation (available from the corresponding datasets studies) as ground truth, although we are aware that there could be some errors in the annotations, since datasets could contain some rare cell subpopulations, that were not identified at the time of the study, or some misclassified cells.

## Simulated Datasets

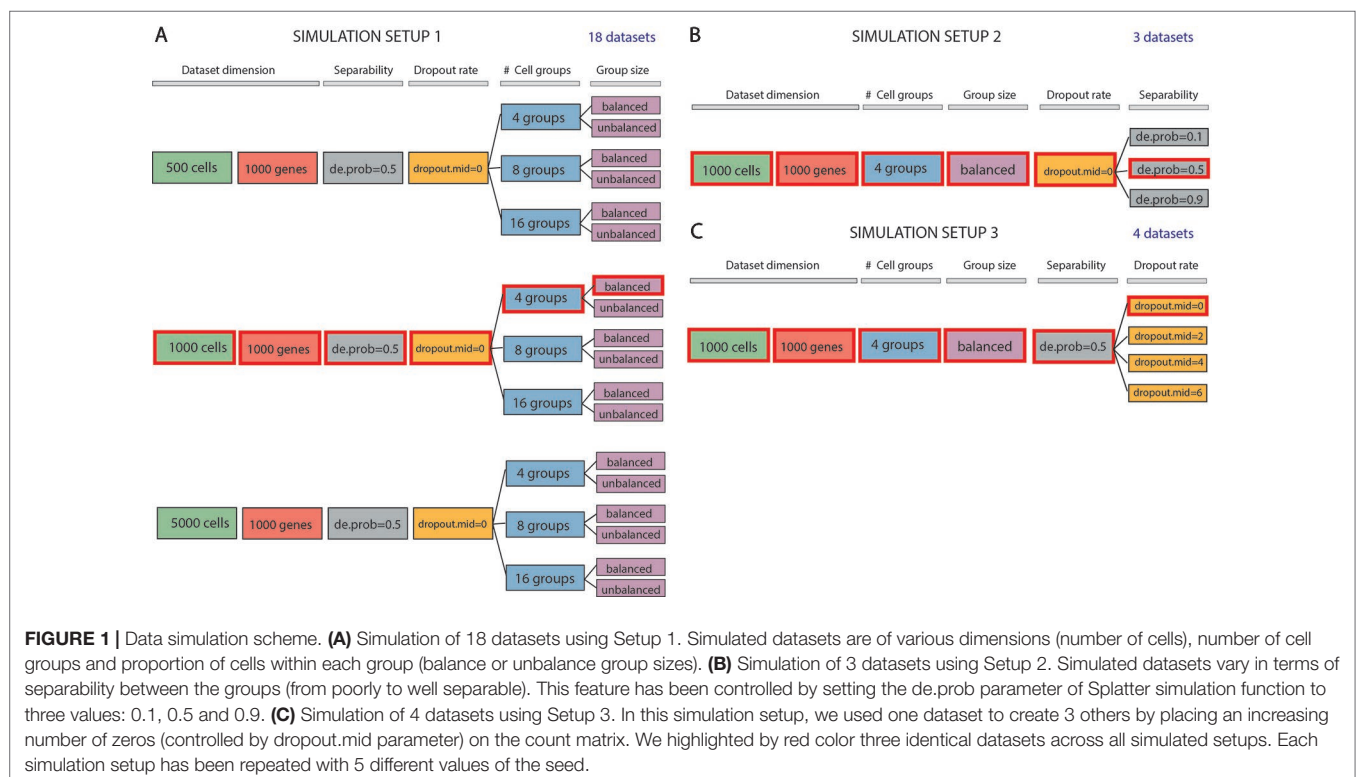
We evaluated methods performance also on synthetic datasets. The simulation study was performed using Splatter package (Zappia et al., 2017). Splatter allows simulating single-cell RNA sequencing count data with a varying number of cells and cell groups, with different degree of cluster separability and varying rate of dropouts. We designed three simulation setups that

allowed us to investigate various aspects of the performance of the methods (see **Figure 1**). Each simulation setup has been repeated 5 times choosing 5 different values of the seed.

In the first simulation setup (**Figure 1A**), we focused on assessing both the scalability (the capacity of each method in handling datasets with an increasing number of cells) and the complexity of the dataset (the ability of each method when the number of true groups increases or when the balancing between each group is disrupted). For this purpose, we simulated counts using three different values for the number of cells: 500, 1000 and 5000; three values for the number of groups: 4, 8, 16 and two possibilities for the number of cells in each of the group: balanced and unbalanced group size. In each of the modes, we set the number of genes to 1000. Therefore, the resulting 18 simulated datasets represent different levels of data complexity and size for the clustering task.

In the second simulation setup (**Figure 1B**), we fixed dataset dimension (1000 cells, 1000 genes) as well as the cell groups (fixed to four groups balanced in sizes) and we investigated the performance of each method with respect to the group separability ranging from poorly to well-separated groups. In such setup, we varied the probability of a gene to be differentially expressed to 0.1, 0.5, and 0.9, to obtain 3 simulated datasets: expression probability close to 1 gives highly separable cell groups that should be less difficult to be detected by any clustering algorithm.

Finally, in the third simulation setup (**Figure 1C**), we investigated the performance of clustering methods in the presence of various rates of missing information. With the number of cells and genes the same as before (1000) and cell groups fixed to four, we varied the rate of zero counts by setting



the midpoint parameter (drop.mid) for dropout logistic function to 0, 2, 4, and 6. In this way, we obtained 4 datasets with varying percentage of dropouts from 20% to 90%.

In each of the 5 runs of simulation, we have kept the synthetic datasets, highlighted in red in **Figures 1A–C** (i.e., those corresponding to 1000 cells, 1000 genes, 4 groups, size-balanced, de.prob = 0.5 and drop.mid = 0), identical across all three setups for easier direct comparison.

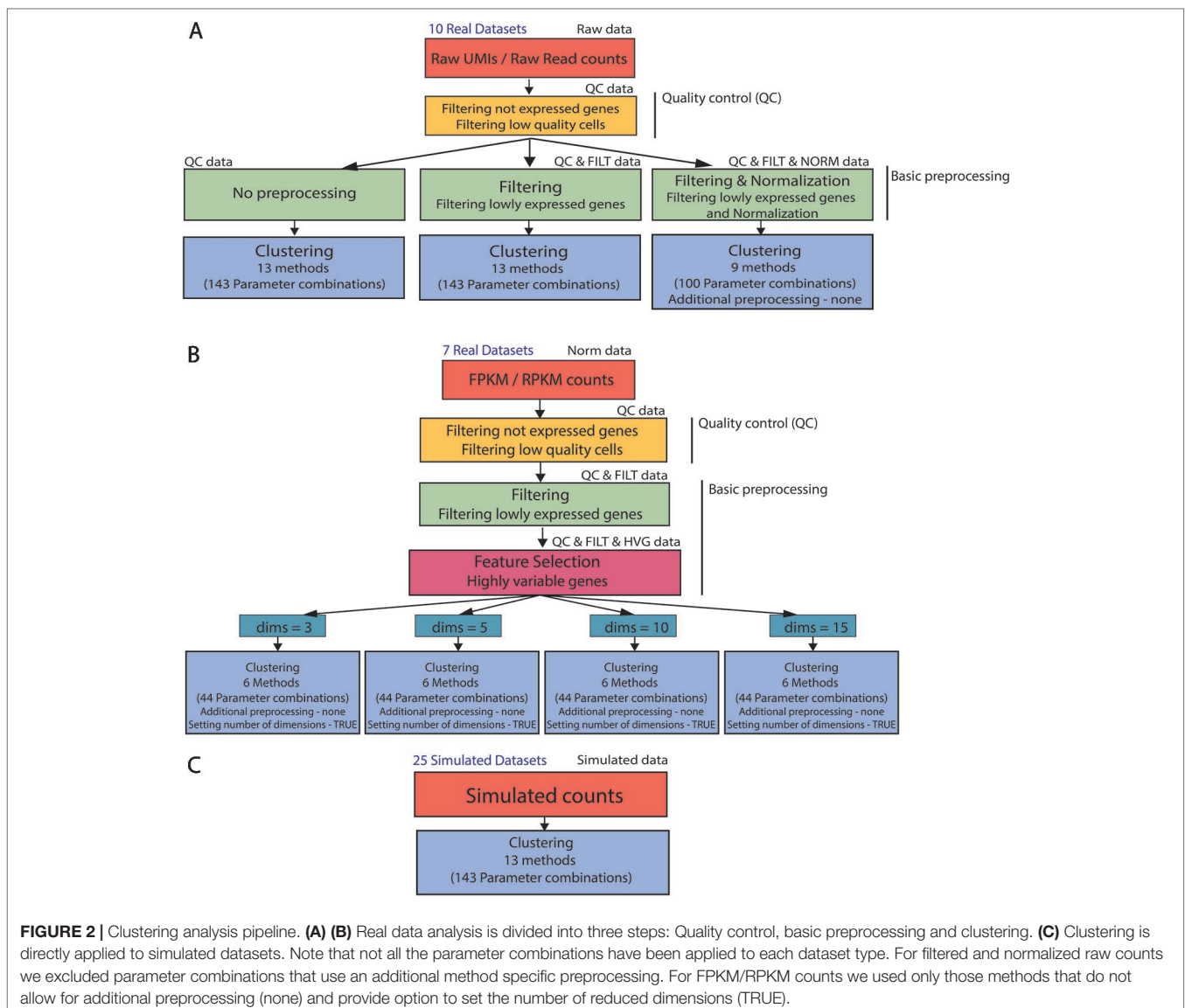
## Analysis Pipeline

In order to analyze real and simulated data, we used the procedure illustrated in **Figure 2**. First, all 17 real datasets (Raw UMI/Raw read counts and FPKM/RPKM counts) underwent the same quality control by filtering not expressed genes and low-quality cells (see **Figures 2A, B**) to remove potential issues from further analysis.

For the raw datasets, we considered three types of basic preprocessing before applying the specific clustering methods (**Figure 2A**). After the basic preprocessing, the clustering methods were applied with specific combinations of the parameters. Note that only a subset of methods (and combination of parameters) can be considered for filtered and normalized counts.

The FPKM/RPKM counts underwent a different basic preprocessing step (see **Figure 2B**) and were then directly clustered. To investigate the influence of the choice in the number of retained dimensions on methods performance, we considered only those methods and those combinations of parameters that allowed us to set the number of reduced dimensions.

In contrast to real data, simulated counts were directly used for clustering (see **Figure 2C**) where all methods and combination of parameters have been considered in the evaluation.



More details about data quality control, basic preprocessing of Raw and FPKM/RPKM counts, methods and parameter settings are described in the next sections.

## Quality Control of Real Datasets

All real datasets underwent an identical quality control step using the scater package (McCarthy et al., 2017). Firstly, we removed features with duplicated gene names and/or not expressed across all the cells as they do not include any useful information. Then, we performed quality control on the cells excluding those with the total number of expressed genes and the total sum of counts more than 3 median absolute deviations below the median across the genes [as suggested in scater documentation (McCarthy and Lun, 2019)]. Cells with the low amount of expressed genes and few counts are likely to be stressed or broken and thus should be removed from the analysis. The resulting dimensions of real datasets before and after quality control are given in **Supplementary Tables 1 and 2**.

## Basic Preprocessing of Real Datasets

After quality control, we applied a basic preprocessing step that mimics some of the most commonly used procedures typically applied before clustering scRNAseq data (McCarthy et al., 2017). In the case of Raw UMIs and Raw read counts, we considered three independent types of basic preprocessing: no preprocessing, filtering, filtering and normalization (see **Figure 2A**). Clearly, in the first case, no further operations were performed on the raw counts. In the second case, we used scater to remove lowly expressed genes that are genes with average expression count (adjusted by library size) equal to 0, where for the library size we mean total sum of the counts per cell. Note that this filtering step did not affect some of the datasets including Baron2016\_m, Klein2015, Zeisel2015, and Romanov2016 (see **Supplementary Table 1**). In the third type, we first applied the filtering as described above, then we performed normalization. Both, Raw UMI counts and Raw read counts were normalized by scran package using deconvolution method. The deconvolution method normalizes data by cell-pooled size factors that account for dropout biases. More details about raw dataset dimensions before and after filtering are given in **Supplementary Table 1**. For illustrative purpose, **Supplementary Figure 1** reports one realization of the tSNE projections of the 10 raw datasets after quality control step that were colored by the corresponding cell group annotations. The inspection of the figure shows the heterogeneity of the datasets with respect to number of cells, number of cell groups and their separation.

In case of FPKM/RPKM counts, the basic preprocessing involved the same gene filtering as for the raw counts followed by high variable gene selection (HVG) (**Figure 2B**). To extract the most informative genes, we used Seurat package (Macosko et al., 2015) that defines most variable genes based on mean-variance dispersion. The dimensions of datasets before and after basic preprocessing are given in **Supplementary Table 2**. **Supplementary Figure 2** shows one realization of the tSNE projections (colored by the corresponding cell group annotations)

of the 7 FPKM/RPKM datasets after quality control and basic preprocessing step.

## Compared Methods and Modes of Usage

In this study, we evaluated 13 different methods aimed to identify cell populations from scRNAseq data. **Table 3** lists the methods that we have considered. For the sake of code compatibility and transparency, we restricted our choice to the methods implemented in the R programming language. Some of the methods have multiple releases and versions. In this evaluation, we only tested the releases with versions reported in **Table 3**. For the sake of completeness, we stress that recently some of the methods listed in the table underwent to a major update which could have partially improved their performance.

Most of the methods (i.e., all except DIMMSC and pcaReduce) considered in this study can be applied by setting different parameter combinations, thus providing potentially different results. Such combinations of parameters allow the user to tune different modes of usage, such as including or not an additional preprocessing step, including or not a dimension reduction procedure, using different criteria for choosing the suitable data dimension, applying different clustering algorithms within the same method, setting or estimating the number of clusters. **Table 4** shows a detailed series of parameters that the user can choose with possible parameter choices. Each row defines valid parameter settings for the specific method. Within the same row, the total number of combinations is given by the product of each possibility (the last column of **Table 4** summarizes the number of combinations). If the method has been reported more than once in the table (i.e., Linnorm and sscClust), it means that some of the parameters worked only with a subset of the settings (i.e., not in a full combinatorial way). By considering all possible combinations, we obtained 143 potential different modes of usage of the 13 tested methods.

As shown in **Table 4**, eight methods (corresponding to 43 parameter combinations) might incorporate an additional

**TABLE 3 |** List of methods compared in the benchmark.

Method	Version	Class of clustering technique	Publication
ascend	v0.9.0	Hierarchical	Senabouth et al. (2019)
CIDR	v0.1.5	Hierarchical	Lin et al. (2017)
DIMMSC	v0.2.1	Model-based	Sun et al. (2018)
Linnorm	v2.6.1	Partitioning	Yip et al. (2017)
monocle3	v2.99.2	Multiple choices	Qiu et al. (2017)
pcaReduce	v1.0	Hierarchical	Zurauskiene and Yau (2016)
RaceID3	v0.1.3	Multiple choices	Herman and Grün (2018)
SC3	v1.10.1	Partitioning	Kiselev et al. (2017)
Seurat	v2.3.4	Graph-based	Macosko et al. (2015)
SIMLR	v1.8.1	Partitioning	Wang et al. (2017)
sincell	v1.14.1	Multiple choices	Julia et al. (2015)
sscClust	v0.1.0	Multiple choices	Ren et al. (2019)
TSCAN	v1.20.0	Model-based	Ji and Ji (2016)

*Versions of the R packages (methods) compared in this study. Methods are based on various clustering techniques that can be categorized based on the cluster-model. Multiple choices indicate that the method allows to cluster cells with more than one clustering technique.*



**TABLE 4 |** Valid configurations in the parameter settings for each method.

Method	Additional preprocessing	Dimension reduction	Setting number of dimensions	Clustering technique	Number of clusters	Combinations
ascend	method specific	internal	TRUE/internal	fixed	estimate	2
CIDR	none	internal	TRUE/internal	fixed	set/estimate	4
DIMMSC	none	none	FALSE	fixed	set	1
Linnorm	none/method specific	tSNE/PCA	TRUE/internal	fixed	set/estimate	16
Linnorm	none/method specific	none	FALSE	hclust	set	2
monocle3	none/method specific	tSNE/UMAP	TRUE/internal	densityPeak/louvain	estimate	16
pcaReduce	none	internal	internal	fixed	set	1
RaceID3	method specific	PCA	TRUE/internal	k-medoids/k-means/hclust	set/estimate	12
SC3	none/method specific	internal	internal	fixed	set/estimate	4
Seurat	method specific	PCA/ICA	TRUE/internal	fixed	estimate	4
SIMLR	none/method specific	internal	TRUE/internal	fixed	set/estimate	8
sincell	none	PCA/ICA/tSNE/ classical-MDS/ nonmetric-MDS	TRUE/internal	max.distance/percent/ knn/k-medoids/ward.D	estimate	50
sscClust	none	iCor	internal	k-means/ADPclust/hclust	set/estimate	6
sscClust	none	iCor	internal	SNN	estimate	1
sscClust	none	PCA	TRUE/internal	k-means/ADPclust/hclust	set/estimate	12
sscClust	none	PCA	TRUE/internal	SNN	estimate	2
TSCAN	method specific	internal	internal	fixed	set/estimate	2

We reported a set of parameters that users can tune in the method such as the additional preprocessing, the dimension reduction strategy, the number of dimensions, the clustering technique and the number of clusters to obtain. In particular, for the key additional processing: none – no additional preprocessing is applied, method specific – an additional preprocessing is applied prior clustering (filtering and/or normalization); for dimension reduction: internal – an internal dimension reduction is applied, none – the method works in the original domain, PCA, tSNE, ICA, iCor or others listed by names – the user can choose a specific method to reduce the dimensionality; for number of dimensions: TRUE or FALSE – method allows or doesn't allow for setting number of reduced dimensions, internal – method use an internal value for the number of dimensions; for clustering technique: fixed – method uses only one clustering technique, otherwise the user can choose among few options that are listed by name; for number of clusters: set or estimate – method allows to set or estimate number of clusters.

preprocessing step (herein, denoted method specific), five methods do not have any specific step (herein denoted none). Out of the 8 methods that include the additional preprocessing step, four methods allow the user to decide it to apply or not (both settings available). Methods differ also in the dimension reduction step either by providing only an internal procedure to reduce dimensions (six methods, herein denoted internal) or allowing for multiple choices for this purpose (five methods, herein denoted with the name of the specific procedures the user can choose, PCA, tSNE, ICA, etc). Note that two methods, DIMMSC and Linnorm have to or can, respectively, work directly in the high-dimensional space (setting herein denoted with none) and one method RaceID3 uses PCA dimension reduction which has been not considered as an internal technique (for more details see methods description in Supplementary Materials). Within all 12 methods that incorporate the dimension reduction step, an internal procedure can be used for selecting the number of reduced dimensions (herein denoted internal). Nine algorithms (63 combinations) also allows to manually set the number of dimensions (herein denoted with TRUE). Those with both options give to the user the possibility of either choosing the dimension or using the internal procedure. In this regard, the setting FALSE is related to methods that do not perform dimension reduction.

Methods can be also customized by the clustering techniques they apply. Some of them are based on a fixed clustering technique (herein denoted fixed), others propose multiple choices in this step (herein denoted with the name of the specific technique the user can choose, k-means, hclust, etc). The group of methods with multiple clustering options include: monocle3 that offers two types of clustering techniques, RaceID3 that utilizes two

partitioning algorithms and a hierarchical clustering algorithm, sincell and sscClust which provide more clustering options. Depending on the clustering technique, methods either require to set the number of clusters by the user (36 combinations, herein denoted set) or provide an internal functionality to estimate it (107 combinations, herein denoted estimate). For more details about specifications, see methods descriptions in **Supplementary Materials**.

Finally, we stress that all 13 methods (with all 143 combinations of parameters) can be applied to non-preprocessed or filtered Raw counts as well as simulated datasets (see **Figure 2**). To avoid performing method-specific normalization on already normalized data, only methods for which the additional preprocessing step can be set to none were used on filtered and normalized Raw counts (i.e., 9 methods with 100 combinations of parameters) (**Figure 2A**) or FPKM/RPKM counts. In addition, according to **Figure 2**, when using normalized FPKM/RPKM counts, we reduced the number of methods and parameter combinations to those which perform dimension reduction step before clustering, and allow setting number of reduced dimensions in that step. In this way, we used a subset of 6 methods and 44 combinations of parameters to be applied on FPKM/RPKM counts (**Figure 2B**).

## Evaluation Metrics

To quantify the agreement between the partition obtained from the considered method and the true partition, we used a well-known and widely used measure, the Adjusted Rand Index (ARI), implemented in the R package mclust (Scrucca et al., 2016). The



values of the ARI range can be negative if the agreement of the partitions is worse than the agreement expected by chance, or between 0 and 1 for clustering better than chance. The exact formulation of the ARI index can be found in (Lawrence and Phipps, 1985).

To evaluate the accuracy of methods in estimating the correct number of clusters, we used symmetric log-modulus transformation defined as follows:

$$L(x) = \text{sign}(x) * \log_{10}(|x| + 1) \quad (1)$$

where  $x$  is the difference between the estimated number of clusters and the true number of cell populations in a given dataset. The positive values of log-modulus transformation mean that the number of estimated clusters was higher than the number of true cell populations. Negative values indicate that methods underestimate the number of clusters whereas zero values denote the equality between the number of estimated clusters and the number of true cell populations.

To identify significant differences in methods performance (ARI Index) when applied after different basic preprocessing types, we used hypothesis testing procedures implemented in stats R package (Hollander and Wolfe, 1973). The Kruskal-Wallis rank sum test was used to assess the difference in methods performance as we vary the basic preprocessing (among QC, QC & FILT, QC & FILT & NORM). The Wilcoxon signed-rank test was used to infer the differences in accuracy with respect to two data basic preprocessing types (QC, QC & FILT). In each context, we computed the Benjamini-Hochberg adjusted p-values (Benjamini and Hochberg, 1995) to correct for multiple comparisons.

Finally, to measure the computational time required by each method to complete its task, we used *Sys.time* function from R that allows reporting time when the method starts and finishes the script. The difference between those time points constituted the computational time of the method in running dataset analysis. Note that computational times have been reported in the unit of minutes followed by  $\log(t+1)$  transformation, where  $t$  is the running time in minutes, and include all the steps that the method needs to cluster a dataset (except data basic preprocessing) together with the loading of the required packages and package dependencies.

## Implementation

This clustering benchmark study was implemented in the R programming language and scripts necessary for the reproducibility were deposited at the time of publication on the GitHub page: [https://github.com/mkrzak/Benchmarking\\_Clustering\\_Methods\\_scRNAseq](https://github.com/mkrzak/Benchmarking_Clustering_Methods_scRNAseq). The repository stores codes for data import, processing, and analysis as well as the information about system requirements and packages to be installed. When performing the analysis, additional HTML reports are produced with a detailed description of data analysis steps. Note that apart from the required methods, the analysis scripts call for other R packages used in plotting and managing R objects. The scripts have been tested on R version 3.5.1 and machine with

specifications—Intel Core i7, 4.00 GHz × 8 and 24 GB RAM which are the minimum system requirements for the analysis.

Moreover, for the sake of completeness and to ensure the reproducibility of our study, we deposited the real and simulated datasets on the following GitHub pages: [https://github.com/DataStorageForReproducibility/Real\\_data\\_for\\_benchmark\\_reproducibility](https://github.com/DataStorageForReproducibility/Real_data_for_benchmark_reproducibility) and [https://github.com/DataStorageForReproducibility/Simulated\\_data\\_for\\_benchmark\\_reproducibility](https://github.com/DataStorageForReproducibility/Simulated_data_for_benchmark_reproducibility). Both directories include RData files as SingleCellExperiment class objects that store the count matrices and the corresponding cell group annotations.

In the clustering benchmark, we set the seed for generating pseudo-random numbers globally and applied it to the execution of any method in order to assure the stability of the solutions and reproducibility of the results. Note that, since the scRNAseq R packages we evaluated are often under continuous development, other version of the methods (R packages) than those reported in **Table 3**, might output slightly different results.

## RESULTS

Results are organized as follows. We first illustrate the performance of the evaluated methods on the 10 raw datasets, then on the 7 normalized FPKM/RPKM counts. Finally, we finish the summary of the main findings obtained on the simulated datasets in the 3 setups described in **Figure 1**.

Within this paper, methods/parameter combinations are referred as string obtained as a concatenation of keys separated by underscores. The concatenation takes the name of the method, the type of additional preprocessing, the dimension reduction technique, the setting of the number of dimensions, the clustering technique and the number of clusters. Each of these keys can take the values reported in **Table 4**.

### Methods Performance on Raw UMI and Raw Read Counts

As mentioned, we independently applied all 13 methods (corresponding to 143 parameter combinations) to the 10 raw counts datasets after using two basic preprocessing types (QC, QC & FILT). Then, we applied only 9 methods (corresponding to a subset of 100 parameter combinations) to the same datasets after applying quality control, filtering and normalization (see the scheme illustrated in **Figure 2**). In the latter case, the 9 methods are those that allow the user to choose none as additional preprocessing to avoid renormalization of already normalized counts (see **Table 4**). To compare the methods across the basic preprocessing procedures, we first show the results corresponding to the combinations that were applied to all three basic preprocessing procedures, then the remaining methods/combinations applied only to QC and QC & FILT data.

Note that some of the methods/parameters combinations failed to cluster some datasets (such cases are marked in grey in **Supplementary Figures 3 and 4**) due to the errors occurred during their execution. The most frequent error messages were reported in **Supplementary Table 3**, for Data type = “Raw counts”. In particular, SIMLR, DIMMSC and Linnorm

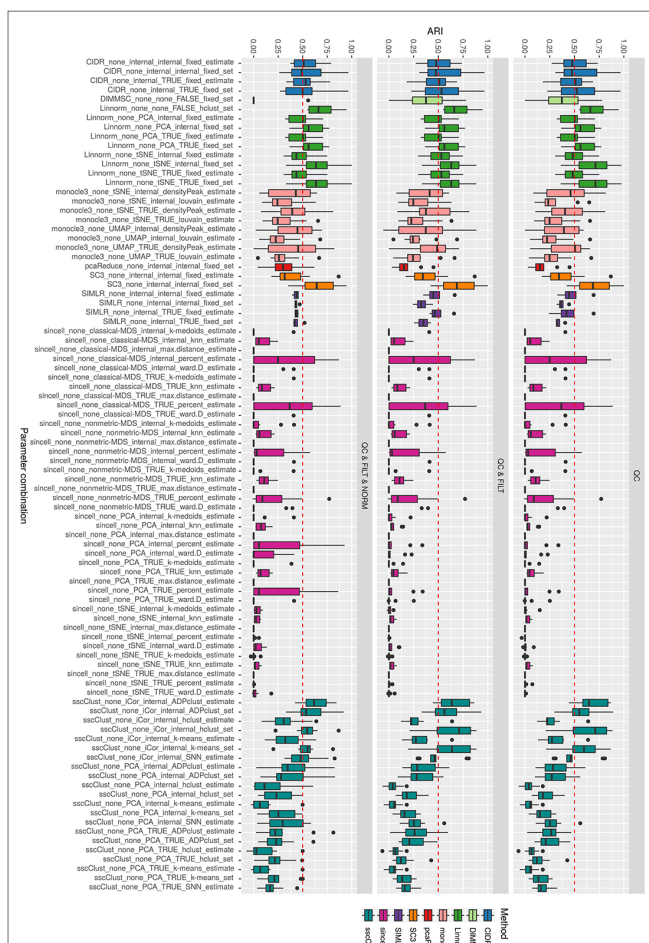
encountered failures in a limited number of cases, therefore we did not consider such datasets in the evaluation of the methods. By contrast, sincell (when ICA was chosen for dimension reduction) reported a significant number of failures, therefore we did not consider such combinations of parameters in the evaluation of sincell. Note that this will limit the overall number of parameter combinations from 143 to 133 (90 combinations applied after all three types of basic preprocessing, 43 applied to QC and QC & FILT data, only).

## Overall Accuracy

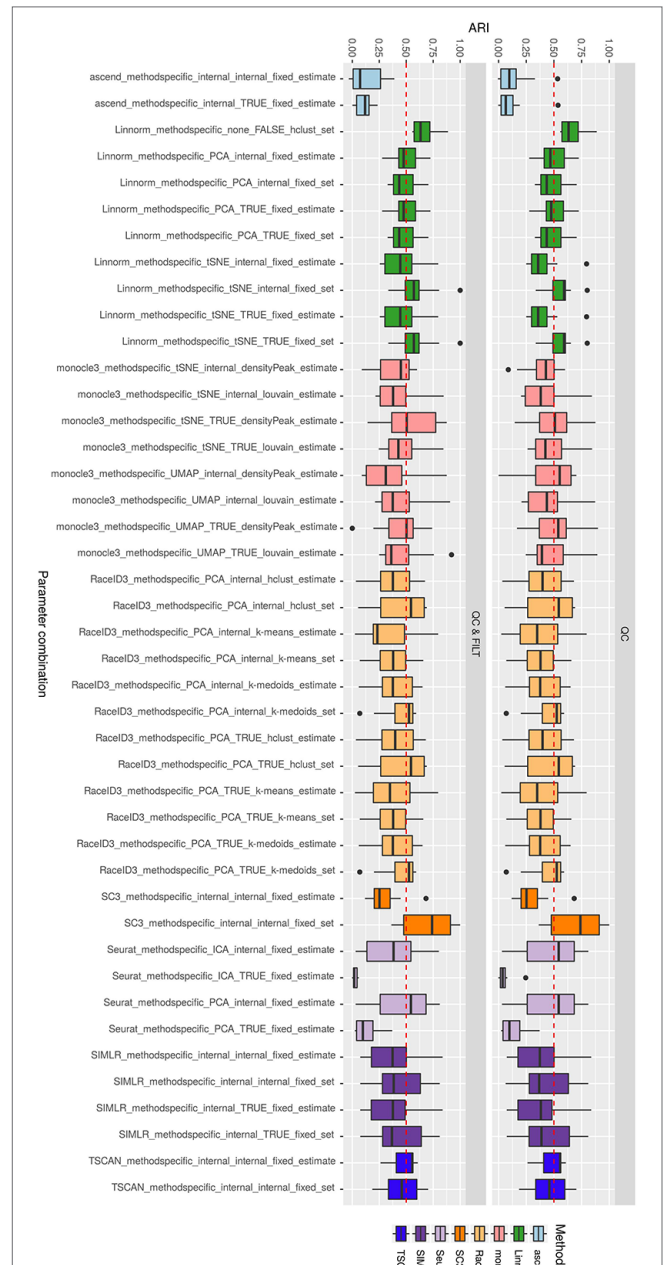
**Figure 3** shows the performance of the 9 methods (90 parameter combinations out of 100) in terms of ARI evaluated across all 10 raw datasets and organized with respect to the type of basic preprocessing. Analogously, **Figure 4** shows the same results corresponding to the remaining 8 methods (43 parameters combinations) independently applied after two basic

preprocessing types. To evaluate the overall accuracy, we first inspected the results regardless of the type of basic preprocessing.

From **Figures 3** and **4**, we can observe that, most of the methods/parameter combinations report a great variability in their performance across the different datasets which proves no all-time winner across the entire set of cases we have analyzed. Some of the methods still performed relatively well (i.e., with most of the results



**FIGURE 3 |** Overall accuracy of methods applied to Raw counts. ARI accuracy for 9 methods with 90 parameter combinations out of 100, independently applied to the 10 raw datasets after the three basic preprocessing types (QC, QC & FILT, QC & FILT & NORM). Box colors distinguish the different methods, although applied with different parameter combinations. Superimposed as reference, a red dashed line at ARI = 0.5.



**FIGURE 4 |** Overall accuracy of methods applied to Raw counts. ARI accuracy for remaining 8 methods with 43 parameter combinations, independently applied to the 10 raw datasets after two basic preprocessing types (QC, QC & FILT). Box colors distinguish the different methods, although applied with different parameter combinations. Superimposed as reference, a red dashed line at ARI = 0.5.

above  $ARI = 0.5$ ) regardless the preprocessing type. This group includes CIDR, Linnorm (with some combinations of parameters), SC3 (when set is chosen in number of clusters), some combinations of sscClust (i.e., when iCor is used for dimension reduction) and TSCAN. On the other hand, few other methods were reporting very poor performance. For example, one of the poorest performance was observed in sincell (with many parameter combinations), ascend, DIMMSC, pcaReduce and Seurat (only when non-internal is chosen for the number of reduced dimensions). Although sincell performed generally poor, the method also showed good performance for few datasets (see, the results over individual datasets showed in **Supplementary Figure 3**).

The analysis of **Figures 3** and **4** also shows that the performance of some methods strongly depends on the particular choice of the parameter settings, i.e. sscClust, Linnorm or Seurat being those whose performance strongly rely on that option. We found such result partially ignored in previous studies, therefore we will investigate it in more detail in *Effect of Parameters Settings on Accuracy*.

### Accuracy in Estimating the Number of Clusters

In order to evaluate the accuracy of a method in estimating the correct number of populations, we used log-modulus transformation in Eq. 1, and we limited the analysis to the 107 methods/parameter combinations that allow setting the option estimate for choosing the number of clusters (see **Table 4**).

**Figures 5** and **6** show the results, respectively for the 69 methods/parameters combinations applied after all three types of preprocessing procedures (i.e., we excluded 10 combination of sincell that reported frequent failures), and for the remaining 28 methods/parameter combinations applied after two basic preprocessing steps.

We observed that most of the methods/parameter combinations either under or overestimated the number of clusters often in a systematic way. In particular, boxes below and above the dashed lines represent parameter combinations which under or overestimated the number of clusters. There are also methods, such as CIDR, some combinations of Linnorm, RaceID3 and TSCAN, that often provide less biased estimates. We also observed that the estimates strongly depend on the specific clustering technique used, as for monocle3, sincell and RaceID3 method, or dimension reduction applied, as for Linnorm. The group of methods that underestimated number of clusters includes monocle3 (when densityPeak is used for clustering), SIMLR method, sincell (with k-medoids and ward.D chosen as clustering techniques), all combinations of sscClust except SNN and RaceID3 (when k-means was applied). A special case of overestimating the number of clusters method was observed with sincell where a large number of cluster was often returned. For example, sincell used with max. distance technique always returned a number of clusters equal to the number of cells in the dataset whereas in combination with knn it also returned a very large number of clusters.

### Effect of Data Basic Preprocessing on Accuracy

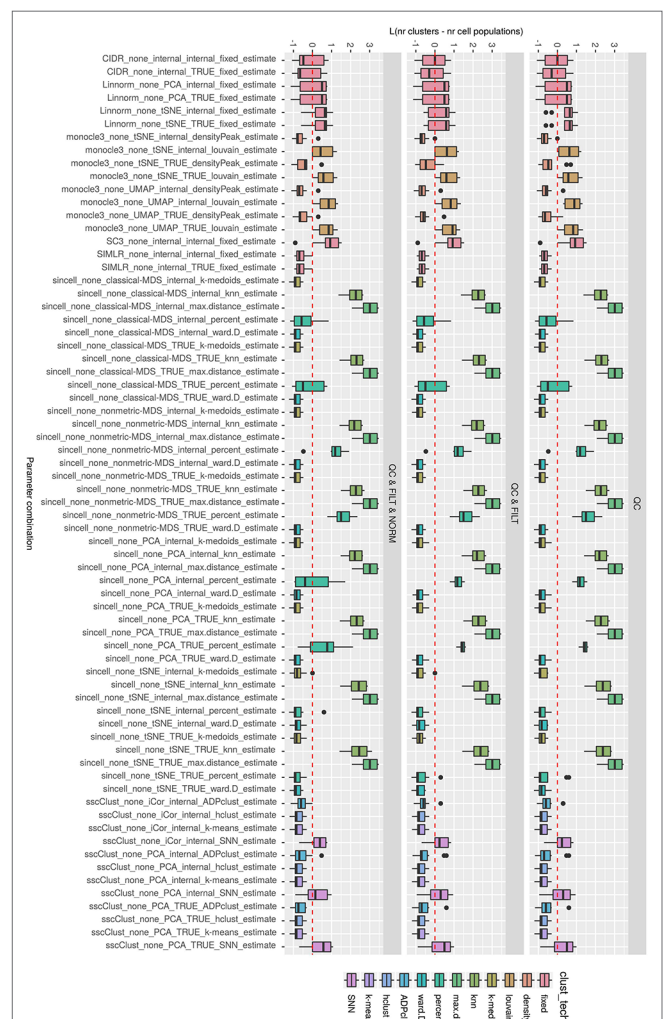
We found that most of the methods performed similarly when changing the preprocessing procedures (see **Figures 3** and **4**), although **Supplementary Figures 3** and **4** showed that some of them (i.e., Linnorm, monocle3, sincell and sscClust) present

slight variability in the performance when data underwent to different preprocessing. However, Kruskal-Wallis rank sum test did not identify any significant difference in the performance of the methods with respect to the three types of basic preprocessing (QC, QC & FILT, QC & FILT & NORM) and Wilcoxon signed-rank test did not identify any significant difference associated to the two types of basic preprocessing (QC, QC & FILT).

### Effect of Parameter Settings on Accuracy

As mentioned above, **Figures 3** and **4** clearly shows that the performance of several methods depends much more on the choice of parameters than on the type of basic preprocessing.

To better investigate this, we computed the PCA of the ARI matrix obtained using the 133 methods/combination as variables and the 10 raw datasets as samples. **Figure 7** shows the results



**FIGURE 5** | Estimation of the number of clusters for methods applied to Raw counts. Boxplots of  $L$  in Eq. 1 for the subset of methods (i.e., 69 parameter combinations) that allows to estimate the number of clusters (and with none preprocessing). Superimposed as a reference, a red dashed line at  $L = 0$ . Parameter combinations with difference below or above 0 resulted into under or overestimation of the number of clusters, respectively.



when the clustering methods were applied to QC & FILT preprocessed data (the figures after the other preprocessing types are very similar, not shown for brevity). Each point depicted in the PCA space represents a particular methods/parameter combination. Therefore, points that are close in the PCA space have similar performance across the 10 datasets. From **Supplementary Figure 5** we can see that the first component is strongly positively correlated with the performance, therefore methods located on the right side of the figure tends to have better performance than those located on the left side, while the second component is not significantly correlated with the ARI. Each panel of **Figure 7** represents the same PCA projection colored by the methods and shaped by one of the parameters of interest. The effect of parameters changes in the performance of

a given method is represented through the spread of the points in the same color. Note that DIMMSC and pcaReduce have only one valid parameter combination thus we do not discuss them in this section, although they are depicted in the figure.

Overall, **Figure 7** confirms the poor performance of sincell and the good performance of SC3, CIDR, TSCAN, and some combinations of Linnorm, as well as the strong impact of parameters setting for many methods (i.e., sscClust, Linnorm, Seurat, SIMLR).

In particular, the analysis of **Figure 7A** shows that the performance strongly depends on whether the number of clusters is estimated or not. Not surprisingly, when using the true number of clusters (parameter set) the performance is better for most of the methods compared to when estimating it (parameter estimate). However, there are few methods that report good overall performance also when the number of clusters is estimated (see for example, CIDR, monocle3 and sscClust).

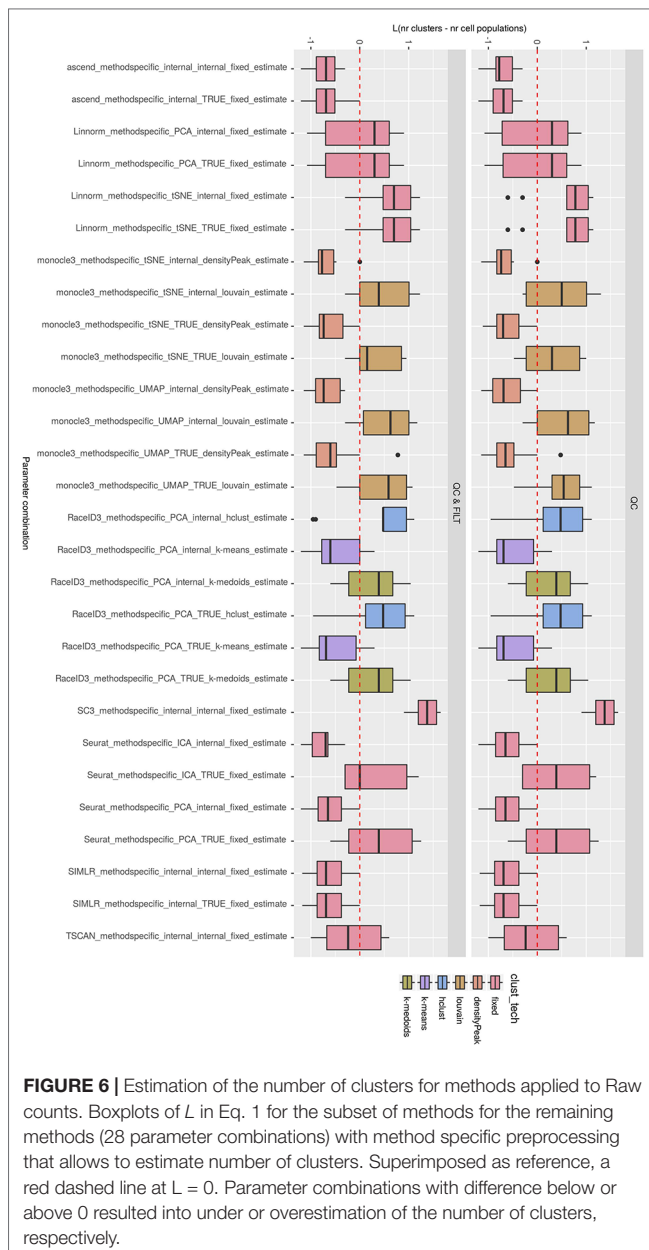
In the same spirit, **Figure 7B** illustrates the effect of an additional preprocessing (that can be either method specific or none) on the methods performance. The figure does not indicate any global difference, but still pointed-out some methods specific variability (i.e. SIMLR showed significantly improved accuracy after such step).

We also superimposed other features, such as dimension reduction or clustering techniques (not shown for brevity). Since such parameters can assume multiple values, the figures do not allow to identify any suggestion that works well for all methods. However, such analysis allowed us to recognize i.e., sscClust with iCor and Seurat with internal number of reduced dimensions, as one of the good performing combinations.

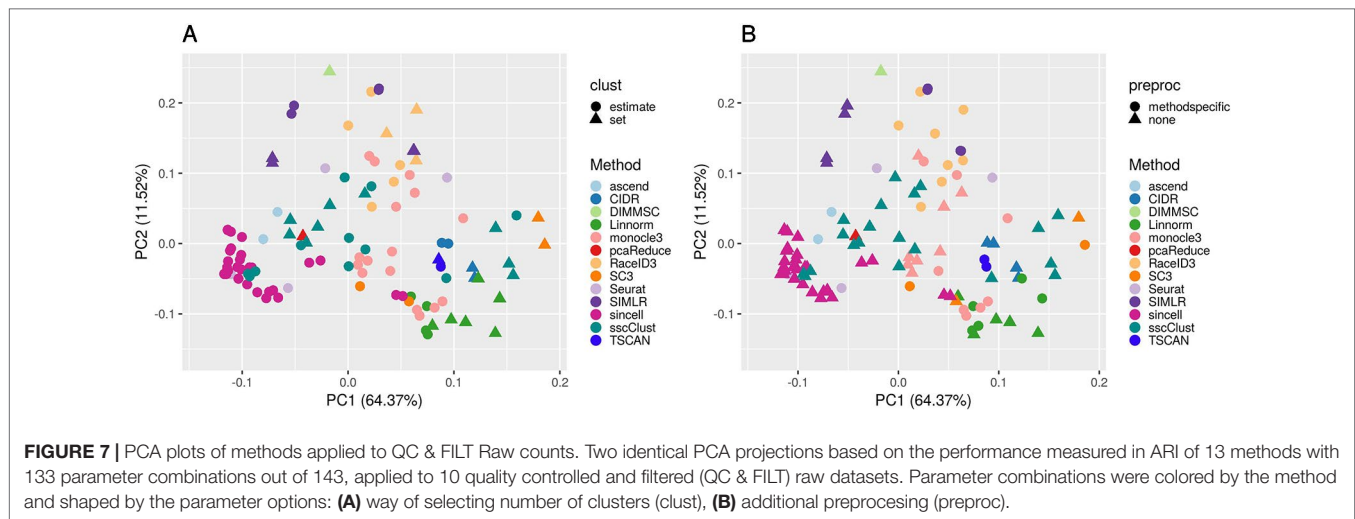
## Computational Time

We compared run times of the methods across all 10 raw datasets in order to assess their scalability and identify potential issues related to a specific dataset.

**Figure 8** reports execution time in minutes on a log plus one scale for the methods applied to QC & FILT preprocessed datasets. As a reference, we superimposed on the figure dashed lines at 1, 10, 60 min, and 10 hours. Overall, computational times varied from a few seconds to tens of minutes or till several hours (at least for some datasets). We distinguish methods that were consistently fast (showing good scalability), methods requiring longer but still reasonable run time with increased data size (showing limited scalability) and methods requiring significant execution time at least in some cases (showing either poor scalability or problems related to the analysis of a specific dataset). ascend, CIDR, monocle3, pcaReduce, RaceID3 (with non-internal number of dimensions), Seurat (with PCA dimension reduction), sincell, sscClust and TSCAN were among the fastest and across the analyzed datasets. Therefore, they were assigned to the first group (with average run time below 2 minutes and maximum time of about 10 minutes). Linnorm and SC3 were assigned to the second group (with average run time about 5 minutes and maximum time between 20 minutes and an hour). Other methods such as DIMMSC, SIMLR, RaceID3 (with internal number of dimensions), Seurat (with ICA dimension reduction) were among the longest, therefore assigned to the







third group (with average run time between 10 minutes and about two hours and maximum time between an hour and more than 10 hours). In the worst case, RaceID3 took about 12 hours before completing the clustering task.

## Methods Performance on FPKM/RPKM Counts

We used 7 FPKM/RPKM datasets to evaluate the performance of the methods/parameter combinations considered in this study with respect to the number of reduced dimensions when using different dimension reduction techniques. Since FPKM/RPKM datasets consist of already normalized counts we limited the study to those methods/parameter combinations that do not use “method specific” as additional preprocessing and that also allow setting the number of reduced dimensions. In total, we tested 44 methods/parameter combinations on each of the four dimensions: 3, 5, 10, and 15.

As in the previous case, we note that some of the methods/parameters combinations failed to cluster some of the datasets (see grey boxes in **Supplementary Figure 6**) due to technical errors reported in **Supplementary Table 3**, for Data type = FPKM/RPKM counts. Note that three of the methods, Linnorm, monocle3 and sincell encountered a significant number of failures with the same error message when used with more than 3 dimensions. We did not consider such cases in further evaluation limiting the overall number of combinations from 44 to 33.

### Overall Accuracy

**Supplementary Figure 7** shows the performance of all 33 methods/parameter combinations applied to FPKM/RPKM datasets. Regardless of the number of dimensions, we can observe variability in the accuracy of the methods similar to what was reported for the raw counts. Most of the methods that were reporting good or poor accuracy on raw counts show similar good/poor performance also on the FPKM/RPKM datasets (as we could have predicted from the results obtained on the QC & FILT & NORM raw datasets). For example, CIDR and sscClust

(with some of the parameter combinations) are among the better-performing methods, whereas sincell with most of the combinations reports poor accuracy (although not in all cases). Additionally, we can also confirm that the performance of some methods depends on the choice of parameter settings.

We also observed a general tendency of the methods to perform poorly on datasets with a high number of cells (more than 1600) (see **Supplementary Figure 6**). Although this relationship was not clearly visible on the raw counts, it could be expected as a consequence of a larger complexity in the data not fully explained by the number of selected features and not fully captured using low dimensional projections.

Finally, we did not observe any systematic differences in the accuracy with respect to the number of reduced dimensions (see **Supplementary Figures 6 and 7**). Some of the methods are either robust to the varying number of dimensions or they do not show any clear preference when using one or another setting. This suggests that data complexity cannot be easily explained by a certain parameter and the performance of the methods are often data specific.

### Accuracy in Estimating Number of Clusters

**Supplementary Figure 8** shows the estimated number of clusters compared with the true one (as computed using Eq. 1) for all methods/combinations that allow the users to estimate such value. We observed a similar tendency in the estimates reported for raw counts. For example, monocle3 (with densityPeak clustering), SIMLR, sincell (with k-medoids and ward.D techniques) or sscClust (all except SNN) tend to underestimate the number of clusters whereas the rest of the combinations of sincell clearly overestimate that value. Moreover, CIDR often provides a less bias estimates that result in a better accuracy (alike on the raw counts).

### Computational Time

**Supplementary Figure 9** reports the running time evaluated for all methods/parameter combinations (for dimension = 3). First, we observe that, since FPKM/RPKM counts underwent to

Finally, note that some of the combinations evaluated on the raw counts, such as RaceID3 or Seurat, were not considered in the FPKM/RPKM evaluation as they do not allow to set none in the additional preprocessing.

Synthetic datasets were used to test the performance of all 143 methods/parameter combinations. We followed three simulation setups in order to simulate the counts (see **Figure 1**) and we repeated the simulation 5 times, each with a different selection of the random seed. Simulation setups mimic different characteristics of scRNAseq datasets i.e. in terms of dimensionality, group structure or levels of noise. In theory, all simulated datasets constitute a different level of complexity for the clustering task.

In the next sections, we will describe the performance of the methods according to the three simulation setups. Note that the overall performance of the methods on the synthetic datasets is much higher than in the real data. This can be related to the fact that simulation models may not always reflect all types of noise present in the real case and thus the clustering task can be less challenging. Despite that, synthetic datasets allowed us to confirm some of the previously identified trends and to recognize the potential limits of the methods.

Simulation Setup 1 has been used to access the performance of the methods depending on three factors: the number of cells present in the dataset, the number of cell groups and their balance in size. **Figure 9** and **Supplementary Figure 11** show the accuracy of the methods for balanced and unbalanced group design, respectively. **Supplementary Figures 12–14** give more details about balanced group design and correspond to the performance on datasets with 4, 8, and 16 number of cell groups, respectively.

On the synthetic datasets, the well performing methods included CIDR, Linnorm, SC3 and some combinations of sscClust (see **Figure 9** and **Supplementary Figure 11**), same as for the real datasets. Similarly, we could confirm the poor performance of methods such as Seurat with an imposed number of dimensions or sincell with tSNE dimension reduction. Additionally, on the synthetic data we observed high accuracy of the DIMMSC method, Seurat with internal number of dimensions and some combinations of monocle3, RaceID3 and SIMLR.

Horizontal box plot showing Log(run time in minutes) for various parameters across different methods. The y-axis lists parameters grouped by method (ascend, CIDR, Linorm, monocl3, RaceID3, SC3, Seurat, SIMLR, sincell, sscClust, TSCAN). The x-axis shows Log(run time in minutes) from 0 to 6. A legend at the bottom identifies methods by color: ascend (blue), CIDR (green), DIMMSC (red), Linorm (orange), monocl3 (purple), RaceID3 (brown), SC3 (pink), Seurat (grey), SIMLR (dark blue), sincell (light blue), sscClust (dark green), TSCAN (black).

Frontiers in Genetics | www.frontiersin.org

and **Supplementary Figure 11**). The only clear exception was SIMLR with several combinations that include cluster number estimation (denoted estimate) which failed on datasets with 5000 number of cells (see the error messages in **Supplementary Table 3**). We observed that many methods were affected by the growing number of simulated cell groups (from 4 to 16 cell groups). In particular, see the methods: CIDR, DIMMSC, Linnorm, SC3, SIMLR, sincell, sscClust, and TSCAN across **Supplementary Figures 12–14**. pcaReduce worked similarly across all three factors (number of cells, number of cell groups, group balance)

(see **Figure 9** and **Supplementary Figure 11**). Seurat accuracy, same for the real datasets, strongly dependent on the number of reduced dimensions (denoted as TRUE/internal).

## Performance on the Simulation Setup 2

In the simulation Setup 2, we varied the separability between the cell groups from lowly to highly separable. Lowly separable groups mean that some of the simulated populations could overlap in space being the most challenging to detect. Separability was controlled by de.prob parameter in the Splatter simulation function.

**Figure 10** shows that some of the methods as CIDR, DIMMSC, SC3, TSCAN, Seurat (with imposed number of dimensions), SIMLR (with estimated number of clusters) and many combinations of sincell behaved similarly and their performance was mostly affected on the datasets with the lowest separability between the cell groups. However, their accuracy was still high meaning in most of the cases ARI above 0.5. The methods that performed well across all the separability modes were some combinations of Linnorm or monocle3, Seurat with internal number of dimensions and SIMLR with set number of clusters. All those methods/parameter combinations provided high accuracy with ARI close to 1.

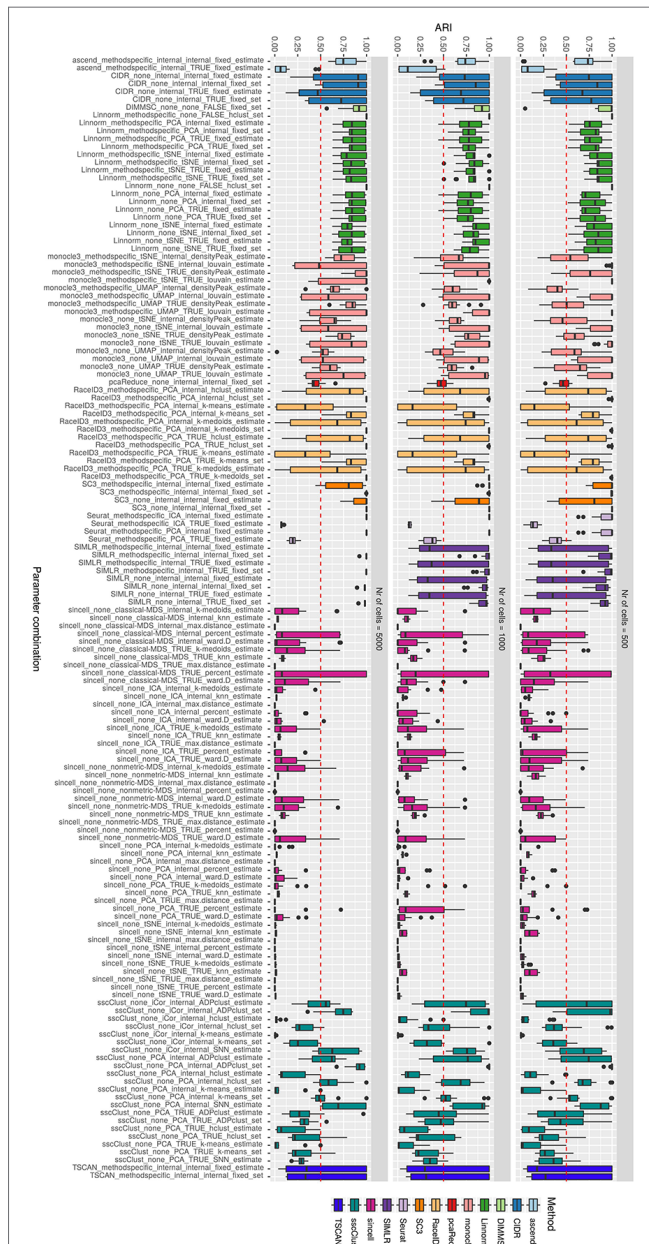
## Performance on the Simulation Setup 3

The third simulation setup was used to access the accuracy of the methods with respect to an increasing number of zero counts placed in the dataset. We simulated percentage of dropouts varying from 20% to 90% by manipulating dropout.mid parameter in the Splatter simulation function.

Overall, we noticed that most of the methods had low accuracy on the datasets with highest magnitude of missing values (dropout.mid = 6) (see **Figure 11**). Although this is an expected result, some of the methods/parameter combinations still performed well in this case (see i.e. monocle3, SC3 and sscClust). Interestingly, monocle3 and sscClust method performed poorly only in particular parameter combinations on the highest dropout rate. For the monocle3 method the bad performing combinations included additional method specific preprocessing and for the sscClust—iCor dimension reduction. Seurat depended highly on the number of dimensions used (either TRUE or internal) and pcaReduce seemed to work moderate across all four ranges of dropouts.

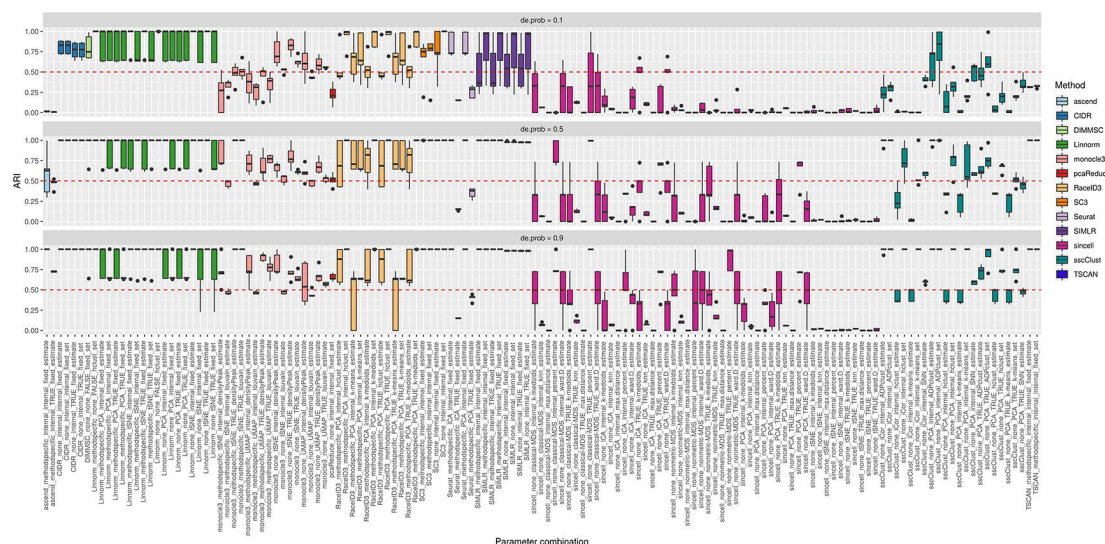
## Computational Time

Computational time for all parameter combinations applied to simulation Setup 1 datasets was reported in the **Supplementary Figure 15**. Some of the methods scaled in time all simulated datasets dimensions while others took longer on the largest datasets (with 5000 number of cells). Note that many of the trends observed here were previously mentioned in the real



**FIGURE 9 |** Overall accuracy of the methods on simulated datasets from Setup1 with balanced group sizes. Performance of 143 parameter combinations on Setup 1 simulated data. Selected results are across all runs.



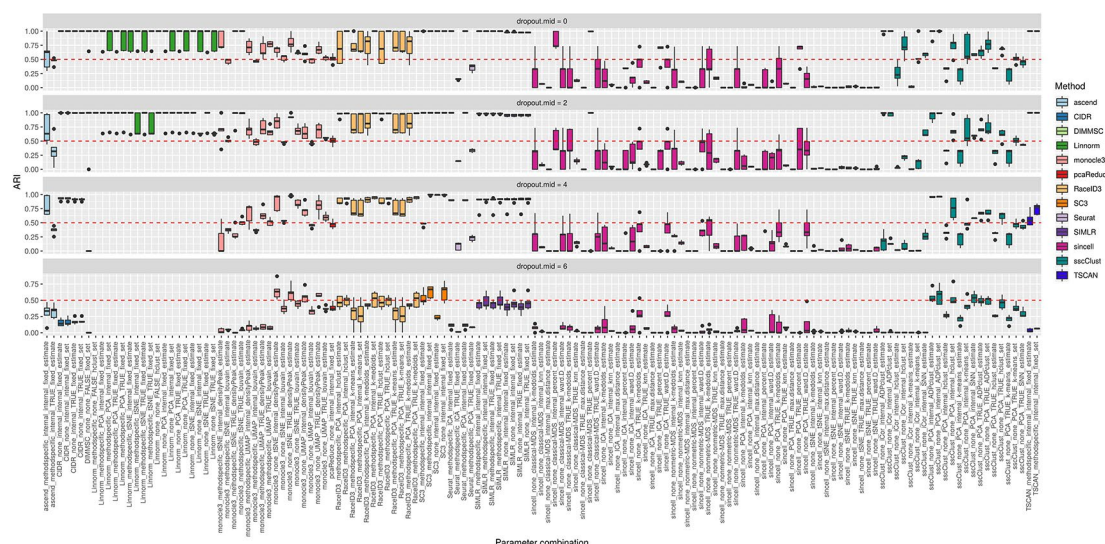


**FIGURE 10 |** Overall accuracy of the methods on simulated datasets from Setup 2. Performance of 143 parameter combinations on simulated data. Selected results are across all runs.

datasets analysis. The fastest group of methods across all datasets dimensions include: ascend, monocle3, pcaReduce, Seurat, many combinations of sincell, sscClust when PCA dimension reduction was used and TSCAN. Other methods like CIDR, DIMMSC, Linnorm with set number of clusters and some combinations of RaceID3 and sincell were still relatively fast in time running for few minutes on datasets with the highest number of cells. SC3, Linnorm (with estimated number of clusters), sincell (when nonmetric-MDS was used as dimensionality reduction) and rest of the combinations of RaceID3 or sscClust took about one hour when applied to the largest simulated datasets whereas SIMLR computational time exceed few hours in that case being the longest method among all.

## DISCUSSION

In this study, we evaluated the performance of several clustering methods on a wide range of real and simulated scRNAseq datasets. Such methods are distributed as open-source R packages and they constitute a significant part of the computational tools nowadays available for inferring the unknown composition of cell populations from scRNAseq data. Our comparison aimed to provide insight into the mode of usage for each of these packages depending on the structural assumptions we are willing to make. We compared the ability of the different packages to infer the unknown number of cell populations, the sensitivity of the methods across different datasets and their computational cost.



**FIGURE 11 |** Overall accuracy of the methods on simulated datasets from Setup 3. Performance of 143 parameter combinations on simulated data. Selected results are across all runs.



For each method we tested different parameter configurations, revealing the great impact of parameter setting on the performance of individual methods. In particular, we found that some of the methods performed relatively well across most of the datasets we have considered and with respect to the different choices of the parameter settings (i.e., CIDR, and several combinations of Linnorm, SC3 and sscClust), or often poorly, as, sincell (with many parameter settings) and ascend. Other methods, such as DIMMSC, monocle3, RaceID3, Seurat, SIMLR and TSCAN, can be placed in the middle in terms of overall performance across all datasets, despite the fact that on few datasets they could have reported good performance. However, we should consider that the field of clustering of scRNAseq data is rapidly evolving. Novel methods are continuously emerging and those that we have compared are undergoing to an extensive revision that might improve their performance. It is not easy to explain why certain methods work better than others, since they perform several steps before applying the clustering algorithm. However, one of the reasons is that some methods were originally developed to analyze scRNAseq data collected under specific protocols (i.e., consisting of datasets with a limited number of cells). Then, the novel challenges (in particular the increasing size and cell heterogeneity) provided by the rapidly evolving scRNAseq technology made them not any more competitive for the complex types of data that are emerging. Reasonably, methods have to be optimized with respect to a specific protocol or dataset size, rather than attempting to find methods that work well on a wide range of scRNAseq conditions. In fact, our study showed that no methods seem to emerge as performing better than others on all datasets. Additionally, our results also showed that there is still space for improving the overall performance of the available methods on large and complex datasets or providing novel and more accurate methods.

We have found that despite different basic preprocessing options, there is no global pre-processing strategy which improves significantly the performance of all methods (packages). Instead, we found that the performance of several methods strongly depends on their parameter settings: in Seurat when varying the number of input dimensions; in SIMLR when estimating the true number of cell groups; in sincell when varying the clustering techniques and in sscClust when changing the dimension reduction step. We believe that the impact of the choice of the method-specific parameters on its performance has been underestimated till now, while it turns out to be crucial when using these methods. Unfortunately, we did not identify a golden rule for choosing the parameters. However, depending on the methods used, we identified some better performing configurations: sscClust performed better with iCor as dimensionality reduction step; Seurat with the internal choice of the number of dimensions; Linnorm and SC3 with a set number of clusters (using the true number of cell populations). On the basis of our results, we suggest that users should be more aware of the different possibilities that several methods offer in terms of parameter choices and modes of usage. Moreover, we recommend them to always evaluate the robustness of their partition with respect to changes in the parameter settings. At the same time, method developers should give more attention in

better documenting all the possibilities that their methods can offer also testing their robustness with respect to changes in the settings. To this purpose, the benchmark pipeline developed for this study can be easily modified to offer an environment where other/novel methods can be evaluated.

We also observed that the poor performance of several methods/parameter combinations is often associated with a poor estimate of the number of clusters (see for instance estimation accuracy of monocle3, SIMLR or sincell). Although a rigorous assessment of the number of cell populations on real data dataset could be debated, our results show that several methods tend to significantly underestimate or overestimate the number of clusters, when compared to the true (usually unknown) cell populations. Therefore, we can say that the estimation of the number of hidden cell populations remains challenging in the scRNAseq data analysis and we hope that novel approaches will provide less biased estimates. Moreover, by comparing the performance of each method when the true number of clusters was imputed with those when it has been estimated from the data, it is possible to quantify the impact that a more accurate estimate of the number of cell populations can have on the overall accuracy.

The dataset dimension and complexity turns out to be clearly influential with respect to the running time of the methods and to the overall performance that the methods can achieve. In particular, SIMLR run time increased together with the sample size and was often the longest among other methods by several orders of magnitude (requiring up to several hours to analyze a given dataset compared to few seconds/minutes for the other methods). Similarly, scalability issues were observed in SC3, although to a less extent. In contrast, other methods/parameter combinations showed a good scalability in their computational time, as ascend, CIDR, monocle3, pcaReduce, RaceID3 (with non-internal number of dimensions), Seurat (with PCA dimension reduction), sincell, TSCAN or sscClust, limiting the computational time to few seconds/minutes. We want to stress that computational issues are becoming particularly important since modern technologies are now allowing to simultaneously sequence thousands or even tens of thousands of cells, thus it is expected that researchers will have to analyze much larger datasets. Hence, it will be important to provide novel methods that have good scalability properties either in terms of running time and/or computational resources required for their execution. This can be achieved either by designing methods with efficient algorithms and by better exploiting the parallel and high-performance computing in their implementation. From a technical point of view, we also observed frequent failures of some methods under particular cases. For example, SIMLR method failed on most of the simulated datasets with 5000 number of cells. We suspect that the method required large amounts of memory on the high-sample datasets than that available in our system. Other failures, like in monocle3, Linnorm and sincell on FPKM/RPKM datasets were related to the choices on the number of reduced dimensions. In fact, all of them encountered technical errors when used with tSNE dimension reduction and more than 3 number of dimensions. Additionally, Linnorm failed on raw and simulated datasets with a high percentage of dropouts (above 70% of zeros in the dataset) suggesting the low capacity of

the method to handle high rates of missing data. Such points are probably less relevant and could be solved with future releases of the methods.

Finally, it is also worth to mention that some of the methods, such as ascend, monocle3, SIMLR, sscClust and some combinations of Linnorm and sincell, showed variability in the clusterization despite the global setting of the seed. The fluctuations can be spotted by looking at the accuracy of methods on the identical datasets across three simulation setups (see results across **Supplementary Figure 12** and **Figures 10** and **11**) or by looking at the accuracy of the methods on datasets not affected by filtering (see **Supplementary Figures 3** and **4**). We notify that the results in such cases might not be easily reproducible. In the spirit of reproducible computational research, the user should be aware of such limits.

## CONCLUSIONS

Concurrently with technical improvements in single-cell RNA sequencing, there is a rapid growth in the development of new methods, in particular, those related to the identification of cellular populations. Newly developed methods differ considerably in their computational design, implemented algorithms and available steps giving the user a large number of options to select parameters and perform a cluster analysis on scRNAseq data. However, such possibilities are often hidden and not fully documented in the software code and their impact has to be better understood.

We are not aware of any comprehensive studies aiming to test various modes of usage of the available methods on large scale datasets that have different experimental complexity in terms of dimensionality, number of hidden cell populations or levels of noise. Our benchmark approach extends the previous comparative studies (Freytag et al., 2018; Duò et al., 2018; Tian et al., 2019) to a broader range evaluation of the algorithms which depends on the parametrization (user-specified parameter choices) and previously mentioned dataset differences. The results presented here showed that the performance of the methods strongly depends on different user-specified parameter settings and that the dataset dimensionality and composition often determines the overall accuracy of the methods. Overall, this means that most of the methods lack of robustness with respect to the tuning parameters or differences among the datasets. We found that both aspects were partially ignored in the previous studies, preventing the user to better understand the potentials and limitations of each method. Although, we did not find a “golden” rule for choosing optimal parameter configurations, our study identified some model-dependent choices which were found more robust than others. Despite that, our study also showed that the overall performance is still far from being optimal. Hence, there is a need for developing novel and more accurate methods, in particular for those datasets containing a very large and heterogeneous amount of cells. Evaluating and improving clustering approaches for scRNAseq data might be beneficial for several areas of biomedical science such as immunology, cell development and cancer see for example Haque et al. (2017).

The analysis of real and simulated datasets confirmed that the high sample size and the high number of cell populations have a great impact on scRNAseq clustering methods. In particular, we found that the estimation of the number of clusters remains challenging. We confirmed these issues in several analyzed cases where the methods either under or overestimated the true number of cell populations and the simulated cell groups. In real scRNAseq applications, overestimation of the number of clusters might be just due to methods identifying previously unknown biologically relevant sub-groups. However, underestimation of the clusters means that methods failed to distinguish accurately differences between populations of cells. Since in scRNAseq clustering we also aim to identify novel and/or rare cell populations, we typically do not know the number of cell populations. The failure to identify the number of sub-groups in a consistent manner is a considerable drawback when it comes to practical applications of such methods. In fact, such failure is usually paid with a lower ARI index. By comparing the performance of each method when the true number of clusters was imputed with those when it has been estimated from the data, one can quantify the impact that a more accurate estimate on the number of groups can have on the overall performance.

With the development of new high-throughput scRNAseq protocols, the data dimensionality grows and one has to consider not only methodological performance but also computational requirements of the different approaches. We have demonstrated that computational cost does not always trade for empirical accuracy and some configurations are just impractical for specific protocols. Since, larger and more complex datasets are going to be produced by novel droplet-based protocols, the computational feasibility needs to be better faced and more attention should be given in designing methods with efficient algorithms and in better exploiting high-performance computing in their implementation.

Taken all together, our systematic evaluation of the methods confirmed some common sense assumptions or expected results, but also identified new potential issues in scRNAseq clustering. The summary of the methods presented here can guide the readers in a number of options that the methods provide also giving awareness about their possible limitations. Moreover, the benchmark pipeline developed for this study is freely available and can be easily modified to add novel methods.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GSE84133, GSE65525, GSE60361, GSE67835, GSE45719, MTAB-3321, MTAB-2600, GSE81861, GSE74672, GSE71585, GSE45719, MTAB-5061, GSE71585, GSE81608, GSE36552, GSE57249, GSE52583.

## AUTHOR CONTRIBUTIONS

MK designed and implemented the clustering benchmark study, performed both real and simulated analysis, selected

and discussed results and wrote the manuscript, YR and LC contributed to the design of the benchmark study, the selection, and discussion of results and the drafting of the manuscript, AB contributed to the selection and discussion of the real and simulated data analysis and provided constructive comments on the benchmark study, CA contributed to the design of the benchmark, guided and supervised all phases of benchmark implementation, selection, and discussion of results and wrote the manuscript. All authors read and approved the manuscript.

## FUNDING

We acknowledge INCIPIT PhD program co-funded by the COFUND scheme (Marie Skłodowska-Curie Actions) grant

## REFERENCES

- Andrews, T. S., and Hemberg, M. (2018). Identifying cell populations with scRNASeq. *Mol. Asp. Med.* 59, 114–122. doi: 10.1016/j.mam.2017.07.002
- Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., et al. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* 3 (4), 346–360. doi: 10.1016/j.cels.2016.08.011
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. (Methodological)* 57 (1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x
- Biase, F. H., Cao, X., and Zhong, S. (2014). Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res.* 24 (11), 1787–1796. doi: 10.1101/gr.177725.114
- Chen, G., Ning, B., and Shi, T. (2019). Single-Cell RNA-Seq technologies and related computational data analysis. *Front. Genet.* 10, 317. doi: 10.3389/fgene.2019.00317
- Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., et al. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U.S.A.* 112 (23), 7285–7290. doi: 10.1073/pnas.1507125112
- Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343 (6167), 193–196. doi: 10.1126/science.1245316
- Duò, A., Robinson, M. D., and Soneson, C. (2018). A systematic performance evaluation of clustering methods for single-cell rna-seq data [version 1; referees: 2 approved with reservations]. *F1000Research* 7, 1141. doi: 10.12688/f1000research.15666.1
- Ester, M., Kriegl, H., Sander, J., and Xu, X. (1996). “A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise,” Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD’96 (Portland: AAAI Press), 226–231. Available at: <http://dl.acm.org/citation.cfm?id=3001460.3001507>.
- Farbehi, N., Patrick, R., Dorison, A., Xaymardan, M., Janbandhu, V., Wystub-Lis, K., et al. (2019). Single-cell expression profiling reveals dynamic flux of cardiac stromal, vascular and immune cells in health and injury. *Elife* 8, e43882. doi: 10.7554/eLife.43882
- Fraley, C., and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* 97 (458), 611–631. doi: 10.1198/016214502760047131
- Freytag, S., Tian, L., Lönnstedt, I., Milica, Ng., and Bahlo, M. (2018). Comparison of clustering tools in r for medium-sized 10x genomics single-cell rna-sequencing data [version 1; referees: 1 approved, 2 approved with reservations]. *F1000Research* 7, 1297. doi: 10.12688/f1000research.15809.1
- Goolam, M., Scialdone, A., Graham, S. J. L., Macaulay, I. C., Jedrusik, A., Hupalowska, A., et al. (2016). Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-Cell mouse embryos. *Cell* 165 (1), 61–74. doi: 10.1016/j.cell.2016.01.047
- agreement n. 665403, EPIGEN project and ADVISE project for financial support.

## ACKNOWLEDGMENTS

MK would like to thank LC for warm hospitality while visiting the School of Mathematics, University of Leeds, to carry out part of this work.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.01253/full#supplementary-material>

- Haque, A., Engel, J., Teichmann, S. A., and Lonnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* 9 (1), 75. doi: 10.1186/s13073-017-0467-4
- Herman, J., and Grün, D. (2018). FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* 15, 379–386. doi: 10.1101/218115
- Hollander, M., and Wolfe, D. A. (1973). Nonparametric statistical methods. *Wiley Series in Probability and Statistics - Applied Probability and Statistics Section*. New York-Sydney-Tokyo-Mexico City: John Wiley & Sons. Available at: <https://books.google.it/books?id=ajxMAAAAMAAJ>.
- Ji, Z., and Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* 44 (13), e117. doi: 10.1093/nar/gkw430
- Julia, M., Telenti, A., and Rausell, A. (2015). Sincell: an R/Bioconductor package for statistical assessment of cell-state hierarchies from single-cell RNA-seq. *Bioinformatics* 31 (20), 3380–3382. doi: 10.1093/bioinformatics/btv368
- Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14 (5), 483–486. doi: 10.1038/nmeth4236
- Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* 20 (5), 273–282. doi: 10.1038/s41576-018-0088-9
- Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161 (5), 1187–1201. doi: 10.1016/j.cell.2015.04.044
- Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C., Illicic, T., Henriksson, J., Natarajan, K. N., et al. (2015). Single Cell RNA-Sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* 17 (4), 471–485. doi: 10.1016/j.stem.2015.09.011
- Lawrence, H., and Phipps, A. (1985). Comparing partitions. *J. Classif.* 2 (1), 193–218. doi: 10.1007/BF01908075
- Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* 49 (5), 708–718. doi: 10.1038/ng3818
- Lin, P., Trupm, M., and Ho, J. W. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* 18 (1), 59. doi: 10.1186/s13059-017-1188-0
- Luecken, M. D., and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* 15 (6), e8746. doi: 10.15252/msb.20188746
- Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161 (5), 1202–1214. doi: 10.1016/j.cell.2015.05.002
- McCarthy, D., and Lun, A. (2019). Quality control with scater. <https://bioconductor.org/packages/release/bioc/vignettes/scater/inst/doc/vignette-qc.html>.

- McCarthy, D. J., Campbell, K. R., Lun, A. T., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33 (8), 1179–1186. doi: 10.1093/bioinformatics/btw777
- Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y. A., and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Censur. *Nat. Methods* 14 (3), 309–315. doi: 10.1038/nmeth4150
- Ren, X., Zheng, L., and Zhang, Z. (2019). SSCC: a novel computational framework for rapid and accurate clustering large-scale single cell RNA-seq data. *Genomics Proteomics Bioinf.* 17 (2), 201–210. doi: 10.1101/344242
- Romanov, R. A., Zeisel, A., Bakker, J., Girach, F., Hellysaz, A., Tomer, R., et al. (2016). Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* 20 (2), 176–188. doi: 10.1038/nn4462
- Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *R. J.* 8 (1), 289–317. doi: 10.32614/RJ-2016-021
- Segerstolpe, A., Palasantza, A., Eliasson, P., Andersson, E. M., Andreasson, A. C., Sun, X., et al. (2016). Single-Cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* 24 (4), 593–607. doi: 10.1016/j.cmet.2016.08.020
- Senabouth, A., Lukowski, S. W., Hernandez, J. A., Andersen, S. B., Mei, X., Nguyen, Q. H., et al. (2019). Ascend: R package for analysis of single-cell RNA-seq data. *Gigascience* 8 (8). doi: 10.1093/gigascience/giz087
- Sun, Z., Wang, T., Deng, K., Wang, X. F., Lafyatis, R., Ding, Y., et al. (2018). DIMM-SC: a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics* 34 (1), 139–146. doi: 10.1093/bioinformatics/btx490
- Svensson, V., Natarajan, K. N., Ly, L. H., Miragaia, R. J., Labalette, C., Macaulay, I. C., et al. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14 (4), 381–387. doi: 10.1038/nmeth4220
- Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13 (4), 599–604. doi: 10.1038/nprot.2017.149
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6 (5), 377–382. doi: 10.1038/nmeth1315
- Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* 19 (2), 335–346. doi: 10.1038/nn4216
- Tian, L., Dong, X., Freytag, S., Le Cao, K. A., Su, S., JalalAbadi, A., et al. (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* 16 (6), 479–487. doi: 10.1038/s41592-019-0425-8
- Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., et al. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509 (7500), 371–375. doi: 10.1038/nature13173
- Vuong, N. H., Cook, D. P., Forrest, L. A., Carter, L. E., Robineau-Charette, P., Kofsky, J. M., et al. (2018). Single-cell RNA-sequencing reveals transcriptional dynamics of estrogen-induced dysplasia in the ovarian surface epithelium. *PLoS Genet.* 14 (11), e1007788. doi: 10.1371/journal.pgen.1007788
- Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* 14 (4), 414–416. doi: 10.1038/nmeth4207
- Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., et al. (2016). RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.* 24 (4), 608–615. doi: 10.1016/j.cmet.2016.08.018
- Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20 (9), 1131–1139. doi: 10.1038/nsmb2660
- Yip, S. H., Wang, P., Kocher, J. A., Sham, P. C., and Wang, J. (2017). Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.* 45 (22), 13097. doi: 10.1093/nar/gkx1189
- Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* 18 (1), 174. doi: 10.1186/s13059-017-1305-0
- Zeisel, A., Munoz-Manchado, A. B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., et al. (2015). Brain structure, cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347 (6226), 1138–1142. doi: 10.1126/science.aaa1934
- Zuraskiene, J., and Yau, C. (2016). pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinf.* 17, 140. doi: 10.1186/s12859-016-0984-y

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Krzak, Raykov, Boukouvalas, Cutillo and Angelini. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing

Tian Mou<sup>1</sup>, Wenjiang Deng<sup>1</sup>, Fengyun Gu<sup>2</sup>, Yudi Pawitan<sup>1\*</sup> and Trung Nghia Vu<sup>1\*</sup>

<sup>1</sup> Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, <sup>2</sup> School of Mathematical Sciences, University College Cork, Cork, Ireland

## OPEN ACCESS

### Edited by:

Monica Bianchini,  
University of Siena, Italy

### Reviewed by:

Max Robinson,  
Institute for Systems Biology (ISB),  
United States  
Yuriy L. Orlov,  
First Moscow State Medical University,  
Russia

### \*Correspondence:

Yudi Pawitan  
yudi.pawitan@ki.se  
Trung Nghia Vu  
Trungnghia.vu@ki.se

### Specialty section:

This article was submitted to  
Bioinformatics and  
Computational Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 06 August 2019

**Accepted:** 05 December 2019

**Published:** 17 January 2020

### Citation:

Mou T, Deng W, Gu F, Pawitan Y and  
Vu TN (2020) Reproducibility of  
Methods to Detect Differentially  
Expressed Genes from Single-Cell  
RNA Sequencing.  
Front. Genet. 10:1331.  
doi: 10.3389/fgene.2019.01331

Detection of differentially expressed genes is a common task in single-cell RNA-seq (scRNA-seq) studies. Various methods based on both bulk-cell and single-cell approaches are in current use. Due to the unique distributional characteristics of single-cell data, it is important to compare these methods with rigorous statistical assessments. In this study, we assess the reproducibility of 9 tools for differential expression analysis in scRNA-seq data. These tools include four methods originally designed for scRNA-seq data, three popular methods originally developed for bulk-cell RNA-seq data but have been applied in scRNA-seq analysis, and two general statistical tests. Instead of comparing the performance across all genes, we compare the methods in terms of the rediscovery rates (RDRs) of top-ranked genes, separately for highly and lowly expressed genes. Three real and one simulated scRNA-seq data sets are used for the comparisons. The results indicate that some widely used methods, such as edgeR and monocle, have worse RDR performances compared to the other methods, especially for the top-ranked genes. For highly expressed genes, many bulk-cell-based methods can perform similarly to the methods designed for scRNA-seq data. But for the lowly expressed genes performance varies substantially; edgeR and monocle are too liberal and have poor control of false positives, while DESeq2 is too conservative and consequently loses sensitivity compared to the other methods. BPSC, Limma, DEsingle, MAST, t-test and Wilcoxon have similar performances in the real data sets. Overall, the scRNA-seq based method BPSC performs well against the other methods, particularly when there is a sufficient number of cells.

**Keywords:** single cell, RNA sequencing, differential expression, rediscovery rate, comparison

## INTRODUCTION

Traditional gene expression profiling with high-throughput RNA-sequencing technology measures the aggregated expression levels of genes from a collection of millions of cells. Such bulk-cell RNA-sequencing cannot capture cellular heterogeneity since there is no cell-specific information (Miao and Zhang, 2016; Jaakkola et al., 2017). Single-cell RNA sequencing (scRNA-seq) has developed rapidly as a powerful technology for studying transcriptomics at the single-cell level (Sandberg, 2014). However, compared to bulk-cell data, scRNA-seq data has a higher level of noise due to both

biological and technical reasons, for example, lower input materials, cell-cycle phase, amplification biases, and the so-called dropout and bursting events (Dal Molin et al., 2017; Jaakkola et al., 2017; Soneson and Robinson, 2018). Such events are caused by the stochastic nature of the gene expression process at the single-cell level (Gong et al., 2018). The dropout events generate zero expression, statistically leading to zero inflation in the gene-expression distribution at a much higher proportion than expected under the standard negative-binomial model commonly assumed in bulk-cell data (Miao et al., 2018). Aggregation of expression in bulk-cell data reduces the effects of these single-cell events.

Differential expression (DE) analysis to discover quantitative changes between different groups or conditions plays an important role for understanding the molecular basis of phenotypic variation. However, due to the unique characteristics of scRNA-seq data, it is not immediately obvious that we can just use standard methods developed for bulk-cell data. A particular challenge is dealing with the large number of low (or zero) read counts in the scRNA-seq data. A previous study (Love et al., 2014) has shown the phenomenon that weakly expressed genes tend to produce more differences than highly expressed genes. For instance, to tackle this issue, several DE methods have been developed for scRNA-seq data, for example, BPSC (Vu et al., 2016), MAST (Finak et al., 2015), and monocle (Qiu et al., 2017). In general, bulk-cell-based DE methods were not originally designed to deal with a large fraction of lowly expressed genes. Yet, in practice, many studies use the bulk-cell-based DE methods for single-cell data, such as edgeR (Wang et al., 2016) or limma (Ziegenhain et al., 2017). Furthermore, various pipelines and workflows of RNA-seq analysis do not consider scRNA-seq data specifically (Lun et al., 2016; Chen et al., 2016; Law et al., 2016) and suggest users apply the bulk-cell-based methods to scRNA-seq data (Zhu et al., 2017).

These bulk-cell-based methods are methodologically sophisticated, and they have been used for scRNA-seq data, but evaluation of their applicability to scRNA-seq data is still uncommon and different studies have reported opposite results. For example, authors in a recent study (Jaakkola et al., 2017) compared five DE methods, including two single-cell-based methods and three bulk-cell-based methods. They concluded that the original DESeq (Anders and Huber, 2010) and limma (Law et al., 2014) are not suitable for scRNA-seq data. In contrast, another comparative study (Miao and Zhang, 2016) declared that DESeq tends to outperform other methods on scRNA-seq data. Most comparative studies (Miao and Zhang, 2016; Dal Molin et al., 2017; Jaakkola et al., 2017; Soneson and Robinson, 2018) agree that bulk-cell-based methods are applicable to scRNA-seq even though there is a lack of agreement in finding DE genes by these DE methods (Wang et al., 2019) and it is difficult to identify the best performing tool for DE analysis of scRNA-seq data (Dal Molin et al., 2017). Therefore, further evaluations of these DE methods, including both bulk-cell- and single-cell-based methods in different aspects, are warranted for better understanding of the methodologies when applied to scRNA-seq studies.

To compare the DE methods, previous studies have used conventional statistics such as type-I error rate, false discovery rate (FDR) and receiver operating characteristic (ROC) curve. Notably, these metrics are applied to the full collection of genes. Reproducibility is also an important metric, although it is sometimes calculated differently in the different studies. For example, a recent study (Miao and Zhang, 2016) assesses the reproducibility of the methods by looking at the average of the overlap of top 1,000 DE genes (ranked by p-value) across 20 replicates. In each replicate, a control group and a testing group are sampled with a different random seed. Another measure of reproducibility (Jaakkola et al., 2017) compares the precision and recall of the detection of all DE genes between the full data set and its subsets.

In this study, we compare the performance of nine DE methods, including both bulk-cell and single-cell-based approaches as well as general statistical tests not specifically designed for RNA-seq data. We focus on the reproducibility of the methods in terms of rediscovery rate (RDR) (Ganna et al., 2014) of top-ranking genes. RDR is defined as the proportion of top-ranking findings detected from a training sample that are replicated in a validation sample. In high-throughput studies, the RDR is determined by both the false positive rate (FPR) and power (Ganna et al., 2014), so it is a convenient and easily understood metric for the comparison of methods. Limiting the assessment to top-ranking genes turns out to be important. Firstly, it follows the data analytic process we perform in practice, where the top-ranked genes are usually considered the most interesting ones for further biological analyses or interpretation. Secondly, some methods perform differently for the top-ranked genes and across all genes. Besides the RDR, type-I error rate or FPR, and ROC are also used as extra metrics for the comparisons.

To get realistic distributional characteristics and capture some diversity in single-data data, we utilize three real scRNA-seq data sets; in addition, we use simulated data from the beta-Poisson model (BPSC), which has been suggested for scRNA-seq data in a recent study (Vu et al., 2016). Because of their distinct distributions, the groups of highly and lowly expressed genes are also considered separately, as the latter is more affected by single-cell specific events such as dropouts.

## RESULTS

We compare nine methods for detecting differentially expressed isoforms, including edgeR (Robinson et al., 2010), DESeq2 (Love et al., 2014), DEsingle (Miao et al., 2018), monocle (Qiu et al., 2017), BPSC (Vu et al., 2016), MAST (Finak et al., 2015), t-test (Welch, 1947), Wilcoxon rank sum test (Hollander et al., 2013), limmatrend (Law et al., 2014). Among those, edgeR, DESeq2 and limmatrend are designed for bulk-cell RNA-seq analysis; and DEsingle, monocle, BPSC, and MAST are developed based on scRNA-seq data. T-test and Wilcoxon rank-sum test are general comparison tests not specific to RNA-seq data. **Table 1** compares the methods in terms of (i) distribution assumption, (ii) original data motivation (bulk-cell or single-cell data), (iii) test statistic,

**TABLE 1** | List of the differential expression analysis methods.

Method	Distribution assumption	Designed for	Test statistic	Run time	Version [Ref.]	Input
BPSC	Beta-Poisson	Single cell	z-test	Hours	0.99.2 (Vu et al., 2016)	CPM
DEsingle	Zero-Inflated Negative Binomial	Single cell	Likelihood ratio test	Hours	1.2.1 (Miao et al., 2018)	raw counts
MAST	Normal (Generalized linear hurdle)	Single cell	Likelihood ratio test	Minutes	1.8.2 (Finak et al., 2015)	$\log_2(\text{CPM}+1)$
monocle	Normal (Generalized additive model)	Single cell	Likelihood ratio test	Minutes	2.10.1 (Qiu et al., 2017)	raw counts
DESeq2	Negative Binomial	Bulk cell	Wald test	Minutes	1.22.2 (Love et al., 2014)	raw counts
edgeR	Negative Binomial	Bulk cell	Quasi-likelihood F-test	Minutes	3.24.3 (Robinson et al., 2010)	raw counts
limmatrend	Normal (linear model)	Bulk cell	Empirical-Bayes Moderated t-statistics	Seconds	(Law et al., 2014)	$\log_2(\text{CPM}+1)$
t-test	Normal	General	t-test	Seconds	(Welch, 1947)	$\log_2(\text{CPM}+1)$
Wilcoxon	Nonparametric	General	Wilcoxon	Minutes	(Hollander et al., 2013)	$\log_2(\text{CPM}+1)$

and (iv) run time for a typical data set used in the comparisons. We also state the exact version of each software tool used in the comparisons.

To get realistic distributional characteristics, the following three real scRNA-seq data sets are used as the basis for simulations. (Different papers and projects use isoform- and gene-level expressions. For simplicity, we shall use the terms “isoform” and “gene” interchangeably.)

- Breast-cancer cell line MDA-MB-231 data set (Athreya et al., 2017) has two groups: control and metformin-treated, 80 cells in each group. The expression estimates of 26,775 isoforms from Cufflinks are used in the analysis.
- Mouse embryonic stem cells (mESCs) belong to two groups from different culture conditions, 94 cells in group 1 and 174 cells in group 2; see the Materials and Methods section for details. The expression estimates of 112,593 isoforms are provided by the *Conquer* project (Soneson and Robinson, 2018).
- Neuronal progenitor cells (NPCs) also form two groups, one from the patient and the other from a healthy donor (Iacono et al., 2018), 360 cells in each group. The expression estimates of 41,020 genes are provided by the bigScaLe project (Iacono et al., 2018).

In addition, we also simulate single-cell data based on the beta-Poisson model (Vu et al., 2016). The variation in sample sizes of the three real data sets, from 160 to 720, allows us to compare the performance of each method at different sample sizes. More details of the methods and data sets are given in the Materials and Methods section.

In each experiment, the comparison focuses on the DE analysis of two predefined groups of cells. Briefly, an equal number of samples is randomly selected from the two groups in the original data set to generate the training set. For each sampled cell from a real data set, all isoforms are taken together; this preserves the statistical dependencies between the isoforms. For the validation set, a different set of samples from both groups is selected. The selection of training and validation sets is repeated 50 times to average out the effect of random selection. Note that the training and validation sets are always disjoint. The nine DE methods are then applied to the training and validation sets separately.

## Type-I Error Control

For each real data set, we generate a null data set by randomly sampling from the two groups combined (i.e., ignoring the group labels). Thus, the null data sets are expected to have no true DE

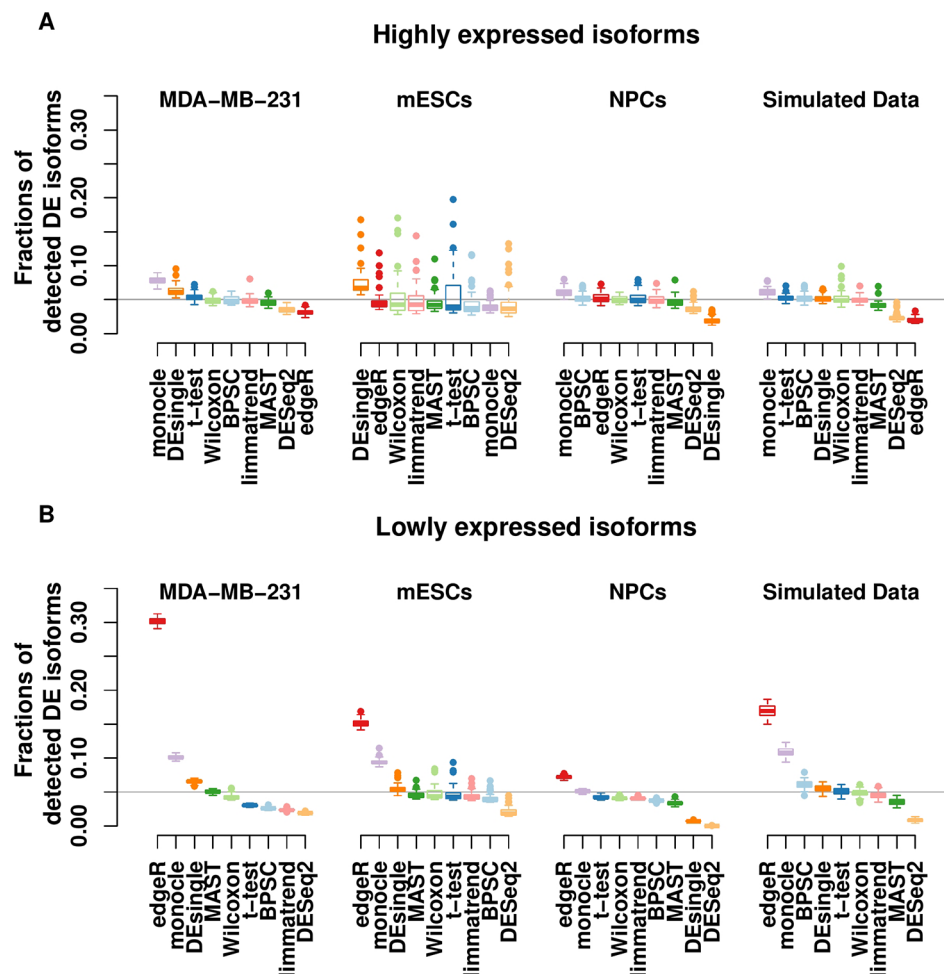
isoforms, and the p-value distribution of each method is expected to be uniform. Theoretically, the p-values should follow a uniform distribution if the null hypothesis is true (Murdoch et al., 2008; Bland, 2013). The uniformity of p-value distribution under the null hypothesis can be used to assess the performance of methods. We calculate the type-I error rate by recording the fraction of the detected DE isoforms that are assigned a significant p-value ( $p < 0.05$ ). This fraction is also known as the FPR. To highlight the effects of the dropout events, which tend to produce low expression and zero inflation, we split the isoforms into two groups based on the expression level: highly expressed isoforms and lowly expressed isoforms. The former refers to the isoforms with an estimated expression above 1 transcripts-per-million (TPM) in more than 25% of the cells, and the remaining isoforms are assigned to the latter. This threshold was also suggested in a recent comparative study of DE methods in scRNA-seq (Soneson and Robinson, 2018).

Results in **Figure 1A** show that for highly expressed isoforms, most methods manage to control the FPR close to the target 0.05. Two single-cell-based methods, monocle and DEsingle, are not stable, as their FPRs fluctuate the most from the expected error rate. As expected, the bulk-cell-based methods, edgeR, DESeq2, and limmatrend, perform well on this group, and DESeq2 is the most conservative.

For the lowly expressed isoforms, DESeq2 is also the most conservative method, **Figure 1B**. It identifies fewer significant isoforms, so the FPR is significantly lower than the expected level (0.05) in all data sets. In contrast, edgeR has the highest FPR, sometimes substantially above the target value. Similarly, monocle also has a large number of false positive findings. The FPR of DEsingle has a slight variation, as it is liberal for MDA-MB-231 data set, conservative for NPCs data set, and performs rather well in the other data sets. Thus, it seems the performance of DEsingle is not stable and highly dependent on data sets. The histograms of p-values (**Figure S1** in the Supplementary report) further illustrate that few methods returned uniformly distributed p-values under the null hypothesis for the lowly expressed isoforms, while most methods have a better uniformity for the highly expressed isoforms.

## The RDR

The RDR is the proportion of the top-ranking DE isoforms in the training set that is found to be significant ( $p < 0.05$ ) in the validation set. The RDR is calculated based on the top 5%, 10%, 20% DE and all isoforms in the training set.



**FIGURE 1 |** Type-I error control for the groups of highly expressed isoforms (A) and lowly expressed isoforms (B) of the three real scRNA-seq data sets and the simulated data set. The values in the y-axis are the fractions of isoforms with  $p < 0.05$  under the null hypothesis. The horizontal line indicates the expected error rate at 0.05. Box plots of the methods in the x-axis are the collection from 50 replicates. The methods are ordered by median false positive rate (FPR) across all replicates. The number of highly expressed isoforms in MDA-MB-231, mESCs, NPCs and simulated data sets are 8,299, 31,895, 10,422, and 8077, respectively. The corresponding number of lowly expressed isoforms are 18,476, 80,698, 30,378, and 1,923.

### RDR Analysis Under the Null Hypothesis

The RDR of the null data sets from the real data in Section 2.1 are reported in **Figure 2**. Panels A and B present the results for the groups of highly expressed isoforms and lowly expressed isoforms, respectively. Under the null hypothesis of no group effect, the expected RDR is 0.05. Similar to the results from the type-I error control in Section 2.1, the RDRs of all methods are generally better for highly expressed isoforms. Monocle and DEsingle are the worst, as their RDRs are often far from 0.05. However, the performances improve for the larger number top DE isoforms. For example, the RDR of monocle for all isoforms in the NPCs data set is very close to the expected value, but it is much higher than 0.05 among the top 5% DE isoforms. Similarly, for the mESCs data set, the RDR of edgeR for all isoforms is close to 0.05, but it is consistently higher than this target value for the smaller number of top DE isoforms. Thus, comparing the performances based on all isoforms could be misleading.

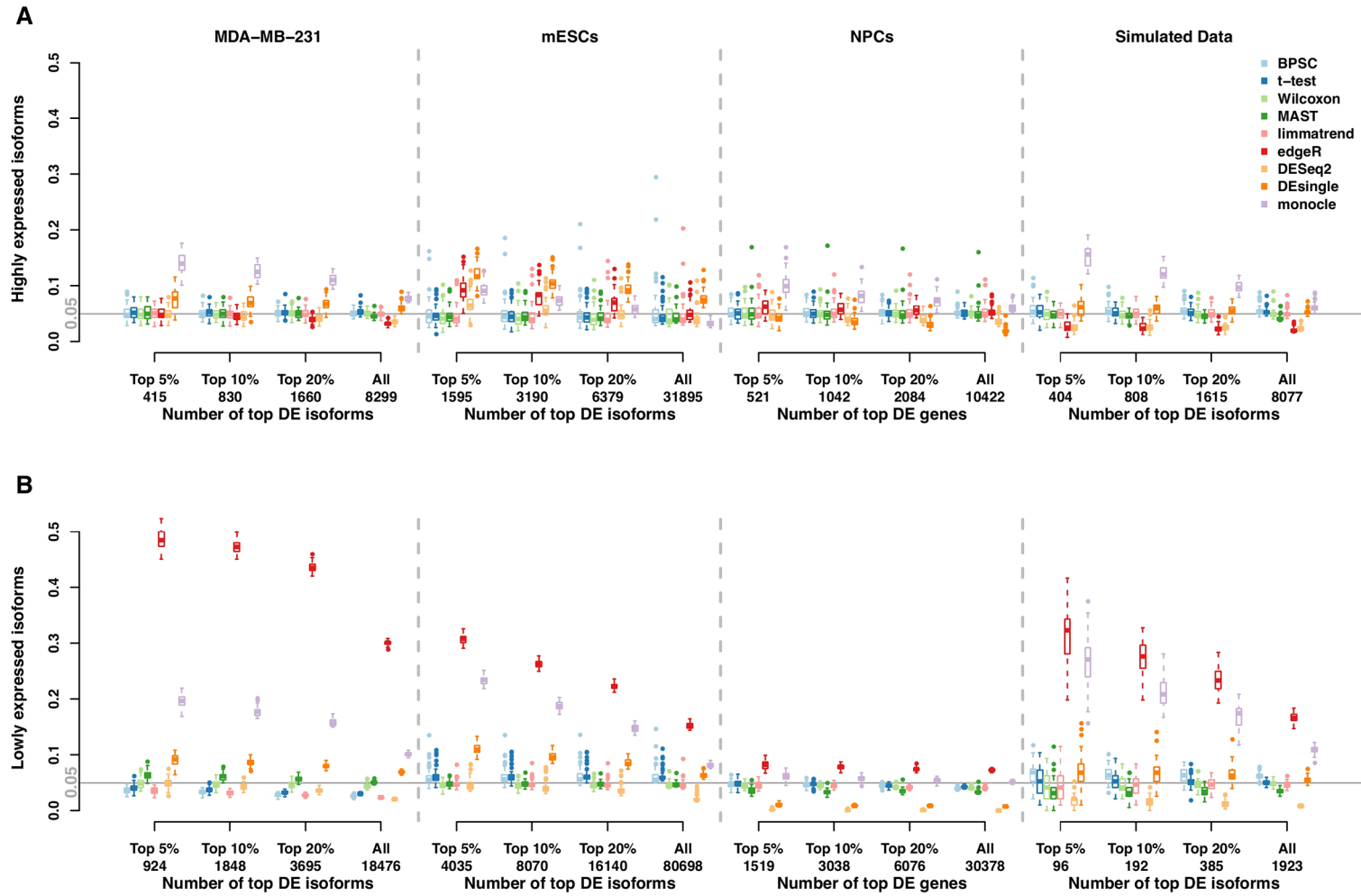
These patterns are much more pronounced for lowly expressed isoforms; see **Figure 2B**. In this case, edgeR performs worst in all data sets; this result is consistent with other studies (Soneson and Robinson, 2018). The performances of DESeq2 still tend to be conservative in both groups of isoforms, while other methods generally have RDR around the expected value.

We further evaluate RDR of the DE methods in the simulated beta-Poisson data set. Results from 50 replicates of the null data sets from the simulated data are reported in the rightmost plots of **Figures 2A, B**. The similar patterns of RDR of DE methods for both isoform groups confirm the results from the real data sets. In particular, monocle has poor performances in both groups, and edgeR does not perform well with lowly expressed isoforms.

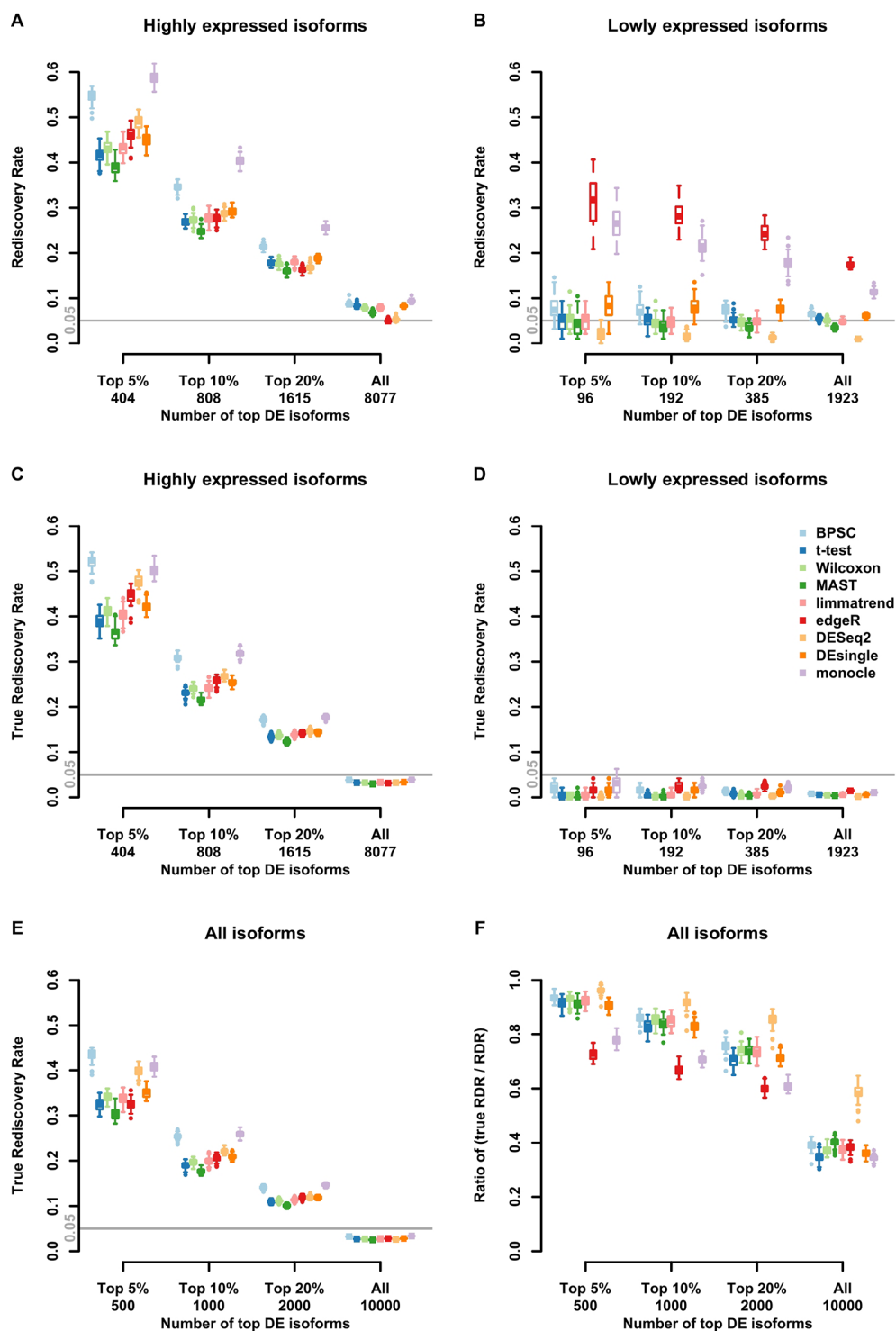
### RDR Analysis Under the Alternative Hypothesis

Results of RDR analysis for the simulated beta-Poisson data under the alternative hypothesis are presented in **Figure 3**. As





**FIGURE 2** | Rediscovery rate (RDR) of differential expression (DE) isoforms in the real and simulated scRNA-seq data sets under the null hypothesis calculated from the top 5%, 10%, 20% DE and all isoforms. (Panels **A** and **B**) present the results of groups of highly expressed isoforms and lowly expressed isoforms, respectively. The number of highly expressed isoforms in MDA-MB-231, mESCs, neuronal progenitor cells (NPCs), and simulated data sets are 8,299, 31,895, 10,422, and 8,077, respectively. The corresponding number of lowly expressed isoforms are 18,476, 80,698, 30,378, and 1,923.



**FIGURE 3 |** Observed rediscovery rate (RDR) and true rediscovery rate (TrueRDR) of differential expression (DE) isoforms in the simulated beta-Poisson data set under the alternative hypotheses calculated among the top 5%, 10%, 20% DE and all isoforms. (Panels **A** and **B**) present the rediscovery rate in the groups of highly and lowly expressed isoforms, respectively. (Panels **C** and **D**) display the true rediscovery rate collected from highly and lowly expressed isoforms separately. (Panels **E**) displays the true rediscovery rate collected from both highly and lowly expressed isoforms. (Panels **F**) presents the ratio between true RDR and observed RDR. The number of highly expressed isoforms in the simulated data set is 8,077, and the number of lowly expressed isoforms is 1,923.

described in more detail in the Materials and Methods section, 5% of the isoforms are randomly selected to be differentially expressed between the two groups (hence true DE isoforms). For highly expressed isoforms (**Figure 3A**), *monocle* and *BPSC* have the highest RDRs across the top 5%, 10%, 20% and all DE isoforms, while *edgeR* is comparable to the rest. *DESeq2* is conservative for the null data sets and the group of lowly expressed isoforms, but its performance is comparable to other methods. For lowly expressed isoforms, *edgeR* and *monocle* produce the highest RDRs compared to other methods (**Figure 3B**). However, remember that from the previous subsection we know these two methods have high false positive rates.

In the simulated data, we in fact know the true DE status, so we can evaluate the true RDR, which is defined as the proportion of the true positives in the validation set among the top DE isoforms identified in the training set. In other words, the true RDR is the intersection of rediscovered genes and true DE genes. This is shown in **Figures 3C, D**. First, let us consider panel D. While there are 5% true DE isoforms, the statistical power for the lowly expressed isoforms is tiny, so very few of the true DE isoforms appear among the top-ranking genes and these isoforms do not produce significant p-values in the validation set. Hence the rediscoveries are mostly false positives. This means that there are reproducible features of the data, such as zero inflation, that consistently create problems for *monocle* and *edgeR* to the point of producing false positives in validation data. These results highlight the challenge in finding true DE among lowly expressed isoforms, or equivalently, the ease of producing false positives.

From **Figure 3C**, the true RDRs of 3 methods including *BPSC*, *monocle* and *DESeq2* are better than the other methods. The overall true RDRs are given in **Figure 3E**, which in this case look similar to the result for highly expressed isoforms, but do not reflect the results for lowly expressed ones. **Figure 3F** shows the ratio of true RDR to observed (RDR). *DESeq2* has the highest ratio among the comparing methods, indicating a good specificity in detecting DE isoforms. However, *DESeq2* generally discovers fewer true DE isoforms, i.e., lower sensitivity, compared to *BPSC*. Two methods of *edgeR* and *monocle* have a lower ratio than the other methods since they have more false discoveries. In the next section, the balance between sensitivity and specificity of the methods are taken into account *via* the ROC curve.

For the real data sets, there are no significant differences in RDR performance for the top 5%, 10%, 20% DE isoforms between nine DE methods in the group of highly expressed isoforms (**Figure S2A** in the Supplementary report). However, similar to the results of the simulated data set, RDRs of *edgeR* and *monocle* are highly liberal, while *DESeq2* tends to be too conservative for the lowly expressed isoforms (**Figure S2B**). We have performed other simulations and analyzed two other datasets that confirmed this observation. This is given in the Supplementary Material and described in the Discussion section.

## ROC Performance

Performances of the DE methods on the simulated data with the alternative hypothesis are also evaluated using the area under the

ROC curve (AUC). In **Figure 4**, the AUC and ROC curves of top 5% DE isoforms and all isoforms over 50 replicates are presented in panels A and B, respectively. For *edgeR* and *monocle*, there are obvious differences between their performances for top 5% DE isoforms and for all isoforms. For the top 5% isoforms, these two methods perform poorly compared to the other methods. However, if all isoforms are considered, the two methods are comparable with the other methods when more isoforms are taken into account. Results for the top 10% and 20% DE isoforms are given in **Figure S3** in the Supplementary report. Among these methods, *BPSC* and *DESeq2* are consistently the top performing methods with the highest AUC values for different sizes of top DE isoform sets. Overall, these results are in agreement with the results from RDR analyses.

## MATERIALS AND METHODS

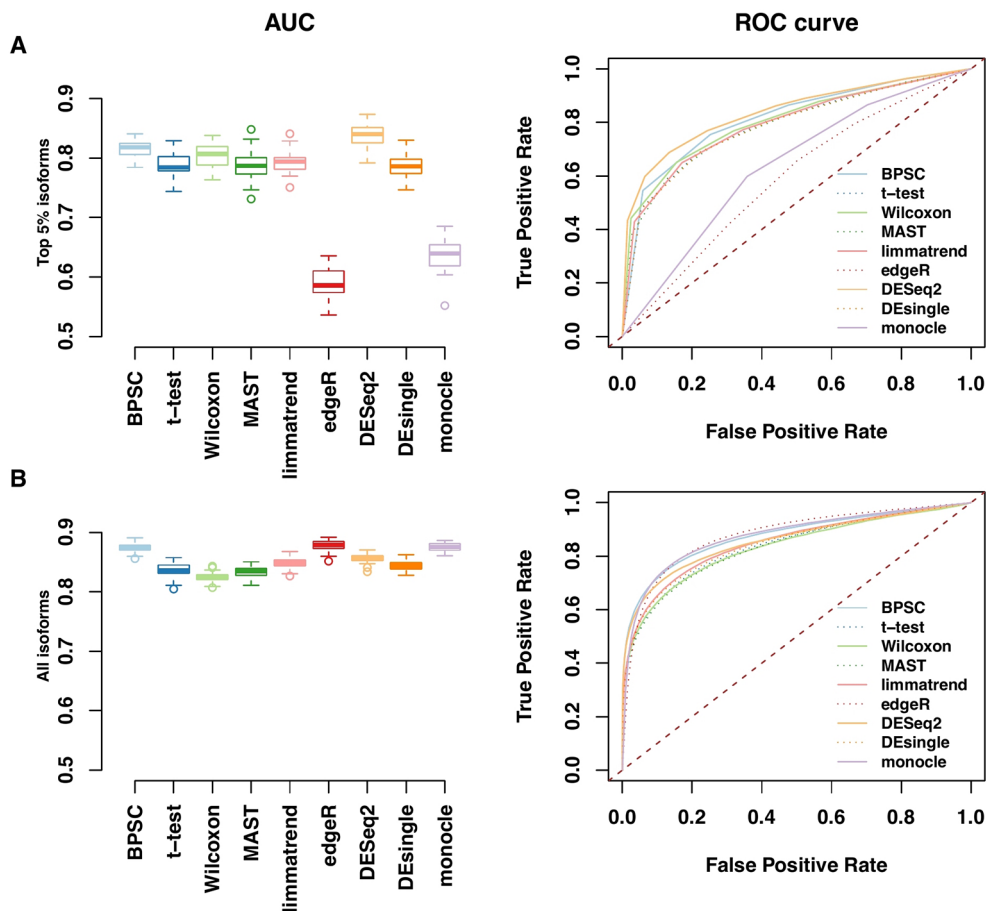
### Experimental and Synthetic Data Sets

To capture the true distributional characteristics of real data, three real scRNA-seq data sets are used for the evaluation of the nine DE methods. The first data set (MDA-MB-231) includes 160 single cells from a triple-negative breast cancer cell line, half of which are treated with metformin. The cells are captured using the Fluidigm C1 system and sequenced on Illumina HiSeq 2500 machines for 80 control and 80 treated cells separately. Then we use *Cufflinks* (Trapnell et al., 2010) to estimate the isoform expression. This data set contains a total of 26,775 isoforms across 160 single cells. The average number of reads per cell is ~649,000.

The second data set (mESCs) is collected from a public scRNA-seq data (GSE60749-GPL13112) in the *Conquer* data set (Soneson and Robinson, 2018), which provides expression estimates of isoforms. The compared single cells are 94 individual v6.5 mouse embryonic stem cells (mESCs) with culture conditions 2i+LIF (group 1) vs. 174 v6.5 mESCs with culture conditions in serum +LIF (group 2). The data are prepared with the C1 System using the SMARTer Ultra Low RNA kit for Illumina Sequencing (Clontech) and protocols provided by Fluidigm. More details of the data can be found in the original paper (Kumar et al., 2014). Then the *Conquer* pipeline estimates isoform abundances using *Salmon* (Patro et al., 2017). This data set contains 112,593 isoforms across 174 single cells in group 1 and 94 single cells in group 2. The average number of reads per cell is ~1.7M, the largest among the 3 data sets.

The third real data set (NPCs) is a subset of GSE102934 data from the NCBI Gene Expression Omnibus (Iacono et al., 2018). This data set has 720 NPCs derived from induced pluripotent stem (iPS) cells, half of which are from a Williams-Beuren patient and the other half are from a healthy donor. The data are sequenced on Illumina HiSeq 2500 platform and then applied massively parallel single-cell RNA sequencing (MARS-Seq) to construct single-cell libraries. This data set contains a total of 41,020 isoforms from 720 single cell, and the average number of reads per cell is 18,600. Thus, this data set has a relatively large number of cells with low sequencing coverage.

The simulated data set for isoform expression of single cells is generated by the beta-Poisson model (Vu et al., 2016). In particular, we generate the counts for each isoform from a



**FIGURE 4 |** Receiver operating characteristic (ROC) and AUC performances for top 5% DE isoforms (A) and all isoforms (B) from the simulated data. Left panels: observed area under the ROC curve (AUC) for all methods; each method has 50 replicates. Right panels: the corresponding ROC curves averaged over 50 replicates.

beta-Poisson distribution with four parameters estimated from the mESCs data set. The four-parameter beta-Poisson model is as follows:

$$BP_4(x|\alpha, \beta, \lambda_1, \lambda_2) = \lambda_2 \text{Poisson}(x|\lambda_1 \text{Beta}(\alpha, \beta)) \quad (1)$$

The mean and variance of the model can be written as

$$\mu = E(X) = \lambda_1 \lambda_2 \phi_1$$

and

$$V \text{ar}(X) = \mu \lambda_2 + \mu^2 \phi_2,$$

where  $\phi_1 = \frac{\alpha}{\alpha + \beta}$  and  $\phi_2 = \frac{\beta}{\alpha(\alpha + \beta + 1)}$ . Crucially, we can modify the parameter  $\lambda_1$  to create mean differences between groups. A more detailed description of the model can be referred to in the original study (Vu et al., 2016).

Beta-Poisson models fitted on the real mESCs data set are used as baseline distributions for simulation. For each isoform, expression values across samples in the control and the treated group are generated from the same beta-Poisson model. To mimic the biological variation, 5% of isoforms are selected to

be differentially expressed between two groups (true DE isoforms). Specifically, the parameter  $\lambda_1$ , which controls the mean of the distribution, is fixed in the control group and multiplied by  $\log_2$  fold change of 1 unit in the treated group. The effect direction is randomly determined for each DE isoform, with equal probability of upregulation and downregulation. In other words, the quantity change between the two compared groups is either two- or half-fold change with equal probability. The simulated data set consists of 80 samples in each of control and treated groups and a total of 10,000 isoforms measured per sample. Library sizes of the single-cell samples are randomly sampled from a range of 1–3 million. We filter out isoforms with zero expression across all samples.

## DE Analysis Methods

Nine DE methods included in this study are categorized into four groups based on different statistical models. These nine methods are selected to cover most statistical models used in recent DE analysis. Regarding other DE methods that are not included in this study, they use similar approach comparing to the nine selected methods. For instance, D3E (Delmans and Hemberg, 2016) utilizes beta-Poisson model which is similar to BPSC;



SCDE (Kharchenko et al., 2014) models the gene expression values using a mixture of negative-binomial distribution for amplification components and a Poisson distribution for dropout events, which is similar to DEsingle; Ballgown (Frazee et al., 2015) is based on the linear modeling strategy which is similar to limma. In this section, we give a brief summary of these nine methods. For more details of the software packages and statistical models, the reader is referred to original publications and related software websites. When applying these tools, we follow standard procedures and parameter settings suggested in software manuals.

### Negative-Binomial-Based Methods

The read counts of an isoform from the technical replicates (repeated sequencing runs of the same sample) are usually modeled to follow a Poisson law (Marioni et al., 2008). However, those from the biological replicates are usually assumed to follow a gamma distribution to accommodate the overdispersion observed in empirical data (Chen et al., 2014). Since the negative binomial (NB) model can be derived as a gamma-Poisson mixture model, several DE methods based on the NB distribution assumption have been developed to accommodate the overdispersion among biological replicates. Note, however, that these theoretical motivations come from bulk-cell RNA-seq data. Two popular methods for this class are edgeR (Robinson et al., 2010) and DESeq2 (Love et al., 2014). The setup is then to assume the expression read counts  $y_{ij} \sim NB(\mu_{ij}, \phi_i)$ , where  $\mu_{ij}$  is the mean and  $\phi_i$  is the dispersion parameter for isoform  $i$  and sample  $j$ . Reliable estimation of the dispersion parameter  $\phi_i$  for each isoform is crucial for detecting DE isoforms. Differences in the estimation of  $\phi_i$  explain the main differences between edgeR and DESeq2.

#### edgeR

A conditional maximum likelihood (CML) is used in edgeR (Robinson et al., 2010) to estimate a common dispersion, which is assumed to be the same for all isoforms. Then this procedure is developed further to allow for the isoform-specific dispersion estimates and an empirical Bayes procedure—approximated by a weighted likelihood—is used to shrink the dispersions toward the common dispersion. The amount of shrinkage is determined by the neighbourhood set that is nearest to isoform  $i$  in average log count-per-million (logCPM). For DE testing, edgeR allows the user to select among different hypothesis tests including quasi-likelihood F-test (edgeQLF) for bulk-cell RNA-seq data and likelihood ratio test (edgeRLFT) for scRNA-seq data as suggested by the developer. However, a recent study (Soneson and Robinson, 2018) shows that edgeQLF performs significantly better than edgeRLFT in scRNA-seq data. Therefore, in this study, we report the results of edgeQLF for the evaluation of edgeR in DE analysis.

#### DESeq2

DESeq2 (Love et al., 2014) uses a similar negative-binomial model as edgeR but facilitates more data-driven shrinkage estimators for dispersion and fold change. DESeq2 assumes the isoforms of similar average expression levels have similar

dispersion and shrinks the isoform-specific dispersion toward a fitted smooth curve by an empirical Bayes approach. To overcome the difficulty in the log fold-change (LFC) estimation for the lowly expressed isoforms, DESeq2 shrinks LFC estimates toward zero when the expression level is low. The shrinkage procedure may result in underestimates of dispersion, thereby producing conservative estimate statistics for the DE test. This helps reduce the FPR at the expense of lower sensitivity.

#### DEsingle

DEsingle (Miao et al., 2018) has another negative-binomial based approach that employs the zero-inflated NB (ZINB) model to discriminate the observed zero values into two parts—constant zeros and zeros from the NB distribution. With the model, DEsingle is designed to overcome the issues of the excessive zero values observed in scRNA-seq data. To detect DE isoforms between two groups, DEsingle first calculates the maximum likelihood estimates (MLE) of two ZINB populations' parameters, then computes the constrained MLE of the two models' parameters under the null hypothesis ( $H_0$ ), and finally uses the likelihood ratio test for testing  $H_0$ .

### Beta-Poisson-Based Methods

#### BPSC

BPSC (Vu et al., 2016) is an analytical procedure based on the beta-Poisson mixture model, which is designed to capture the property of scRNA-seq data. The model is integrated into the generalized linear model (GLM) framework for DE analysis. The sophisticated four-parameter beta-Poisson model is as shown in Eq. (1). The iterative weighted least-squares (IWLS) algorithm is used to estimate the model parameters.

### Normal-Based Methods

#### Limma

Limma (Law et al., 2014) method is based on linear modelling which was originally designed for gene expression microarray data, but has recently been extended to RNA-seq data. In this study, we use limmatrend (Law et al., 2014), a version of limma where the empirical Bayes procedure is modified to incorporate a mean-variance trend for DE analysis. In a recent study of DE analysis of scRNA-seq data (Soneson and Robinson, 2018), limmatrend has the best performances among other versions of limma, such as voomlimma.

#### Monocle

Monocle (Qiu et al., 2017) is a tool originally designed for scRNA-seq data for identifying DE genes that vary across different cell types or across a so-called “pseudo-time.” The mean expression level of each isoform is modeled by generalized additive models (GAMs) which relate one or more predictor variables to a response variable as

$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m),$$

where  $Y$  is a response variable, and  $x_i$ 's are predictor variables. The function  $g$  is a link function, typically the log function, and  $f_i$ 's are nonparametric functions, such as cubic splines or other

smoothing functions. Gene expression level across cells is modeled by a Tobit model; with some approximations, *monocle*'s GAM is thus

$$E(Y) = s(\psi_t(b_x, s_i)) + \epsilon,$$

where  $\psi_t(b_x, s_i)$  is the assigned pseudo-time of a cell and  $s$  is a cubic smoothing function with (by default) three effective degrees of freedom.  $\epsilon$  is the error term that is normally distributed with a mean of zero. The DE test is performed with a  $\chi^2$ -approximation of the likelihood ratio test.

### MAST

MAST (Finak et al., 2015) uses a hurdle model tailored to scRNA-seq data. It is a two-part GLM that simultaneously models the gene expression rate (how many cells express the gene) by logistic regression and the expression level by Gaussian distribution. The DE testing is then done using the likelihood ratio test.

### T-Test

T-test (Welch, 1947) is a general comparison method that is used to compare the means of two groups. One of the most common assumptions made when doing a t-test is the normality of data distribution. Empirically, scRNA-seq data are highly skewed, but the t-test is known to have a certain robustness against skewness, so it is still worth comparing against other sophisticated methods.

## Nonparametric Methods

### Wilcoxon Rank Sum Test

Wilcoxon rank sum test (Hollander et al., 2013) (also known as Mann-Whitney test) is a nonparametric test that is used to determine whether the two independent samples come from the same distribution. The main idea of the test is to compare the sum of the ranks for the observations which come from different samples.

## DISCUSSION

We have performed a systematic comparison of nine different statistical methods for DE analysis of scRNA-seq data. To get realistic distributional characteristics, three real scRNA-seq data sets are used as the basis for generating the data. A beta-Poisson model-based simulated data set is also performed to assess the performance of each method. The nine methods are evaluated by the type-I error control, the ROC curve and the RDR under both null and alternative hypotheses. Our results show that lowly expressed isoforms are generally the source of strong differences between methods. Most methods except *monocle* have good RDR performances for highly expressed isoforms.

EdgeR and *monocle* tend to produce extremely small p-values for lowly expressed isoforms, leading to many false positives. Notably, these two methods perform very poorly compared to the other methods for top DE isoforms. These results are consistent with other recent studies (Dal Molin et al., 2017; Sonesson and Robinson, 2018). DESeq2, a bulk-cell-based method with a shrinkage procedure, works rather well over all

isoforms on both the real scRNA-seq data and the simulated data. However, DESeq2 is highly conservative for lowly expressed isoforms, so its sensitivity is always lower than the other methods for all three real data sets. The performances of BPSC are comparable to DESeq2 in all analyses but less conservative. Other methods including limmatrend, t-test, Wilcoxon, MAST, and DEsingle perform reasonably in both real and simulated data sets.

To validate our results, we analyzed two extra public real scRNA-seq data sets including one data set with 164 single cells from H7 human cell-line generated by the SMARTer C1 protocol and another big data set contain 2,027 intestinal single cells of mouse from the CEL-Seq protocol. The results in **Figure S6-S8** show the consistency of the comparison analyses for different types of scRNA-seq data for the new small data set. But for the new big data set, *monocle* and DESeq2 show particularly low sensitivity for lowly expressed isoforms in **Figure S6D-S8D**. The details of these data and results are referred to the **Supplementary Material**.

We also investigated further the performances of the DE methods for the group of lowly expressed isoforms. We first checked the relationship between the performance of the Wilcoxon test, one of the most stable DE methods, and the signal strength in different log fold-change (LFC) 1, 2, 3, and 4 using the simulated dataset. Results in the **Figure S4** show that the RDR of Wilcoxon is a function of signal strength where it achieves a higher RDR for the data with a higher LFC. The low signal in the simulated data in **Figure 3D** had made the differences of true RDR for different methods inconspicuous. So we generated another simulation data set using the same procedure described in 3.1 but with a high signal strength  $LFC = \pm 4$ , then applied all 9 methods on the simulated lowly expressed genes. The results (**Figure S5**) confirmed that for the lowly expressed isoforms, DESeq2 is too conservative and consequently loses sensitivity compared to the other methods.

The nine methods compared in this study are selected to cover most statistical models used in recent DE analysis. Although some DE methods are not included in this study, they use similar approach to those we included. For instance, D3E (Delmans and Hemberg, 2016) utilizes beta-Poisson model which is similar to BPSC; SCDE (Kharchenko et al., 2014) models the gene expression values using a mixture of NB distribution for amplification components and a Poisson distribution for dropout events, which is similar to DEsingle; Ballgown (Frazee et al., 2015) is based on the linear modeling strategy which is similar to limma.

The main strengths of our comparison method include (i) the use of three real scRNA-seq data sets in order to capture the true distributional characteristics and the diversity of single-cell data; (ii) the use of the RDR metric for top-rank genes. This is consistent with the data analysis process of identifying the list of interesting genes. In some cases we show that considering the full collection of genes will lead to misleading comparisons; (iii) Separate results of highly and lowly expressed genes, as these two groups have distinct distributions and the methods vary more in their performances for lowly expressed genes. In summary,

performances of DE methods do vary, so we need to pay attention in choosing the method to use, and, at least for highly expressed genes, some methods designed for bulk-cell RNA-seq analysis do not necessarily perform worse than those specifically designed for scRNA-seq data. Finally, as shown the figures, the number of lowly expressed genes is not trivial, so our results also highlight the need for further development of methods to deal with these genes.

## CONCLUSION

There are large differences in the performance of methods for detecting DE in single-cell RNA-seq data. This is driven partly by the expression level of genes. For highly expressed genes, many bulk-cell-based DE methods perform well against single-cell-based methods. But, for lowly expressed genes, the performance of the methods varies, so a careful check of the gene expression level should be made before choosing a DE method in analyses. This is to ensure that the chosen method is appropriate for your data. We found edgeR and monocle to have poor control of false-positives on lowly expressed genes, so we do not recommend these two methods for such genes. DESeq2 tends to be too conservative, so it sacrifices sensitivity for higher specificity. According to the simulation results, BPSC performs well against the other methods, particularly when there is a sufficient number of cells. RDR for top-rank genes is a useful metric for assessing performance of DE methods, sometimes giving different results compared to analysis of the full set of genes. We suggest to be considered in choosing DE methods to use, performances of DE methods in scRNA-seq data strongly depend on the expression level of genes.

## REFERENCES

- Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11, R106. doi: 10.1186/gb-2010-11-10-r106
- Athreya, A. P., Kalari, K. R., Cairns, J., Gaglio, A. J., Wills, Q. F., Niu, N., et al. (2017). Model-based unsupervised learning informs metformin-induced cell-migration inhibition through an AMPK-independent mechanism in breast cancer. *Oncotarget* 8, 27199. doi: 10.18632/oncotarget.16109
- Bland, M. (2013). Do baseline p-values follow a uniform distribution in randomised trials? *PloS One* 8, e76010. doi: 10.1371/journal.pone.0076010
- Chen, Y., Lun, A. T., and Smyth, G. K. (2014). "Differential expression analysis of complex RNA-seq experiments using edgeR," in *Statistical analysis of next generation sequencing data* (New York: Springer), 51–74.
- Chen, Y., Lun, A. T., and Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *FResearch* 5, 1438. doi: 10.12688/f1000research.8987.2
- Dal Molin, A., Baruzzo, G., and Di Camillo, B. (2017). Single-cell RNA-sequencing: assessment of differential expression analysis methods. *Front. In Genet.* 8, 62. doi: 10.3389/fgene.2017.00062
- Delmans, M., and Hemberg, M. (2016). Discrete distributional differential expression (D3E)-a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinf.* 17, 110. doi: 10.1186/s12859-016-0944-6
- Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16, 278. doi: 10.1186/s13059-015-0844-5

## DATA AVAILABILITY STATEMENT

The raw data of two data sets mESC (GSE60749-GPL13112) and NPCs (GSE102934) are published by the original studies and publicly available from NCBI Gene Expression Omnibus repository. The gene expression data of these data sets, MDA-MB-231 data set, simulated data sets and the two supplementary data sets can be found at: <https://github.com/Tianmou/scRNAseq-DE-comparison>.

## AUTHOR CONTRIBUTIONS

YP and TM designed the study. TM, TV, and YP performed the analysis and wrote the manuscript. WD performed the acquisition of MDA-MB-231 data. FG performed a part of simulation studies. All authors read and approved the final manuscript.

## FUNDING

This work was partially supported by funding from the Swedish Cancer Fonden, the Swedish Research Council (VR) and the Swedish Foundation for Strategic Research (SSF).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.01331/full#supplementary-material>

- Frazer, A. C., Perte, G., Jaffe, A. E., Langmead, B., Salzberg, S. L., and Leek, J. T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat. Biotechnol.* 33, 243. doi: 10.1038/nbt.3172
- Ganna, A., Lee, D., Ingelsson, E., and Pawitan, Y. (2014). Rediscovery rate estimation for assessing the validation of significant findings in high-throughput studies. *Briefings In Bioinf.* 16, 563–575. doi: 10.1093/bib/bbu033
- Gong, W., Kwak, I. Y., Pota, P., Koyano-Nakagawa, N., and Garry, D. J. (2018). DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinf.* 19, 220. doi: 10.1186/s12859-018-2226-y
- Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods* Vol. 751 (Hoboken, NJ, USA: John Wiley & Sons).
- Iacono, G., Mereu, E., Guillaumet-Adkins, A., Corominas, R., Cuscó, I., Rodríguez-Esteban, G., et al. (2018). bigSCale: an analytical framework for big-scale single-cell data. *Genome Res.* 28, 878–890. doi: 10.1101/gr.230771.117
- Jaakkola, M. K., Seyedsnasrollah, F., Mehmood, A., and Elo, L. L. (2017). Comparison of methods to detect differentially expressed genes between single-cell populations. *Briefings In Bioinf.* 18, 735–743. doi: 10.1093/bib/bbw057
- Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11, 740. doi: 10.1038/nmeth.2967
- Kumar, R. M., Cahan, P., Shalek, A. K., Satija, R., DaleyKeyser, A. J., Li, H., et al. (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* 516, 56. doi: 10.1038/nature13920
- Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15, R29. doi: 10.1186/gb-2014-15-2-r29

- Law, C. W., Alhamdoosh, M., Su, S., Smyth, G. K., and Ritchie, M. E. (2016). RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *FResearch* 5, 1408. doi: 10.12688/f1000research.9005.2
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8
- Lun, A. T., Chen, Y., and Smyth, G. K. (2016). "It's DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR," in *Statistical Genomics* (New York: Springer), 391–416.
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18, 1509–1517. doi: 10.1101/gr.079558.108
- Miao, Z., and Zhang, X. (2016). Differential expression analyses for single-cell RNA-seq: old questions on new data. *Quant. Biol.* 4, 243–260. doi: 10.1007/s40484-016-0089-7
- Miao, Z., Deng, K., Wang, X., and Zhang, X. (2018). DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* 34, 3223–3224. doi: 10.1093/bioinformatics/bty332
- Murdoch, D. J., Tsai, Y. L., and Adcock, J. (2008). P-values are random variables. *Am. Stat.* 62, 242–245. doi: 10.1198/000313008X332421
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417. doi: 10.1038/nmeth.4197
- Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y. A., and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* 14, 309–315. doi: 10.1038/nmeth.4150
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616
- Sandberg, R. (2014). Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* 11, 22. doi: 10.1038/nmeth.2764
- Soneson, C., and Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* 15, 255–261. doi: 10.1038/nmeth.4612
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511. doi: 10.1038/nbt.1621
- Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen, M., et al. (2016). Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* 32, 2128–2135. doi: 10.1093/bioinformatics/btw202
- Wang, Y. J., Schug, J., Won, K. J., Liu, C., Naji, A., Avrahami, D., et al. (2016). Single-cell transcriptomics of the human endocrine pancreas. *Diabetes* 65, 3028–3038. doi: 10.2337/db16-0405
- Wang, T., Li, B., Nelson, C. E., and Nabavi, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinf.* 20, 40. doi: 10.1186/s12859-019-2599-6
- Welch, B. L. (1947). The generalization of students' problem when several different population variances are involved. *Biometrika* 34, 28–35. doi: 10.2307/2332510
- Zhu, X., Wolfgruber, T. K., Tasato, A., Arisdakessian, C., Garmire, D. G., and Garmire, L. X. (2017). Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *Genome Med.* 9, 108. doi: 10.1186/s13073-017-0492-3
- Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., et al. (2017). Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* 65, 631–643. doi: 10.1016/j.molcel.2017.01.023

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Mou, Deng, Gu, Pawitan and Vu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Mclmpute: Matrix Completion Based Imputation for Single Cell RNA-seq Data

Aanchal Mongia<sup>1</sup>, Debarka Sengupta<sup>1,2\*</sup> and Angshul Majumdar<sup>3</sup>

<sup>1</sup> Department of Computer Science and Engineering, Indraprastha Institute of Information Technology Delhi, New Delhi, India,

<sup>2</sup> Center for Computational Biology, Indraprastha Institute of Information Technology Delhi, New Delhi, India, <sup>3</sup> Department of Electronics and Communications Engineering, Indraprastha Institute of Information Technology Delhi, New Delhi, India

## OPEN ACCESS

### Edited by:

Indrajit Saha,  
National Institute of Technical  
Teachers' Training and Research, India

### Reviewed by:

Kumardeep Chaudhary,  
Icahn School of Medicine at Mount  
Sinai, United States  
Sumit Kumar Bag,  
National Botanical Research Institute  
(CSIR), India  
Yuriy L. Orlov,  
Russian Academy of Sciences, Russia  
Shaoli Das,  
National Institutes of Health (NIH),  
United States

### \*Correspondence:

Debarka Sengupta  
debarka@iitd.ac.in

### Specialty section:

This article was submitted to  
Bioinformatics and Computational  
Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 08 August 2018

**Accepted:** 10 January 2019

**Published:** 29 January 2019

### Citation:

Mongia A, Sengupta D and  
Majumdar A (2019) Mclmpute: Matrix  
Completion Based Imputation for  
Single Cell RNA-seq Data.  
Front. Genet. 10:9.  
doi: 10.3389/fgene.2019.00009

**Motivation:** Single-cell RNA sequencing has been proved to be revolutionary for its potential of zooming into complex biological systems. Genome-wide expression analysis at single-cell resolution provides a window into dynamics of cellular phenotypes. This facilitates the characterization of transcriptional heterogeneity in normal and diseased tissues under various conditions. It also sheds light on the development or emergence of specific cell populations and phenotypes. However, owing to the paucity of input RNA, a typical single cell RNA sequencing data features a high number of dropout events where transcripts fail to get amplified.

**Results:** We introduce mclmpute, a low-rank matrix completion based technique to impute dropouts in single cell expression data. On a number of real datasets, application of mclmpute yields significant improvements in the separation of true zeros from dropouts, cell-clustering, differential expression analysis, cell type separability, the performance of dimensionality reduction techniques for cell visualization, and gene distribution.

**Availability and Implementation:** [https://github.com/aanchalMongia/Mclmpute\\_scRNAseq](https://github.com/aanchalMongia/Mclmpute_scRNAseq)

**Keywords:** scRNA-seq, dropouts, imputation, matrix completion, Nuclear norm minimization

## 1. BACKGROUND AND INTRODUCTION

In contrast to traditional bulk population-based expression studies, single-cell transcriptomics provides more precise insights into the functioning of individual cells. Over the past few years, this powerful tool has brought in transformative changes in the conduct of functional biology (Wagner et al., 2016). With single-cell RNA sequencing (scRNA-seq) we are now able to discover subtypes within seemingly similar cells. This is particularly advantageous for characterizing cancer heterogeneity (Patel et al., 2014; Tirosh et al., 2016), identification of new rare cell type and understanding the dynamics of transcriptional changes during development (Tang et al., 2010; Yan et al., 2013; Biase et al., 2014).

Despite all the goodness, scRNA-seq technologies suffer from a number of sources of technical noise. Most important of these is insufficient input RNA. Due to small quantities transcripts are frequently missed during the reverse transcription step. As a direct consequence, these transcripts are not detected during the sequencing step (Kharchenko et al., 2014). Often times the lowly

expressed genes are the worst hit. Excluding these genes from the analysis may not be the best solution as many of the transcription factors and cell surface markers are sacrificed in this process (van Dijk et al., 2017). Added to that, variability in dropout rate across individual cells or cell types works as a confounding factor for a number of downstream analyses (Sengupta et al., 2016; Li et al., 2017). Hicks et al. (2015) showed, on a number of scRNA-seq datasets, that the first principal components highly correlate with the proportion of dropouts across individual transcriptomes. In summary, there is a standing need for efficient methods to impute scRNA-seq datasets.

Very recently, efforts have been made to devise imputation techniques for scRNA-seq data (Table S6). Most notable of among these are MAGIC (van Dijk et al., 2017), scImpute (Li and Li, 2018), and drImpute (Kwak et al., 2017). MAGIC uses a neighborhood based heuristic to infer the missing values based on the idea of heat diffusion, altering all gene expression levels including the ones not affected by dropouts. On the other hand, scImpute first estimates which values are affected by dropouts based on Gamma-Normal mixture model and then fills the dropout values in a cell by borrowing information of the same gene in other similar cells, which are selected based on the genes unlikely affected by dropout events. The overall performance of scImpute has been shown to be superior to MAGIC. Parametric modeling of single-cell expression is challenging due to our lack of knowledge about possible sources of technical noise and biases (Sengupta et al., 2016). Moreover, there is a clear lack of consensus about the choice of the probability density function. Another method, DrImpute, repeatedly identifies similar cells based on clustering and performs imputation multiple times by averaging the expression values from similar cells, followed by averaging multiple estimations for final imputation. We propose mcImpute (Figure 1), an imputation algorithm for scRNA-seq data which models gene expression as a low-rank matrix and sprouts in values in place of dropouts in the process of recovering the full gene expression data from sparse single-cell data. This is done by applying soft-thresholding iteratively on singular values of scRNA-seq data. One of the salient features of mcImpute is that it does not assume any distribution for gene expression.

We first evaluate the performance of mcImpute in separating “true zero” counts from dropouts on single-cell data of myoblasts (Trapnell et al., 2014) (We call it Trapnell dataset). On the same dataset, we assess the impact of imputation on differential genes prediction. We further investigate mcImpute’s ability to recover artificially planted missing values in a single cell expression matrix of mouse neurons (Usoskin et al., 2015). Accurate imputation should enhance cell type identity i.e., the transcriptomic similarity between cells of identical type. We, therefore, quantify cell type separability as a metric and assess its improvement. In addition to these, we also test the impact of imputation on cell clustering. Four independent datasets Zeisel (Zeisel et al., 2015), Jurkat-293T (Zheng et al., 2017), Preimplantation (Yan et al., 2013) and Usoskin (Usoskin et al., 2015), for which cell type annotations are available and another dataset, Trapnell et al. (2014) for which bulk RNA-seq data has been provided (required for validation of differential genes

prediction and separation of “true zeros” from dropouts), are used for this purpose. McImpute clearly serves as a crucial tool in the scRNA-seq pipeline by significantly improving all the above-mentioned metrics and outperforming the state-of-the-art imputation methods in the majority of experimental conditions.

With the advent of droplet-based, high-throughput technologies (Macosko et al., 2015; Zheng et al., 2017), library depth is being compromised to curb the sequencing cost. As a result, scRNA-seq datasets are being produced with an extremely high number of dropouts. We believe that mcImpute’s great performance, will provide an adequate solution for the dropouts problem.

## 2. RESULTS

We performed computational experiments to evaluate the efficacy of our proposed imputation technique comparing mcImpute with a number of existing imputation methods for single cell RNA data: scImpute, drImpute, and MAGIC.

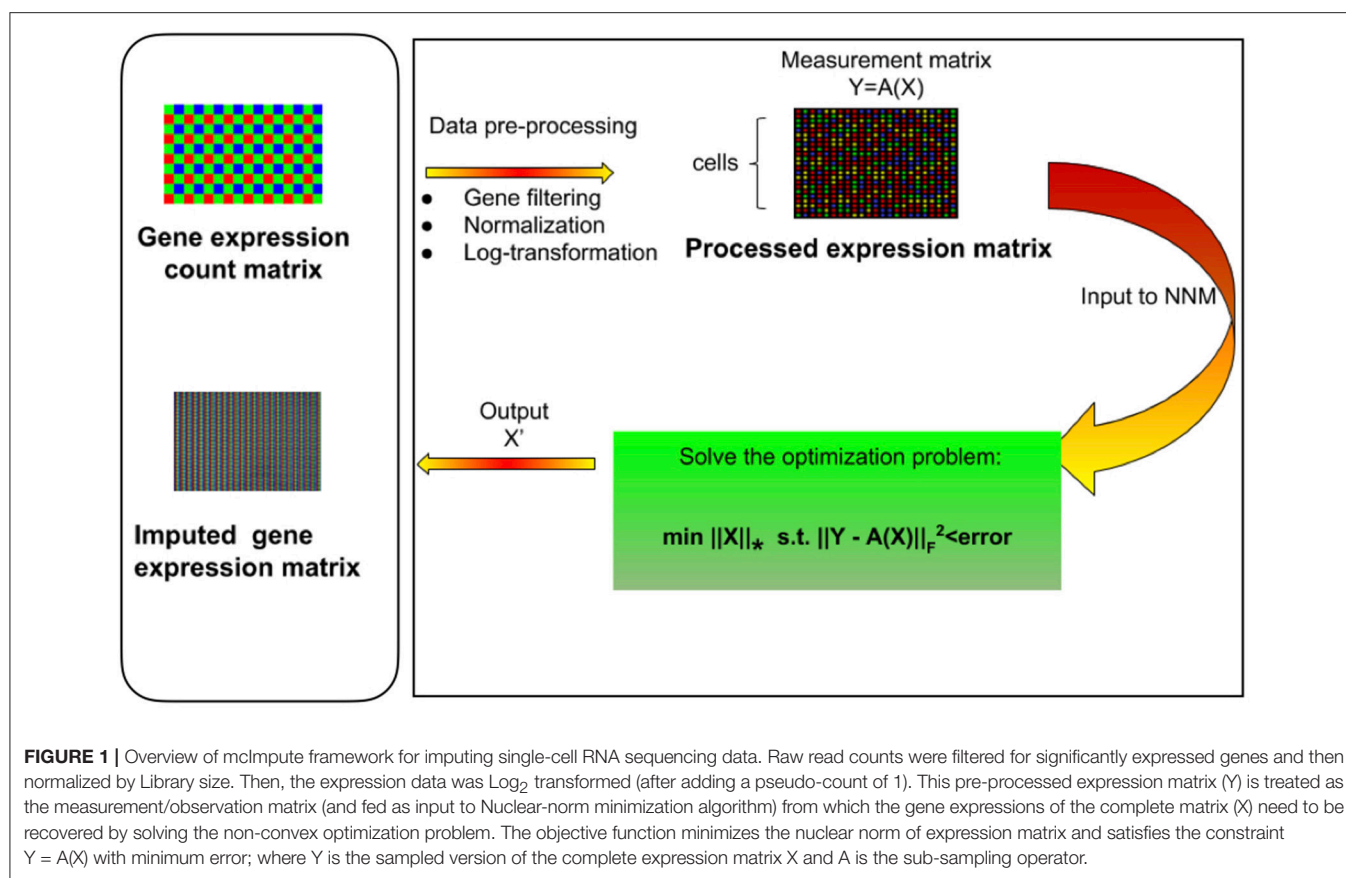
### 2.1. Dropouts vs. True Zeros

The inflated number of zero counts in scRNA-seq data could either be biologically driven or due to lack of measurement sensitivity in sequencing. The transcript which is not detected because of failing to get amplified in the sequencing step essentially corresponds to a “false zero” in the finally observed count data and needs to be imputed. A reasonable imputation strategy which has this discriminating property should keep the “true zero” counts (where the genes are truly expressed and have no transcripts from the beginning) untouched, while at the same time attempt to recover the dropouts.

The goodness of an imputation strategy can be formally confirmed by observing two factors. First, whether the imputation method is able to impute the true zero counts in the expression data as is or not; Second, if it can fill-in the dropouts with biologically meaningful expression counts or not; showing an increasing difference between the zero counts observed in unimputed data and the imputed one with expression amplification.

We investigate the performance of mcImpute in distinguishing “true zero” counts from dropouts on Trapnell data (Trapnell et al., 2014), for which the bulk-counterpart was available and hence, we could pull out low-to-medium expression genes from the corresponding bulk data for validation. Of note, to differentiate between the “true” and “false” zeros, we have used the matched bulk-expression profiles; as it is a well-known fact that bulk-RNA seq data has limited or no dropouts events as the corresponding experiments involve millions of cells. The fraction of zero counts was observed for genes with expression ranging from zero to 500 for unimputed and imputed gene-expression data. It should be noted that an imputed count value ranging from 0 to 0.5 is taken as an imputed zero, rendering minor flexibility to all imputation techniques.

Given the nature of this analysis, gene filtering in single cell expressions has been skipped. DrImpute could not be taken into



account since we could not programmatically mute the gene filtering step in its pipeline.

We observe (Figure 2A, Table S1) that with low expression genes, all imputation strategies successfully impute the “true zeros” while, as the gene expression amplifies, un-imputed matrix still exhibits large fraction of zeros, which essentially correspond to dropouts and only mcImpute and scImpute are able to curtail the fraction of zeros, thus recovering the dropouts back. As can be observed, MAGIC although successfully imputes the “true zeros”; it fails to recover most of the dropouts in the expression data.

## 2.2. Improvement in Clustering Accuracy

A correct interpretation of single-cell expression data is contingent on the accurate delineation of cell types. Bewildering level of dropouts in scRNA-seq data often introduces batch effect, which inevitably traps the clustering algorithm. A reasonable imputation strategy should fix these issues to a great extent. In a controlled setting, we, therefore, examined if the proposed method enhanced clustering outcomes. For this, we ran  $K$ -means on first 2 principal component genes of log-transformed expression profiles featured in each dataset (Figure S5). Since the prediction from this clustering algorithm tends to change with the choice of initial centroids, which are chosen at random, we analyze the results on 100 runs of  $k$ -means to get reliable and robust results. We set the number of annotated cell types as the

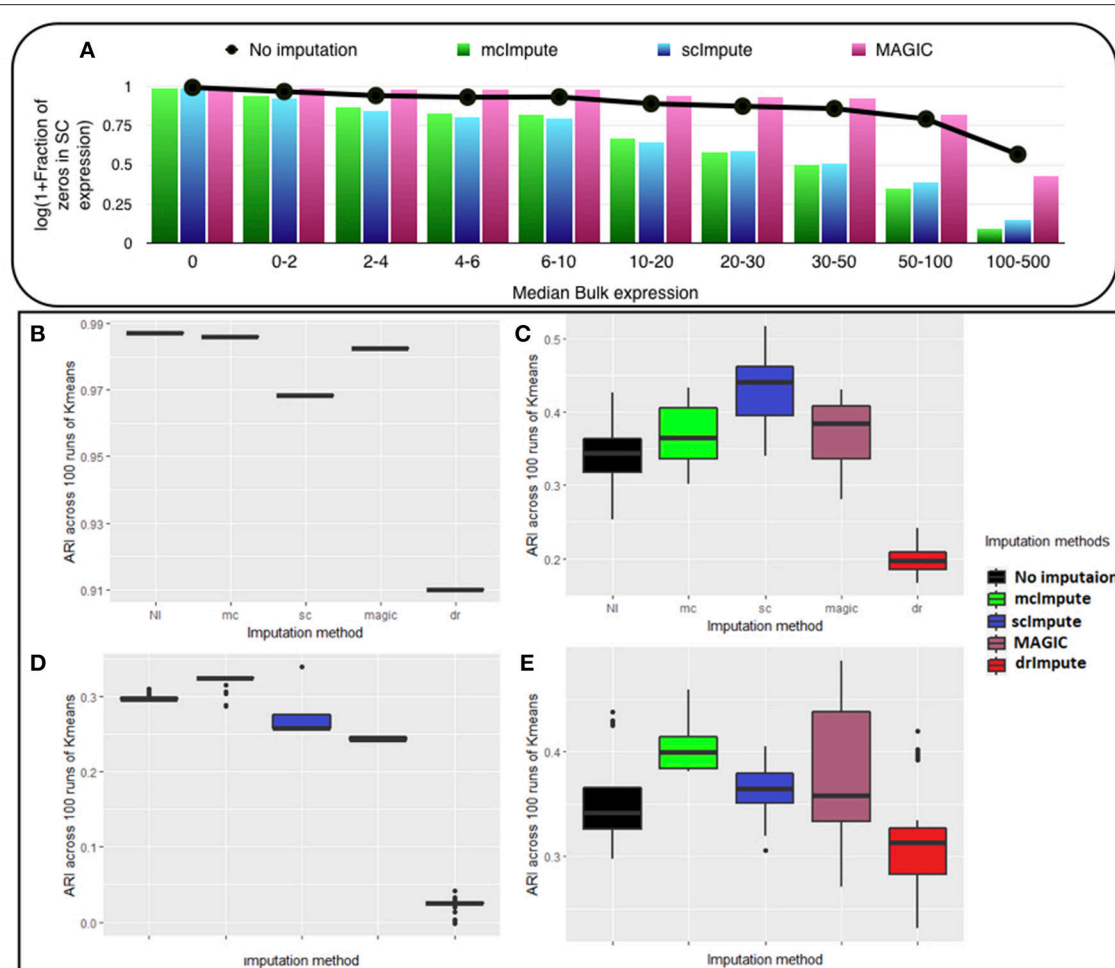
value of  $K$  for every data. Adjusted Rand Index (ARI) was used to measure the correspondence between the clusters and the prior annotations.

McImpute based re-estimation best separates the four groups of mouse neural single cells from Usoskin dataset and brain cells from Zeisel dataset, and clearly shows comparable improvement on other datasets too (Figures 2B–E, Table S2). The striking difference between Jurkat and 293T cells made them trivially separable through clustering, leading to same ARI across all 100 runs. Still, mcImpute was able to better maintain the ARI in comparison to other imputation methods.

## 2.3. Matrix Recovery

In this set of experiments, we study the choice of matrix completion algorithm – matrix factorization (MF) or nuclear norm minimization (NNM). Both the algorithms have been explained in section Materials and Methods.

The experiments are carried out on the processed Usoskin dataset (Usoskin et al., 2015). We artificially removed some counts at random (sub-sampling) in the data to mimic dropout cases and used our algorithms (MF and NNM) to impute the missing values. (Figures 3A–C) and Table S3 show the variation of Normalized Mean Squared Error (NMSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to compare our two methods for different sub-sampling ratios. This is the



**FIGURE 2 |** McImpute shows remarkable improvement in separation of “true zeros” from dropouts and clustering of single cells **(A)** Separation of “true zeros” from dropouts: plot showing fraction of zero counts (values between 0 and 0.5) in single cell expression matrix against the median bulk expression. The genes are divided into 10 bins based on median bulk genes expression (first bin corresponds to zero expression genes) **(B–E)** Boxplots showing the distribution of ARI calculated on 100 runs of k-means clustering algorithm on first two principal components of single cell expression matrix for datasets **(B)** Jurkat-293T **(C)** Preimplantation **(D)** Usoskin, and **(E)** Zeisel.

standard procedure to compare matrix completion algorithms (Keshavan et al., 2010; Marjanovic and Solo, 2012).

We are showing the results for Usoskin dataset, but we have carried out the same analysis for other datasets and the conclusion remained the same. We find that the nuclear norm minimization (NNM) method performs slightly better than the matrix factorization (MF) technique; so we have used NNM as the workhorse algorithm behind McImpute.

## 2.4. Improved Differential Genes Prediction

Optimal imputation of expression data should improve the accuracy of differential expression (DE) analysis. It is a standard practice to benchmark DE calls made on scRNA-Seq data against calls made on their matching bulk counterparts (Kharchenko et al., 2014). To this end, we used a dataset of myoblasts, for which matching bulk RNA-Seq data were also available (Trapnell et al., 2014). For simplicity, this dataset has been referred to as

the Trapnell dataset. DE and non-DE genes were identified using edgeR (Zhou et al., 2014) package in R.

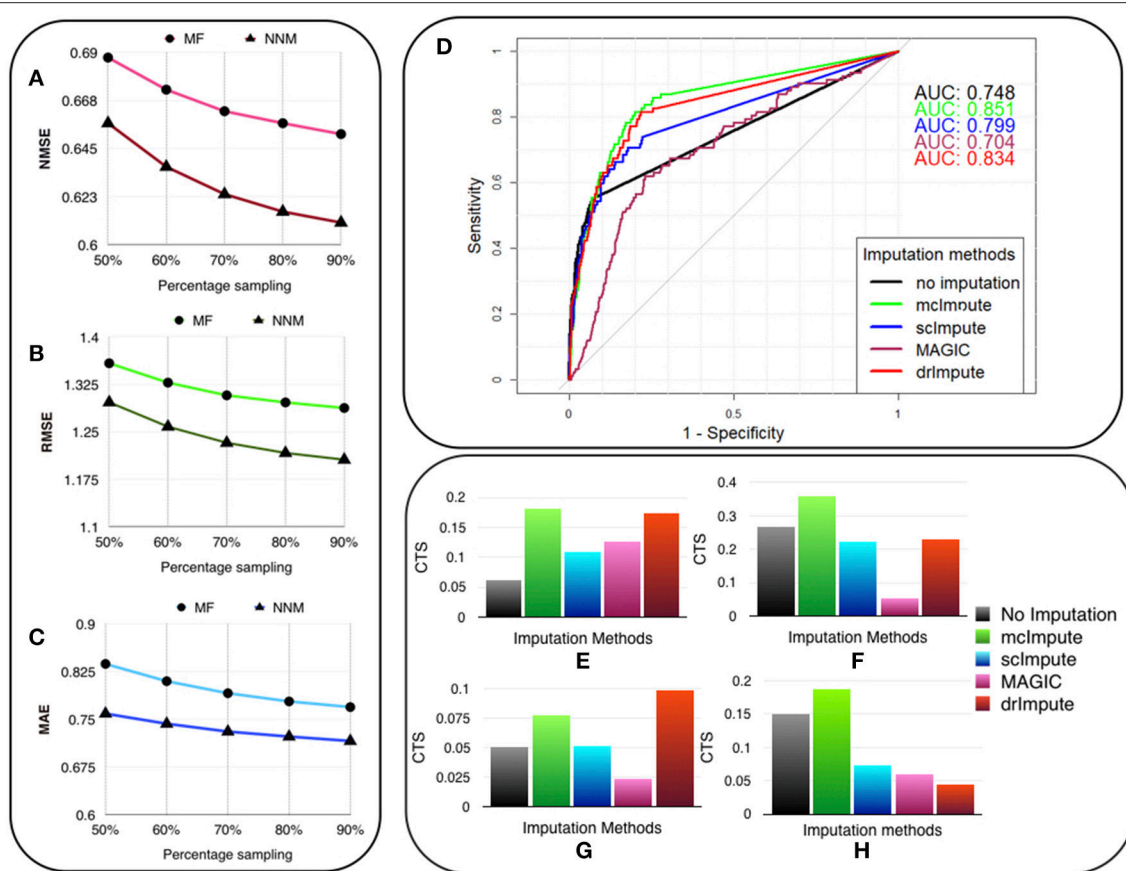
We used the standard Wilcoxon Rank-Sum test for identifying differentially expressed genes from matrices imputed by various methods. Congruence between bulk and single cell-based DE calls were summarized using the Area Under the Curve (AUC) values yielded from the Receiver Operating Characteristic (ROC) curves (**Figure 3D**). Among all the methods mcImpute performed best with an AUC of 0.85.

For each method, the AUC value was computed on the identical set of ground truth genes. We had to make an exception only for drImpute as it applies the filter to prune genes in its pipeline. Hence AUC value for drImpute was computed based on a smaller set of ground truth genes.

## 2.5. Improvement in Cell Type Separability

Downstream analysis becomes much easier if expression similarities between cells of identical type are considerably





**FIGURE 3 |** McImpute recovers the original data from their masked version with low error, performs best in prediction of differentially expressed genes and significantly improves CTS score. Variation of (A) NMSE, (B) RMSE, and (C) MAE with sampling ratio using MF (Matrix factorization) and NNM (Nuclear norm minimization) on Usoskin dataset showing NNM performing better than MF algorithm. (D) ROC curve showing the agreement between DE genes predicted from scRNA and matching bulk RNA-Seq data (Trapnell et al., 2014). DE calls were made on expression matrix imputed using edgeR. (E–H) 2D-Axis bar plot depicting improvement in Cell type separabilities between (E) Jurkat and 293T cells from Jurkat-293T dataset; (F) 8cell and BXC cell types from Preimplantation dataset; (G) NP and NF cells from Usoskin dataset; and (H) S1pyramidal and Ependymal from Zeisel dataset. Refer Table S4 for absolute values.

higher than that of cells coming from different subpopulations. To this end, we define the cell-type separability score as follows:

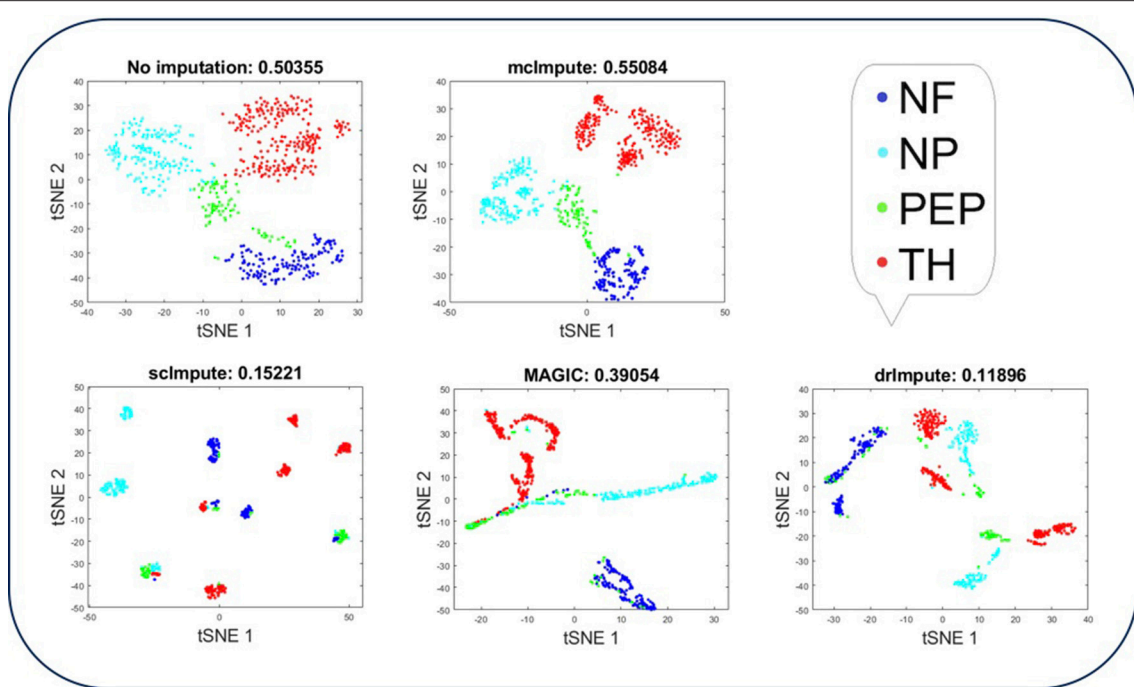
For any two cell groups, we first find the median of Spearman correlation values computed for each possible pair of cells within their respective groups. We call the average of the median correlation values the intra-cell type scatter. On the other hand, inter-cell type scatter is defined as the median of Spearman correlation values computed for pairs such that in each pair, cells belong to two different groups. The difference between the intra-cell scatter and inter-cell type scatter is termed as the cell-type separability (CTS) score. We computed CTS scores for two sample cell-type pairs from each dataset. In more than 80 % (13 out of 16) of test cases, mclmpute yielded significantly better CS values (Figures 3E–H, Table S4).

## 2.6. Cell Visualization

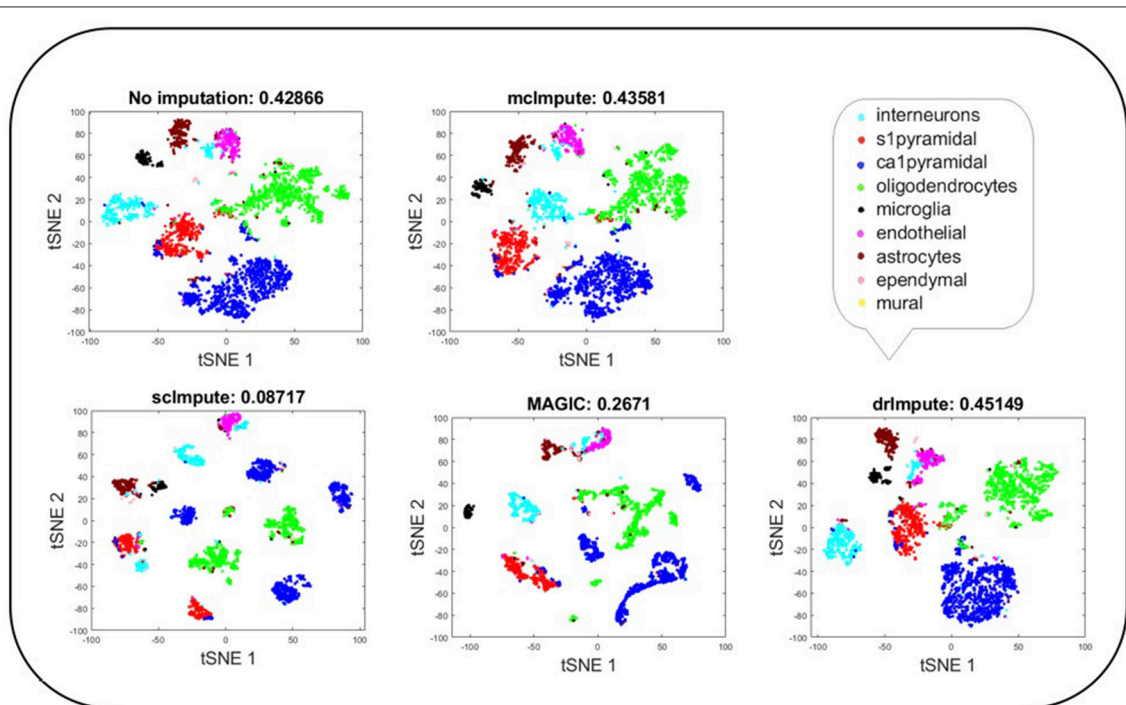
Representing scRNA-seq data visually would involve reducing the gene-expression matrix to a lower dimensional space and

then plotting each cell transcriptome in that reduced two or three-dimensional space. Two well-known techniques for dimensionality reduction are PCA and t-SNE (Holland, 2008; Maaten and Hinton, 2008). It has been shown that t-Distributed Stochastic Neighbor Embedding (t-SNE) is particularly well suited and effective for the visualization of high-dimensional datasets (Liu et al., 2017). So, we use t-SNE (Figures 4, 5) on Usoskin and Zeisel expression matrices to explore the performance of dimensionality reduction, both without and with imputation. The cells are visualized in 2-dimensional space, coloring each subpopulation by its annotated group, both before and after imputation. To quantify the groupings of cell transcriptomes, we use an unsupervised clustering quality metric, silhouette index. The average silhouette values for each method have been shown in the plot titles (Figures 4, 5 and Figures S3, S4).

T-SNE analysis depicts that mclmpute brings all four groups of mouse neural cells from Usoskin dataset closest to each other in comparison to other methods and performs fairly well, competing with drImpute on Zeisel dataset too.



**FIGURE 4 |** Plot showing t-SNE visualization and average silhouette values for Usoskin dataset before and after imputation. McImpute improves the visual distinguishability the most for all groups of mouse neural single cells amongst all imputation strategies. The neuronal types were defined as neurofilament containing (NF), non-peptidergic nociceptors (NP), peptidergic nociceptors (PEP), and tyrosine hydroxylase containing (TH).



**FIGURE 5 |** Plot showing t-SNE visualization and average silhouette values for Zeisel dataset before and after imputation. Both mcImpute and drImpute bring brain cells closer, at the same time maintaining the structure of gene expressions.

## 2.7. Improvement in Distribution of Genes

It has been shown that for single-cell gene expression data, in the ideal condition all genes should obey  $CV = mean^{-1/2}$  (Klein et al., 2015) (CV: coefficient of variation), following a Poisson distribution as depicted by the green diagonal line (**Figures 6, 7**). This is because individual transcripts are sampled from a pool of available transcripts for CEL-Seq. This accounts for technical noise component which obeys Poissonian statistics (Grün et al., 2014), and thus the CV is inversely proportional to the square root of the mean. Since this result has only been shown for single-cell data with transcript numbers, this experiment has not been analyzed for Jurkat-293T and Zeisel datasets for which the individual RNA molecules were counted using unique molecular identifiers (UMIs).

We model CV as a function of mean expression for all genes to analyze how various imputation methods affect the relationship between them. The results (**Figures 6, 7**) show that both mcImpute and drImpute succeed to restore the relationship between CV and mean to a great extent (improving the dependency of the CV on the mean expression level to be more consistent with Poissonian sampling noise), while others do not.

## 3. DISCUSSION

Single-cell RNA seq technologies have opened up numerous possibilities for analysis at the single-cell resolution. But, low amount of starting RNA is a major limitation of the technology which results in frequent missing of transcripts in the reverse transcription step (dropout events). This dropout problem in single-cell RNA-seq data makes the expression matrix highly sparse; which in turn hinders the downstream analysis.

To overcome the dropout problem in single-cell data, we take motivation from various areas of applied sciences (including computer vision Tomasi and Kanade, 1992, control Mesbahi and Papavasilopoulos, 1997, machine learning Abernethy et al., 2006; Amit et al., 2007; Argyriou et al., 2007, etc) where recovery of an unknown low-rank matrix from very limited information is of interest. The problem is akin to that of recommendation systems (e.g. in Netflix movie recommendations and Amazon product recommendations) (Bell and Koren, 2007; Bennett and Lanning, 2007; SIGKDD, 2007), where there is a database of ratings given by users to movies/products. Since the users typically rate only a small subset of items, not all the ratings are available; which makes the user-movie rating matrix sparse. Also, the matrix is assumed to be of low-rank because there are not too many independent parameters on which the users generally rate the movie. The objective is to estimate the ratings of all the users on all the movies. If the new movie rating predictions can be done accurately, recommendation accuracy increases. There is a pretty straightforward link between both the Netflix problem and dropout problems. Therefore, imputation to single-cell expression matrix can be efficiently performed by Low-rank approximation. (Koren et al., 2009; Majumdar and Ward, 2011).

One could argue about the low-rank origin of the gene expression data. It should be noted that numerous studies have

suggested that genes do not work in isolation (Staiger et al., 2013), but as part of a complex regulatory network (Silver et al., 2013). This inter-dependency has been analyzed in the form of associated network structures (Xiong et al., 2005; Gill et al., 2010) and is best reflected by the gene-gene correlations (Weckwerth et al., 2004; Klebanov and Yakovlev, 2007; Reynier et al., 2011; Najafov and Najafov, 2018). It is so believed that such high levels of correlation are caused by sharing of regulatory programs among different genes (Ye et al., 2013). Also, it has previously been shown that a small number of interdependent biophysical functions trigger the functioning of transcription factors, which in turns influence the expression levels of genes, resulting in a highly correlated data matrix (Kapur et al., 2016). On the other hand, cells coming from same tissue source also lie on differential grades of the variability of a limited number of phenotypic characteristics. Therefore, it is just to assume that the gene expression values lie on a low-dimensional linear subspace and the data matrix thus formed may well be thought as a low-rank matrix.

We attempt to give another mathematical justification on the Low-rank assumption of the gene-expression in **Figure S2** by showing that the maximum information of the expression-data is held in its first few singular values; hence the rank of the expression matrix (number of non-zero singular values) should be low.

In specific, we used Nuclear Norm-based Matrix Completion for imputing single-cell RNA seq data. The algorithm models the single-cell gene expression as a low-rank matrix and recovers the full gene expression from partial information by thresholding the singular values of expression matrix iteratively. The recovery process sprouts-in appropriate expressions in place of dropouts; keeping the biologically silent expression values intact.

Apart from taking care of biologically silent genes, the proposed algorithm performs competitively with the state-of-the-art methods in improving the clustering accuracy of cells, identifying differentially expressed genes, enhancing cell type separability, improving the dimensionality reduction, etc.

Our method is particularly suitable for single-cell data since it does not assume anything about the statistical property of the expression or the dropouts and can be seamlessly incorporated into the single-cell analysis pipeline. We have also demonstrated that our method clearly distinguishes between biological and technical silencing.

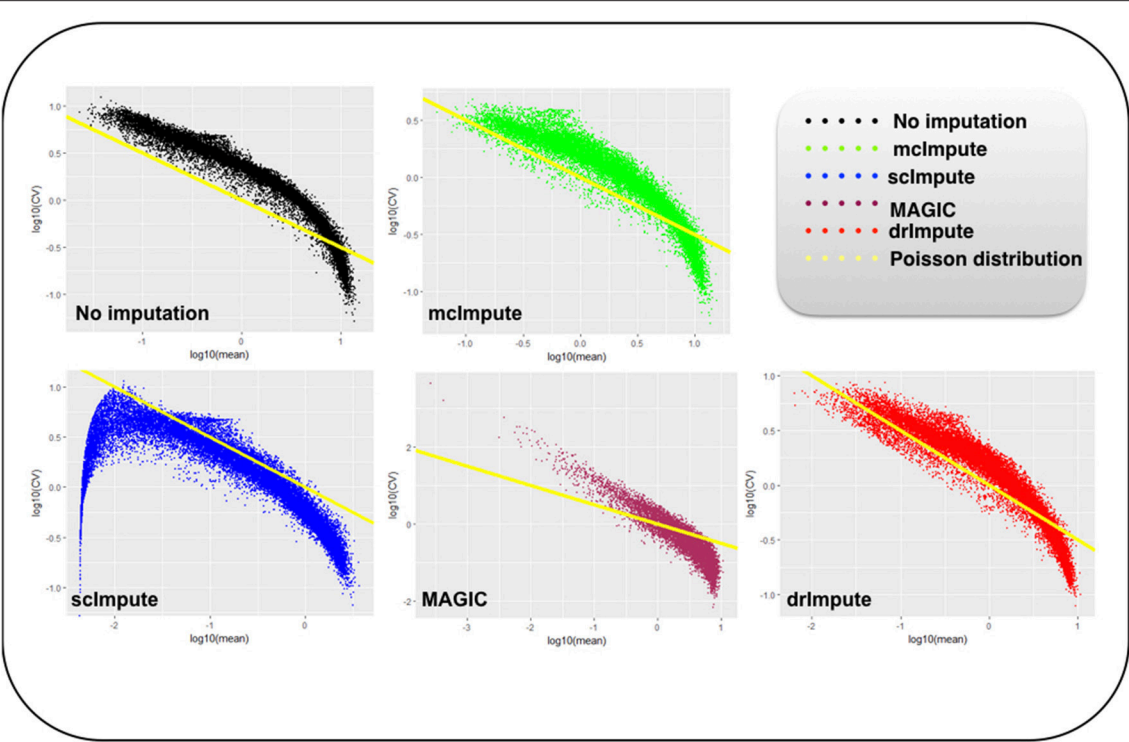
The algorithm has some scope of improvement when it comes to handling scRNA-seq datasets with large sample sizes. As can be seen in **Table S5**, the running time of our algorithm is comparatively more than that of MAGIC and drImpute; although much less than that of scImpute.

## 4. DATA AND METHODS

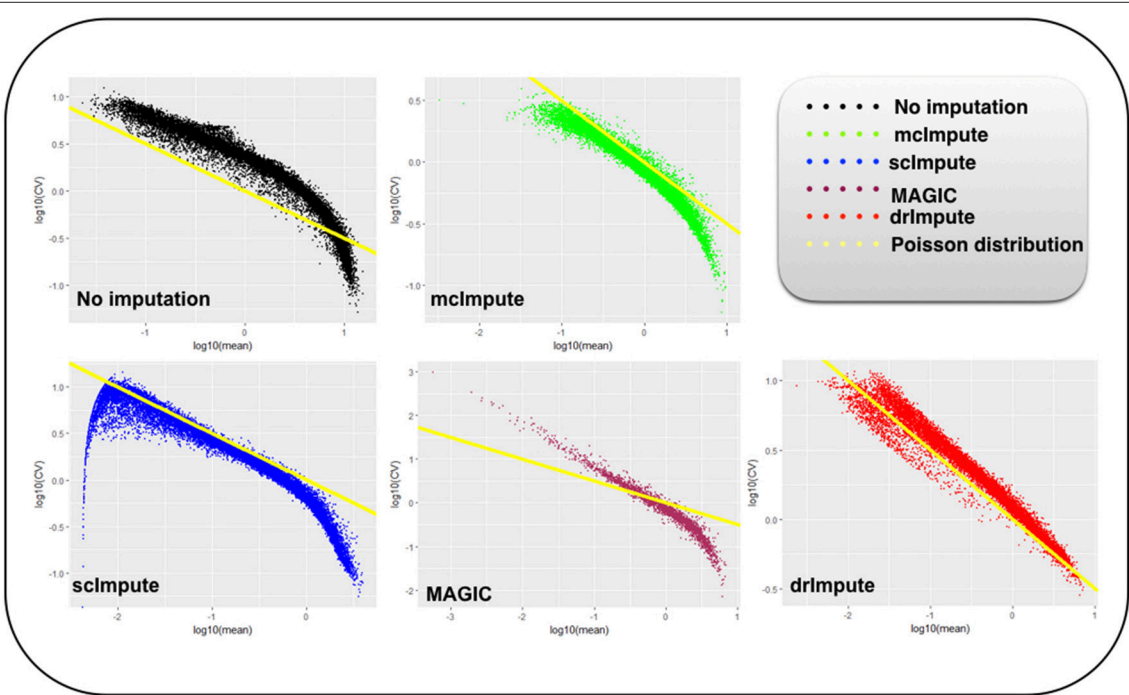
### 4.1. Dataset Description

We used five scRNA-seq datasets from four different studies for performing various experiments (**Table S7**).

- **Jurkat-293T:** This dataset contains expression profiles of Jurkat and 293T cells, mixed *in vitro* at equal proportions



**FIGURE 6 |** Plot showing  $\log_{10}(\text{CV})$  vs.  $\log_{10}(\text{mean})$  relationship between genes for Preimplantation dataset before and after imputation.



**FIGURE 7 |** Plot showing  $\log_{10}(\text{CV})$  vs  $\log_{10}(\text{mean})$  relationship between genes for Usoskin dataset before and after imputation.



(50:50). All  $\sim 3,300$  cells of this data are annotated based on the expressions of cell-type specific markers (Zheng et al., 2017). Cells expressing CD3D are assigned Jurkat, while those expressing XIST are assigned 293T. This dataset is also available at 10x Genomics website ([https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/jurkat:293t\\_50:50](https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/jurkat:293t_50:50)).

- **Preimplantation:** This is an scRNA-seq data of mouse preimplantation embryos. It contains expression profiles of  $\sim 300$  cells from zygote, early 2-cell stage, middle 2-cell stage, late 2-cell stage, 4-cell stage, 8-cell stage, 16-cell stage, early blastocyst, middle blastocyst, and late blastocyst stages. The first generation of mouse strain crosses was used for studying monoallelic expression. We downloaded the count data from Gene Expression Omnibus (GSE45719) (Yan et al., 2013).
- **Zeisel:** Quantitative single-cell RNAseq has been used to classify cells in the mouse somatosensory cortex (S1) and hippocampal CA1 region based on 3005 single cell transcriptomes (Zeisel et al., 2015). Individual RNA molecules were counted using unique molecular identifiers (UMIs) and confirmed by single-molecule RNA fluorescence *in situ* hybridization (FISH). A divisive biclustering method based on sorting points into neighborhoods (SPIN) was used to discover molecularly distinct, 9 major classes of cells. Raw data is available under the accession number GSE60361.
- **Usoskin:** This data of mouse neurons (Usoskin et al., 2015) was obtained by performing RNA-Seq on 799 dissociated single cells dissected from the mouse lumbar dorsal root ganglion (DRG) distributed over a total of nine 96-well plates. After Principal component analysis (PCA) of expression magnitudes across all cells and genes, 622 cells were classified as neurons, 68 cells had an ambiguous assignment and 109 cells were non-neuronal. We take into account the 622 neuronal clusters of mouse lumbar DRG- neurofilament containing (NF), non-peptidergic nociceptors (NP), peptidergic nociceptors (PEP), and tyrosine hydroxylase containing (TH). RPM normalized counts are available under the accession number GSE59739.
- **Trapnell:** This is an scRNA-seq data of primary human myoblasts (Trapnell et al., 2014). Differentiating myoblasts were cultured and cells were dissociated and individually captured at 24-h intervals. 50–100 cells at each of the four time points were captured following serum switch using the Fluidigm C1 microfluidic system. This data is available at Gene Expression Omnibus under the accession number GSE52529. Of note, this dataset has been used for the experiments which require the Bulk-counterpart of the gene-expression data i.e., “Dropout vs true-zeros” and “Differential genes prediction.”

## 4.2. Data Preprocessing

Steps involved in preprocessing of raw scRNA-seq data are enumerated below.

- **Data filtering:** It is ensured that data has no bad cells and if a gene was detected with  $\geq 3$  reads in at least 3 cells we considered it expressed. We ignored the remaining genes.
- **Library-size Normalization:** Expression matrices were normalized by first dividing each read count by the total

counts in each cell, and then by multiplying with the median of the total read counts across cells.

- **Log Normalization:** A copy of the matrices were  $\log_2$  transformed following the addition of 1 as pseudo-count.
- **Imputation:** Further, log transformed expression matrix was used as input to mcImpute. The algorithm returns imputed log transformed matrix, normalized matrix (after applying reverse of log operation on imputed log-transformed expressions), and the count matrix after imputation.

A brief overview of the complete mcImpute pipeline has been shown in Figure 1.

## 4.3. Low-Rank Matrix Completion: Definition

Our problem is to complete a partially observed gene expression matrix  $X$  where columns represent genes and rows, individual cells. The complete matrix is constituted by the known and the yet unknown values. We can assume that the single cell data that we have acquired,  $Y$  is a sampled version of the complete expression matrix  $X$ . Mathematically, this is expressed as,

$$Y = A(X) \quad (1)$$

Here  $A$  is the sub-sampling operator. It is a binary mask that has 0's where the counts of complete expression data  $X$  have not been observed and 1's where they have been. The values of  $A$  are element-wise multiplied to the complete expression matrix  $X$  so that  $Y$  (the sub-sampled data) is a sparse representation of  $X$  and has expression values only at positions where gene expression is observed. Our problem is to recover  $X$ , given the observations  $Y$ , and the sub-sampling mask  $A$ . It is known that  $X$  is of low-rank.

It should be noted that matrix completion is a well studied framework. In this work, we consider two algorithms for efficient imputation of scRNA-seq expression data: Matrix factorization (Koren et al., 2009) and Nuclear norm minimization?

## 4.4. Matrix Factorization

Matrix factorization is the most straightforward way to address the low-rank matrix completion problem; it has previously been used for finding lower dimensional decompositions of matrices (Lee and Seung, 2001). Say  $X$  is of dimensions  $m \times n$ , but is known to have a rank  $r$  ( $< m, n$ ). In that case, one can express  $X_{m \times n}$  as a product of two matrices  $U_{m \times r}$  and  $V_{r \times n}$ . Therefore the complete problem (1) can be formulated as,

$$Y = A(X) = A(UV) \quad (2)$$

Estimating  $U$  and  $V$  from (2) tantamount to recovering  $X$ . The two matrices  $U$  and  $V$  can be solved by minimizing the Frobenius norm of the following cost function.

$$\min_{U,V} \|Y - A(UV)\|_F^2 \quad (3)$$

Since this is a bi-linear problem, one cannot guarantee global convergence. However, it usually works in practice. It has been used for solving recommender systems problems (Koren et al., 2009), where (3) was solved using stochastic

gradient descent (SGD). SGD is not an efficient techniques and requires tuning of several parameters. In this work, we will solve (3) in a more elegant fashion using Majorization-Minimization (MM) (Sun et al., 2017). The basic MM approach and its geometrical interpretation has been diagrammatically represented (Figure S1). It depicts the solution path for a simple scalar problem but essentially captures the MM idea.

For our given problem, the cost function to be minimized is given as  $J(X) = \|Y - A(X)\|_F^2$ ; the majorization step basically decouples the problem (from  $A$ ), so that we can solve the optimization problem by solving

$$\min_{U,V} \|B - UV\|_F^2 \quad (4)$$

where  $B_{k+1} = X_k + \frac{1}{a}A^T(Y - A(X_k))$  at each iteration  $k$ . Here,  $X_k$  is the matrix at iteration  $k$  and  $a$  is a scalar parameter in the MM algorithm.

This (4) is solved by alternating least squares (Hastie et al., 2015), i.e., while updating  $U$ ,  $V$  is assumed to be constant and while updating  $V$ ,  $U$  is assumed to be constant.

$$U_k \leftarrow \min_U \|B - U_{k-1}V_{k-1}\|_F^2 \quad (5)$$

$$V_k \leftarrow \min_V \|B - U_kV_{k-1}\|_F^2 \quad (6)$$

Since the log-transformed input (with pseudo count added) expressions would never be negative, we have imposed a non-negativity constraint on the recovered matrix  $X$ , so that it does not contain any negative values.

The matrix factorization algorithm has been summarized in Algorithm 1. The initialization of factor  $V$  is done by keeping  $r$  right singular vectors of  $X$  in  $V$  obtained by performing singular value decomposition (SVD) of  $X$ , where  $r$  is the approximate rank of the expression matrix to be recovered.

---

#### Algorithm 1 Matrix completion using matrix factorization

---

```

1: procedure MATRIX-FACTORIZATION( $Y, A, r$ )
2:   Initialize:  $X = \text{random}$ ,  $a$ ,  $V$  (SVD initialization),  $k$  and  $l$ .
3:   For loop 1, iterate ( $k$ )
4:      $B_k = X_{k-1} + \frac{1}{a}A^T(Y - A \circ X_{k-1})$ 
5:     For loop 2, iterate ( $l$ )
6:        $U_l \leftarrow \min_U \|B_k - U_{l-1}V_{l-1}\|_F^2$ 
7:        $V_l \leftarrow \min_V \|B_k - U_lV_{l-1}\|_F^2$ 
8:     End loop 2
9:      $X_k = U_kV_k$ 
10:     $X_k \leftarrow X_k^+$ 
11:  End loop 1
```

---

## 4.5. Nuclear Norm Minimization

The problem depicted in (3) is non-convex. Hence, there is no guarantee for global convergence. Also one needs to know the approximate rank of the matrix  $X$  in order to solve it, which

is unknown in this case. To combat this issues, researchers in applied mathematics and signal processing proposed an alternative solution. They would directly solve the original problem (1) with a constraint that the solution is of low-rank. This is mathematically expressed as,

$$\min_X \text{rank}(X) \text{ such that } Y=A(X) \quad (7)$$

However, this turns out to be NP hard problem with doubly exponential complexity. Therefore, studies in matrix completion (Candes and Recht, 2009; Candès and Tao, 2010) proposed relaxing the NP hard rank minimization problem to its closest convex surrogate: nuclear norm minimization.

$$\min_X \|X\|_* \text{ such that } Y=A(X) \quad (8)$$

Here  $\|\cdot\|_*$  is the nuclear norm and is defined as the sum of singular values of data matrix  $X$ . It is the  $l_1$  norm of the vector of singular values of  $X$  and is the tightest convex relaxation of the rank of matrix, and therefore its ideal replacement.

This is a semi-definite programming (SDP) problem. Usually its relaxed version (Quadratic Program) is solved (Candès and Plan, 2010) with the unconstrained Lagrangian version.

$$\min_X \|Y - A(X)\|_F^2 + \lambda \|X\|_* \quad (9)$$

Here,  $\|\cdot\|_*$  is the nuclear norm and  $\lambda$  is called the Lagrange multiplier. The problem (9) does not have a closed form solution and needs to be solved iteratively.

To solve (9), we invoke MM once more. Here  $J(X) = \|Y - A(X)\|_F^2 + \lambda \|X\|_*$ , we can express (9) in the following fashion in every iteration  $k$

$$\min_X \|B - X\|_F^2 + \lambda \|X\|_* \quad (10)$$

where  $B_{k+1} = X_k + \frac{1}{a}A^T(Y - A(X_k))$ .

Using the inequality  $\|Z_1 - Z_2\|_F \geq \|s_1 - s_2\|_2$ , where  $s_1$  and  $s_2$  are singular values of the matrices  $Z_1$  and  $Z_2$  respective, we can solve the following instead of solving the minimization problem (10).

$$\min_{s_X} \|s_B - s_X\|_2^2 + \lambda \|s_X\|_1 \quad (11)$$

Here,  $s_B$  and  $s_X$  are the singular values of  $B$  and  $X$ , respectively and  $\|s_X\|_1$  is the  $l_1$  norm or the sum of absolute values of  $s_X$ . It has been shown that problem (10) is minimized by soft thresholding the singular values with threshold  $\lambda/2$ . The optimal update is given by

$$s_X = \begin{cases} s_B + \lambda/2 & \text{when } s_B \leq -\lambda/2 \\ 0 & \text{when } |s_B| \leq \lambda/2 \\ s_B - \lambda/2 & \text{when } s_B \geq \lambda/2 \end{cases} \quad (12)$$

or more compactly by

$$s_X = \text{soft}(s_B, \lambda/2) = \text{sign}(s_B) \max(0, |s_B| - \lambda/2) \quad (13)$$

**Algorithm 2** Matrix completion via nuclear norm minimization

```

1: procedure MATRIX-NNM( $Y, A$ )
2:   Initialize:  $X = \text{random}, a$ 
3:   For loop, iterate ( $k$ )
4:      $B_k = X_{k-1} + \frac{1}{a} A^T (Y - A \circ X_{k-1})$ 
5:     Compute SVD (singular value decomposition) of
        $B: B_k = USV^T$ 
6:     Soft threshold the singular values:
        $\Sigma = \text{soft}(S, \lambda/2)$   $\triangleright$  refer equation 13
7:      $X_k = U\Sigma V^T$ 
8:      $X_k \leftarrow X_k^+$ 
9:   End loop 1

```

We found that the algorithm is robust to values of  $\lambda$  as long as it is reasonably small ( $< 0.01$ ).

Here too, we have imposed the non-negativity constraint on  $X$  since expressions cannot be smaller than zero. The Nuclear Norm Minimization algorithm has been depicted in Algorithm 2.

## 5. CONCLUSION

As an inevitable consequence of a steep decline in single cell library depth, dropout rates in scRNA-seq data have skyrocketed. This works as a confounding factor (Hicks et al., 2015), thereby hindering cell clustering and further downstream analyses. A good imputation strategy would handle the Dropouts problem gracefully and thereby has the potential to facilitate the discovery of new rare cell subtypes within seemingly similar cells. This, in turn, can be helpful for characterizing cancer heterogeneity and understanding the dynamics of transcriptional changes during development. The proposed mcImpute algorithm, without making any assumption about the expression data distribution, recovers dropouts by simultaneously retaining the true zero counts and shows comparable performance on a number of

measures including clustering accuracy, cell type separability, differential gene prediction, cell visualization, gene distribution, etc.

We believe that McImpute, by far is the most intuitive way of catering the dropouts problem. It can seamlessly be integrated and serve as a key component in single-cell RNA seq pipeline.

Currently, imputation and clustering are together a piecemeal two-step process—imputation followed by clustering. In the future, we would like to incorporate both clustering and imputation as a joint optimization problem.

## 6. SOFTWARE

The source code of mcImpute is shared at [https://github.com/aanchalMongia/McImpute\\_scRNAseq](https://github.com/aanchalMongia/McImpute_scRNAseq).

## DATA AVAILABILITY STATEMENT

The details of datasets for this study has been given in section 4.

## AUTHOR CONTRIBUTIONS

DS and AnM led the study, contributed to the statistical analysis and design of the experiments. AaM analyzed and interpreted the scRNA-seq data and performed the experiments. All authors read and reviewed the manuscript.

## ACKNOWLEDGMENTS

This manuscript has been submitted to the preprint server-bioRxiv (Mongia et al., 2018).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.00009/full#supplementary-material>

## REFERENCES

- Abernethy, J., Bach, F., Evgeniou, T., and Vert, J. P. (2006). Low-rank matrix factorization with attributes. *arXiv preprint cs/0611124*.
- Amit, Y., Fink, M., Srebro, N., and Ullman, S. (2007). "Uncovering shared structures in multiclass classification," in *Proceedings of the 24th International Conference on Machine Learning* (Corvallis, OR: ACM), 17–24.
- Argyriou, A., Evgeniou, T., and Pontil, M. (2007). "Multi-task feature learning," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 41–48.
- Bell, R. M. and Koren, Y. (2007). "Improved neighborhood-based collaborative filtering," in *KDD Cup and Workshop at the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Jose, CA: Citeseer), 7–14.
- Bennett, J., and Lanning, S. (2007). "The netflix prize," in *Proceedings of KDD Cup and Workshop Vol 2007* (New York, NY), 35.
- Biase, F. H., Cao, X., and Zhong, S. (2014). Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell rna sequencing. *Gen. Res.* 24, 1787–1796. doi: 10.1101/gr.17725.114
- Candès, E. J., and Plan, Y. (2010). Matrix completion with noise. *Proc. IEEE* 98, 925–936. doi: 10.1109/JPROC.2009.2035722
- Candès, E. J., and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* 9, 717–772. doi: 10.1007/s10208-009-9045-5
- Candès, E. J., and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theor.* 56, 2053–2080. doi: 10.1109/TIT.2010.2044061
- Gill, R., Datta, S., and Datta, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinform.* 11:95. doi: 10.1186/1471-2105-11-95
- Grün, D., Kester, L., and Van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nat. Methods* 11:637. doi: 10.1038/nmeth.2930
- Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015). Matrix completion and low-rank svd via fast alternating least squares. *J. Mach. Learn. Res.* 16, 3367–3402.
- Hicks, S. C., Teng, M., and Irizarry, R. A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell rna-seq data. *bioRxiv [preprint]*. doi: 10.1101/025528
- Holland, S. M. (2008). *Principal Components Analysis (pca)*. Athens, GA: Department of Geology, University of Georgia, 30602–2501.

- Kapur, A., Marwah, K., and Alterovitz, G. (2016). Gene expression prediction using low-rank matrix completion. *BMC Bioinformatics* 17:243. doi: 10.1186/s12859-016-1106-6
- Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. *IEEE Trans. Inf. Theor.* 56, 2980–2998. doi: 10.1109/TIT.2010.2046205
- Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11, 740–742. doi: 10.1038/nmeth.2967
- Klebanov, L., and Yakovlev, A., (2007). Diverse correlation structures in gene expression data and their utility in improving statistical inference. *Ann. Appl. Stat.* 1, 538–559. doi: 10.1214/07-AOAS120
- Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201. doi: 10.1016/j.cell.2015.04.044
- Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer* 8, 30–37. doi: 10.1109/MC.2009.263
- Kwak, I. Y., Gong, W., Koyano-Nakagawa, N., and Garry, D. (2017). Drimpute: imputing dropout events in single cell rna sequencing data. *bioRxiv [preprint]*. doi: 10.1101/181479
- Lee, D. D., and Seung, H. S. (2001). “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems 13*, eds T. K. Leen, T. G. Dietterich, and V. Tresp (Vancouver, BC: MIT Press), 556–562.
- Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* 49, 708–718. doi: 10.1038/s41467-018-03405-7
- Li, W. V., and Li, J. J. (2018). An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nat. Commun.* 9:997. doi: 10.1038/ng.3818
- Liu, S., Maljovec, D., Wang, B., Bremer, P. T., and Pascucci, V. (2017). Visualizing high-dimensional data: Advances in the past decade. *IEEE Trans. Visual. Comp. Grap.* 23, 1249–1268. doi: 10.1109/TVCG.2016.2640960
- Maaten, L. v. d., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9, 2579–2605.
- Macosko, E. Z., Basu, A., Satija, R., Nemes, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002
- Majumdar, A., and Ward, R. (2011). Some empirical advances in matrix completion. *Signal Process.* 91, 1334–1338. doi: 10.1016/j.sigpro.2010.12.005
- Marjanovic, G., and Solo, V. (2012). On lq optimization and matrix completion. *IEEE Trans. Signal Process.* 60, 5714–5724. doi: 10.1109/TSP.2012.2212015
- Mesbahi, M., and Papavassilopoulos, G. P. (1997). On the rank minimization problem over a positive semidefinite linear matrix inequality. *IEEE Trans. Autom. Control* 42, 239–243. doi: 10.1109/9.554402
- Mongia, A., Sengupta, D., and Majumdar, A. (2018). Mcimpute: matrix completion based imputation for single cell rna-seq data. *bioRxiv [preprint]*. doi: 10.1101/361980
- Najafav, J., and Najafav, A. (2018). GECCO: gene expression correlation analysis after genetic algorithm-driven deconvolution. *Bioinformatics* 35, 156–159. doi: 10.1093/bioinformatics/bty623
- Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401. doi: 10.1126/science.1254257
- Reynier, F., Petit, F., Paye, M., Turrel-Davin, F., Imbert, P. E., Hot, A., et al. (2011). Importance of correlation between gene expression levels: application to the type i interferon signature in rheumatoid arthritis. *PLoS ONE* 6:e24828. doi: 10.1371/journal.pone.0024828
- Sengupta, D., Rayan, N. A., Lim, M., Lim, B., and Prabhakar, S. (2016). Fast, scalable and accurate differential expression analysis for single cells. *bioRxiv [preprint]*. doi: 10.1101/049734
- SIGKDD (2007). *Kdd Cup 2007*. Available at online: <https://www.kdd.org/kdd-cup/view/kdd-cup-2007>. (Accessed December 15, 2018).
- Silver, M., Chen, P., Li, R., Cheng, C.-Y., Wong, T.-Y., Tai, E.-S., et al. (2013). Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two asian cohorts. *PLoS Genet.* 9:e1003939. doi: 10.1371/journal.pgen.1003939
- Staiger, C., Cadot, S., Györfy, B., Wessels, L. F., and Klau, G. W. (2013). Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Front. Genet.* 4:289. doi: 10.3389/fgene.2013.00289
- Sun, Y., Babu, P., and Palomar, D. P. (2017). Majorization-minimization algorithms in signal processing, communications, and machine learning. *Trans. Sig. Proc.* 65, 794–816. doi: 10.1109/TSP.2016.2601299
- Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., et al. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell rna-seq analysis. *Cell Stem Cell* 6, 468–478. doi: 10.1016/j.stem.2010.03.015
- Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science* 352, 189–196.
- Tomasi, C., and Kanade, T. (1992). Shape and motion from image streams under orthography: a factorization method. *Int. J. Comp. Vis.* 9, 137–154. doi: 10.1126/science.aad0501
- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nat. Biotech.* 32:381. doi: 10.1038/nbt.2859
- Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnberg, P., Lou, D., et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nat. Neurosci.* 18:145. doi: 10.1038/nn.3881
- van Dijk, D., Nainys, J., Sharma, R., Kathail, P., Carr, A. J., Moon, K. R., et al. (2017). Magic: a diffusion-based imputation method reveals gene-gene interactions in single-cell rna-sequencing data. *BioRxiv [preprint]*. doi: 10.1101/111591
- Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* 34, 1145–1160. doi: 10.1038/nbt.3711
- Weckwerth, W., Loureiro, M. E., Wenzel, K., and Fiehn, O. (2004). Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 7809–7814. doi: 10.1073/pnas.0303415101
- Xiong, M., Feghali-Bostwick, C. A., Arnett, F. C., and Zhou, X. (2005). A systems biology approach to genetic studies of complex diseases. *FEBS Lett.* 579, 5325–5332. doi: 10.1016/j.febslet.2005.08.058
- Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1131–1139. doi: 10.1038/nsmb.2660
- Ye, G., Tang, M., Cai, J. F., Nie, Q., and Xie, X. (2013). Low-rank regularization for learning gene expression programs. *PLoS ONE* 8:e82146. doi: 10.1371/journal.pone.0082146
- Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnberg, P., La Manno, G., Jureus, A., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science* 347, 1138–1142. doi: 10.1126/science.aaa1934
- Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049. doi: 10.1038/ncomms14049
- Zhou, X., Lindsay, H., and Robinson, M. D. (2014). Robustly detecting differential expression in rna sequencing data using observation weights. *Nucleic Acids Res.* 42, e91–e91. doi: 10.1093/nar/gku310

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Mongia, Sengupta and Majumdar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Expression Profile Analysis Identifies a Novel Five-Gene Signature to Improve Prognosis Prediction of Glioblastoma

Wen Yin<sup>1</sup>, Guihua Tang<sup>2</sup>, Quanwei Zhou<sup>1</sup>, Yudong Cao<sup>1</sup>, Haixia Li<sup>3</sup>, Xianrong Fu<sup>1</sup>, Zhaoping Wu<sup>1</sup> and Xingjun Jiang<sup>1\*</sup>

<sup>1</sup> Department of Neurosurgery, Xiangya Hospital of Central South University, Changsha, China, <sup>2</sup> Department of Clinical Laboratory, Hunan Provincial People's Hospital (First Affiliated Hospital of Hunan Normal University), Changsha, China, <sup>3</sup> Department of Operative Nursing, Xiangya Hospital of Central South University, Changsha, China

## OPEN ACCESS

### Edited by:

Monica Bianchini,  
University of Siena, Italy

### Reviewed by:

Nitish Kumar Mishra,  
University of Nebraska Medical  
Center, United States  
Sen Peng,  
Translational Genomics Research  
Institute, United States  
Max Shpak,  
St David's Medical Center,  
United States

### \*Correspondence:

Xingjun Jiang  
jiangxj@csu.edu.cn;  
jxjyz@163.com

### Specialty section:

This article was submitted to  
Bioinformatics and Computational  
Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 20 December 2018

**Accepted:** 17 April 2019

**Published:** 03 May 2019

### Citation:

Yin W, Tang G, Zhou Q, Cao Y,  
Li H, Fu X, Wu Z and Jiang X (2019)  
Expression Profile Analysis Identifies  
a Novel Five-Gene Signature  
to Improve Prognosis Prediction  
of Glioblastoma.  
Front. Genet. 10:419.  
doi: 10.3389/fgene.2019.00419

Glioblastoma multiforme (GBM) is the most aggressive primary central nervous system malignant tumor. The median survival of GBM patients is 12–15 months, and the 5 years survival rate is less than 5%. More novel molecular biomarkers are still urgently required to elucidate the mechanisms or improve the prognosis of GBM. This study aimed to explore novel biomarkers for GBM prognosis prediction. The gene expression profiles from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) datasets of GBM were downloaded. A total of 2241 overlapping differentially expressed genes (DEGs) were identified from TCGA and GSE7696 datasets. By univariate COX regression survival analysis, 292 survival-related genes were found among these DEGs ( $p < 0.05$ ). Functional enrichment analysis was performed based on these survival-related genes. A five-gene signature (PTPRN, RGS14, G6PC3, IGFBP2, and TIMP4) was further selected by multivariable Cox regression analysis and a prognostic model of this five-gene signature was constructed. Based on this risk score system, patients in the high-risk group had significantly poorer survival results than those in the low-risk group. Moreover, with the assistance of GEPIA <http://gepia.cancer-pku.cn/>, all five genes were found to be differentially expressed in GBM tissues compared with normal brain tissues. Furthermore, the co-expression network of the five genes was constructed based on weighted gene co-expression network analysis (WGCNA). Finally, this five-gene signature was further validated in other datasets. In conclusion, our study identified five novel biomarkers that have potential in the prognosis prediction of GBM.

**Keywords:** glioblastoma, differentially expressed genes, gene signature, prognosis, TCGA, GEO

## INTRODUCTION

Glioblastoma multiforme (GBM) is the most common and aggressive primary central nervous system malignant tumor with high morbidity and mortality. According to genomic abnormalities and gene expression, GBM can be divided into four molecular subtypes: classical, mesenchymal, neural, and proneural, which lay a foundation for understanding its inherent heterogeneity (Verhaak et al., 2010; Ma et al., 2018). In the United States, the incidence of GBM is 2.96 cases/100,000 population/year (Jhanwar-Uniyal et al., 2015). Although there are several treatment options, including surgery, radiotherapy and chemotherapy, the median survival of GBM

patients remains 12–15 months, and the 5 years survival rate is less than 5% (Wen and Kesari, 2008; Ostrom et al., 2013).

With the development of next-generation sequencing technologies, many specific molecular signatures have been identified to better understand the molecular pathogenesis of GBM (Aldape et al., 2015). As a result, many potential diagnostic and prognostic biomarkers have been discovered that enable a more specific classification and a more precise outcome prediction of GBM. Some molecular markers including MGMT (O6-methylguanine DNA methyltransferase), IDH (isocitrate dehydrogenase), EGFR (epidermal growth factor receptor), and PTEN (phosphatase and tensin homolog) have been routinely tested in GBM patients clinically (van den Bent et al., 2017; Binabaj et al., 2018). More importantly, these molecular signatures have contributed to personalized therapeutic approaches and targeted anti-GBM therapies (Huang et al., 2017; Szopa et al., 2017). However, considering the poor prognosis of GBM, novel molecular biomarkers and new therapeutic strategies are still urgently required to elucidate the mechanisms of GBM or increase overall patient survival.

Previous studies have shown that gene expression profile analysis could detect gene signatures to predict the outcome for malignancy tumors (Luo et al., 2018; Mao et al., 2018; Zeng et al., 2018). Shergalis et al. (2018) discovered that 20 genes were overexpressed and correlated with poor survival outcomes in GBM patients by bioinformatics analysis using data from The Cancer Genome Atlas (TCGA) project. Bao et al. (2014) identified a nine-gene signature to predict the prognosis of glioma patients based on mRNA expression profiling from the Chinese Glioma Genome Atlas (CGGA) database. Therefore, it is necessary to understand the development and progression of GBM by identifying GBM-related genes and to investigate of their potential clinical roles and molecular mechanisms.

In this study, RNA-Seq data from TCGA and microarray data from the Gene Expression Omnibus (GEO) database of GBM were downloaded. Based on the overlapping differentially expressed genes (DEGs), the genes related to prognosis were screened. By using Cox regression, we developed a five-gene signature based risk score to demonstrate the association between gene expression and the prognosis of GBM. Moreover, we validated this signature in the GEO dataset and TCGA array dataset. These results might be able to provide new reference for the prognostic predication of GBM.

## MATERIALS AND METHODS

### Data Source

The GBM RNA sequencing (RNA-seq) dataset and corresponding clinical follow-up information were downloaded from TCGA database (March, 2018). Subtype data of GBM were downloaded from UCSC Xena<sup>1</sup>. A total of 159 patients, including 154 samples of primary GBM patients and five samples of normal brain tissue were extracted for subsequent analysis.

<sup>1</sup><http://xena.ucsc.edu/>

Gene expression microarray data GSE7696 (Lambiv et al., 2011), including 71 samples of primary GBM patients and four samples of normal brain tissue, were downloaded from the National Center of Biotechnology Information (NCBI) Gene Expression Omnibus<sup>2</sup>. The dataset was based on the GPL570 platform of [HG-U133\_Plus\_2] Affymetrix Human Genome U133 Plus 2.0 Array (Affymetrix, Santa Clara, CA, United States).

### Differential Expression Analyses

Then, gene profiles were standard normalized within and among samples, respectively. Because the numerical distribution of RPKM (reads per kilo-base per million mapped reads) is too wide, the final expression level of a gene was defined as the  $\log_2(x + 1)$  of the raw expression level. Next, the DEGs between the tumor and normal samples were calculated by the limma package ( $\text{Padj} < 0.05$  and  $|\log_2\text{FC}| > 1$ ). The Venn diagram was produced by the VennDiagram R package (Chen and Boutros, 2011).

### Identification and Selection of Survival-Related Genes

Only the patients with detailed follow-up times were extracted for subsequent survival analyses. Univariate Cox regression survival analysis using the Survival package in R was performed to identify survival-related genes (Yang et al., 2016). Genes were selected with a  $p$ -value of less than 0.05.

### GO and KEGG Annotation of Survival-Related Genes

Gene Ontology (GO) enrichment and KEGG (Kyoto Encyclopedia of Genes and Genomes) analysis were performed on the survival-related genes (Ogata et al., 1999; Wanggou et al., 2016; Li et al., 2018). DAVID (The Database for Annotation, Visualization, and Integrated Discovery) (Dennis et al., 2003) software and the clusterProfiler package (Yu et al., 2012) in R were used to annotate and visualize GO terms and KEGG pathways.

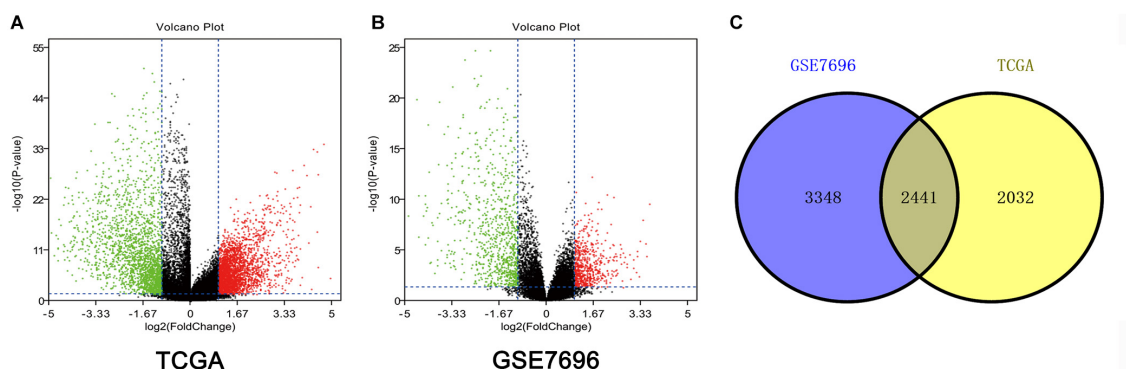
### Gene Signature Identification and Risk Score System Establishment

Based on the top 100 survival-related genes in TCGA dataset, multivariable Cox proportional hazard regression analysis was performed to establish a risk score formula (O'Quigley and Moreau, 1986). As previously reported, a prognosis risk score formula could be constructed on the basis of a linear combination of the expression level (exp) multiplied by a regression coefficient ( $\beta$ ) derived from the multivariate cox regression model.

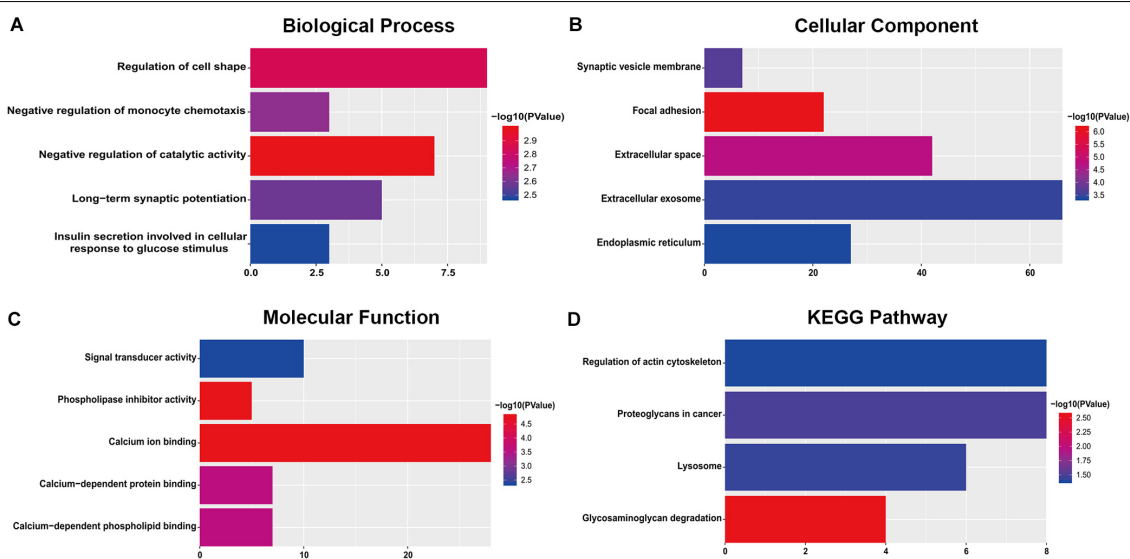
$$\text{Risk Score (RS)} = \text{expPTPRN} * \beta_{\text{PTPRN}} + \text{expRGS14} * \beta_{\text{RGS14}} \\ + \text{expG6PC3} * \beta_{\text{G6PC3}} + \text{expIGFBP2} * \beta_{\text{IGFBP2}} + \text{expTIMP4} * \beta_{\text{TIMP4}}$$

Based on the formula, the risk score of each GBM patient was calculated, and then GBM patients were divided into high-risk score and low-risk score groups. The receiver operating characteristic (ROC) curve analysis was conducted using the R

<sup>2</sup><http://www.ncbi.nlm.nih.gov/geo/>



**FIGURE 1 |** Identification of DEGs among TCGA and GEO datasets of GBM. **(A)** Volcano plots of DEGs in TCGA dataset. **(B)** Volcano plots of DEGs in GSE7696 dataset. **(C)** The Venn diagram of overlapping DEGs among TCGA and GSE7696 datasets.



**FIGURE 2 |** The most significantly enriched GO annotations and KEGG pathways of genes related to survival. The length of the bars represents the number of genes, and the color of the bars corresponds to the  $p$ -value according to legend. **(A)** Top 5 significantly enriched biological process. **(B)** Top 5 significantly enriched cellular component. **(C)** Top 5 significantly enriched molecular function. **(D)** Top 5 significantly enriched KEGG pathways.

package “pROC.” After choosing an optimal cut-off point with the maximal sensitivity and specificity, the survival differences between the low-risk and high-risk groups were assessed by the Kaplan–Meier analysis with log-rank test. Similarly, to evaluate the predictive power of the five-gene signature in internal dataset, we assessed the gene signature within each subtype (classical, mesenchymal, neural, and proneural).

## Analysis in GEPIA and Exploring Co-expression by WGCNA

The expression levels of the five genes were acquired with the assistance of GEPIA<sup>3</sup>, which is a newly developed interactive web server for analyzing the RNA sequencing expression data of 23 types of cancers and normal samples from TCGA

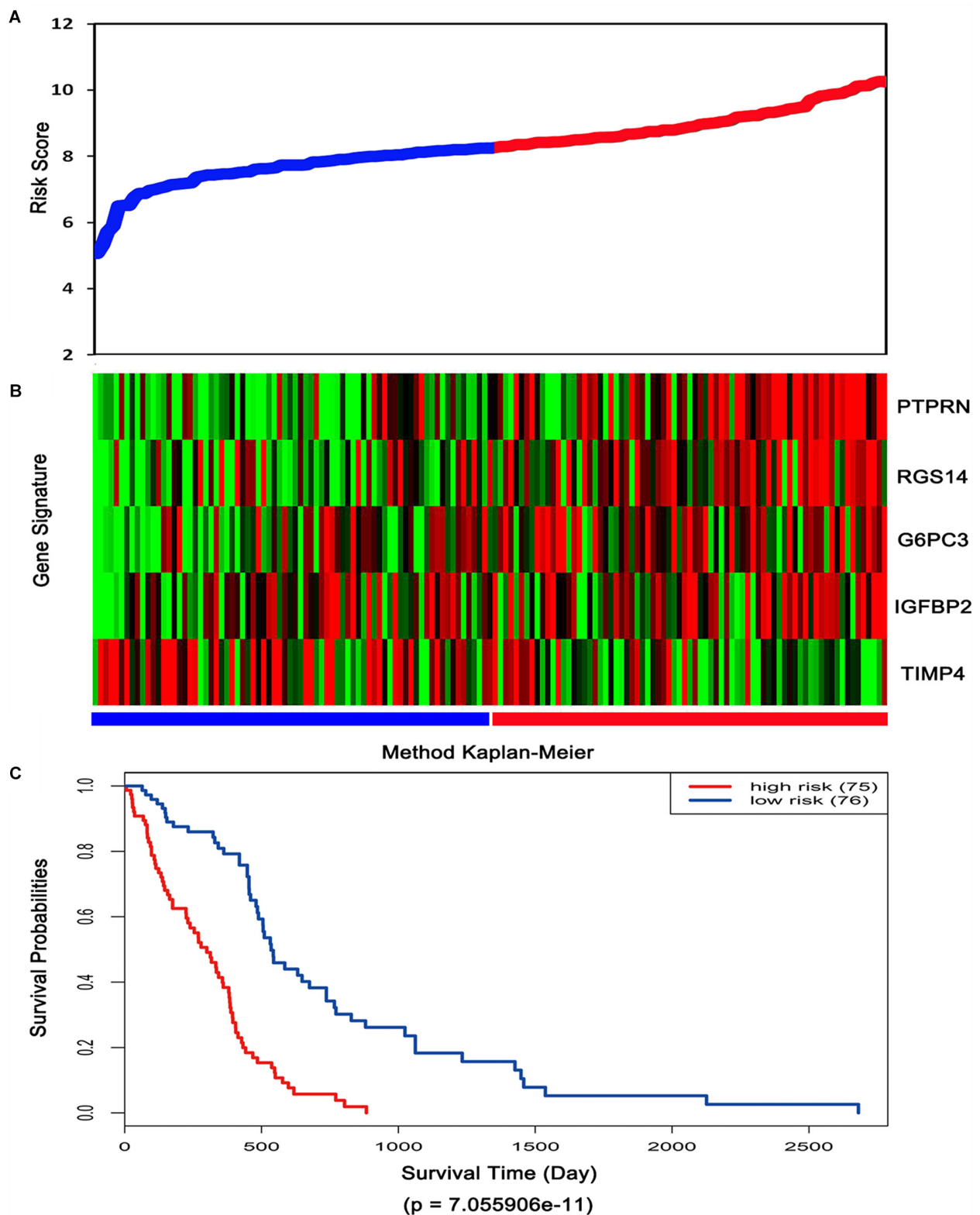
<sup>3</sup><http://gepia.cancer-pku.cn/>

**TABLE 1 |** Information about the five genes screened to build the risk score system.

Genes	Coefficient	HR	95% CI	P-value
PTPRN	0.50894	1.66353	1.4010–1.9753	6.35e-09
RGS14	0.54671	1.72757	1.2026–2.4816	0.00309
G6PC3	1.20753	3.34520	1.9960–5.6063	4.57e-06
IGFBP2	0.25845	1.29492	1.1096–1.5112	0.00104
TIMP4	−0.20684	0.81315	0.6951–0.9513	0.00976

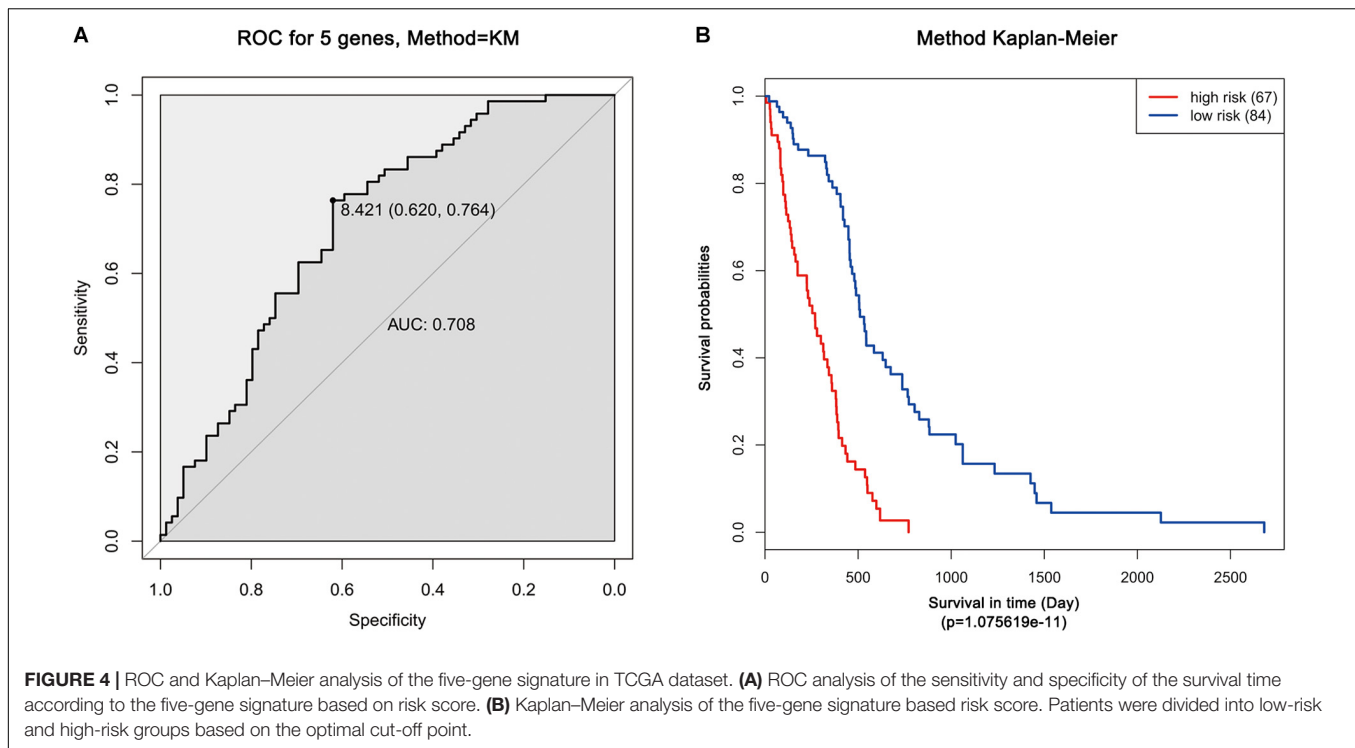
and the GTEx projects according to the standard processing pipeline (Tang et al., 2017).

To explore the regulatory network of the five genes, all the overlapped DEGs were analyzed by WGCNA (Ahn et al., 2016; Chen et al., 2018). Finally, the co-expression network of the



**FIGURE 3 |** Risk score analysis, expression distribution and survival analysis of the five-gene signature in TCGA dataset. **(A)** The five-gene signature risk score distribution. **(B)** The heat-map of the five-gene expression profiles. Red indicates a higher expression and green indicates a lower expression. Blue bar: low-risk group. Red bar: high-risk group. **(C)** Kaplan-Meier analysis using the median risk score cut-off which divided patients into low-risk and high-risk groups.





five genes was constructed based on WGCNA and visualized by Cytoscape 3.6.1 (Shannon et al., 2003).

## Validation of the Five-Gene Prognostic Signature by the GEO Dataset and TCGA Microarray Dataset

Dataset GSE13041 from the GEO and TCGA microarray dataset were used to validate this five-gene prognostic signature (Lee et al., 2008). The GSE13041 dataset including 188 samples of GBM patients and the TCGA microarray dataset including 498 samples of GBM patients were both based on the Affymetrix Human Genome U133A Array platform (GPL97). The ROC curves and Kaplan-Meier analyses were used to validate the prognostic value of the five-gene for GBM patients.

## RESULTS

### Differentially Expressed Genes (DEGs) in TCGA and GSE7696

Altogether, 4473 DEGs in TCGA dataset (Figure 1A) and 5789 DEGs in the GSE7696 dataset (Figure 1B) were screened by the limma package. The 2241 overlapping DEGs were screened for further analysis (Figure 1C).

### Survival-Related Genes in GBM

In TCGA dataset, every overlapped DEG was evaluated by univariate Cox regression survival analysis. Altogether, 292 significantly changed genes were considered -survival-related

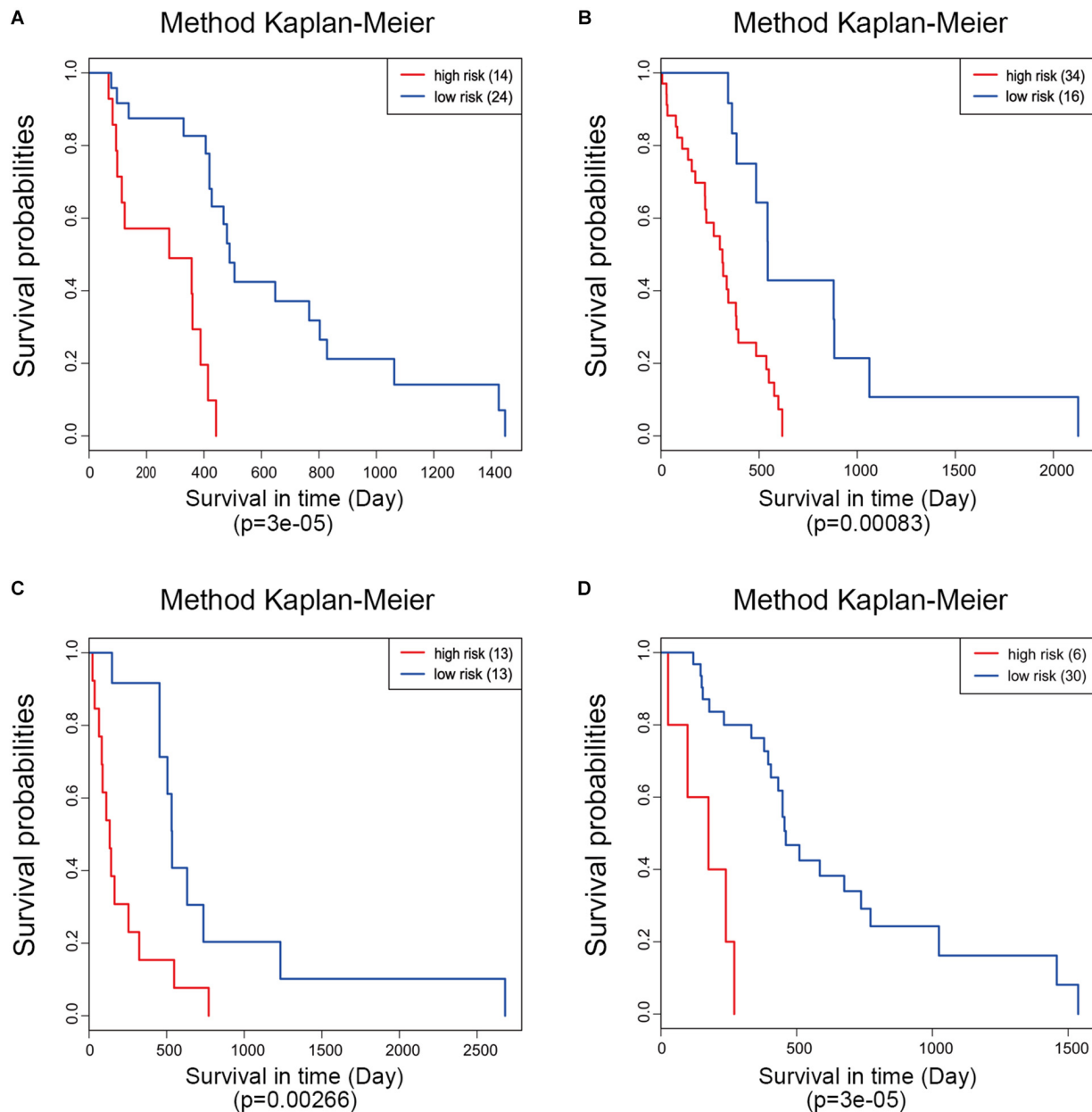
genes by the threshold of  $p < 0.05$ . The top 100 survival-related genes are shown in **Supplementary Table 1**.

## Go and KEGG Analysis of Survival-Related Genes

For the “biological processes” (BP), negative regulation of catalytic activity, regulation of cell shape, negative regulation of monocyte chemotaxis, long-term synaptic potentiation and insulin secretion involved in cellular response to glucose stimulus were the commonly enriched categories (Figure 2A). For the “cellular component” (CC), the enriched categories were correlated with focal adhesion, extracellular space, synaptic vesicle membrane, extracellular exosome, and endoplasmic reticulum (Figure 2B). For the “molecular function” (MF), those genes mainly showed enrichment in calcium ion binding, phospholipase inhibitor activity, calcium-dependent protein binding, calcium-dependent phospholipid binding, and signal transducer activity (Figure 2C). KEGG pathway enrichment analysis suggested that glycosaminoglycan degradation was the most significant pathway. These genes also participated in following pathways: proteoglycans in cancer, lysosome, and regulation of the actin cytoskeleton (Figure 2D).

## Risk Score System Based on Five-Gene Signature

After multivariate Cox regression analysis was conducted for these 100 genes, five genes (PTPRN, RGS14, G6PC3, IGFBP2, and TIMP4) were selected as signature genes that can optimally predict the overall survival of patients with GBM (Table 1). To comprehensively investigate the association between these five



**FIGURE 5 |** Kaplan–Meier analysis of the five-gene signature in different molecular subtypes of glioblastoma. Classical (A), mesenchymal (B), neural (C), and proneural (D).

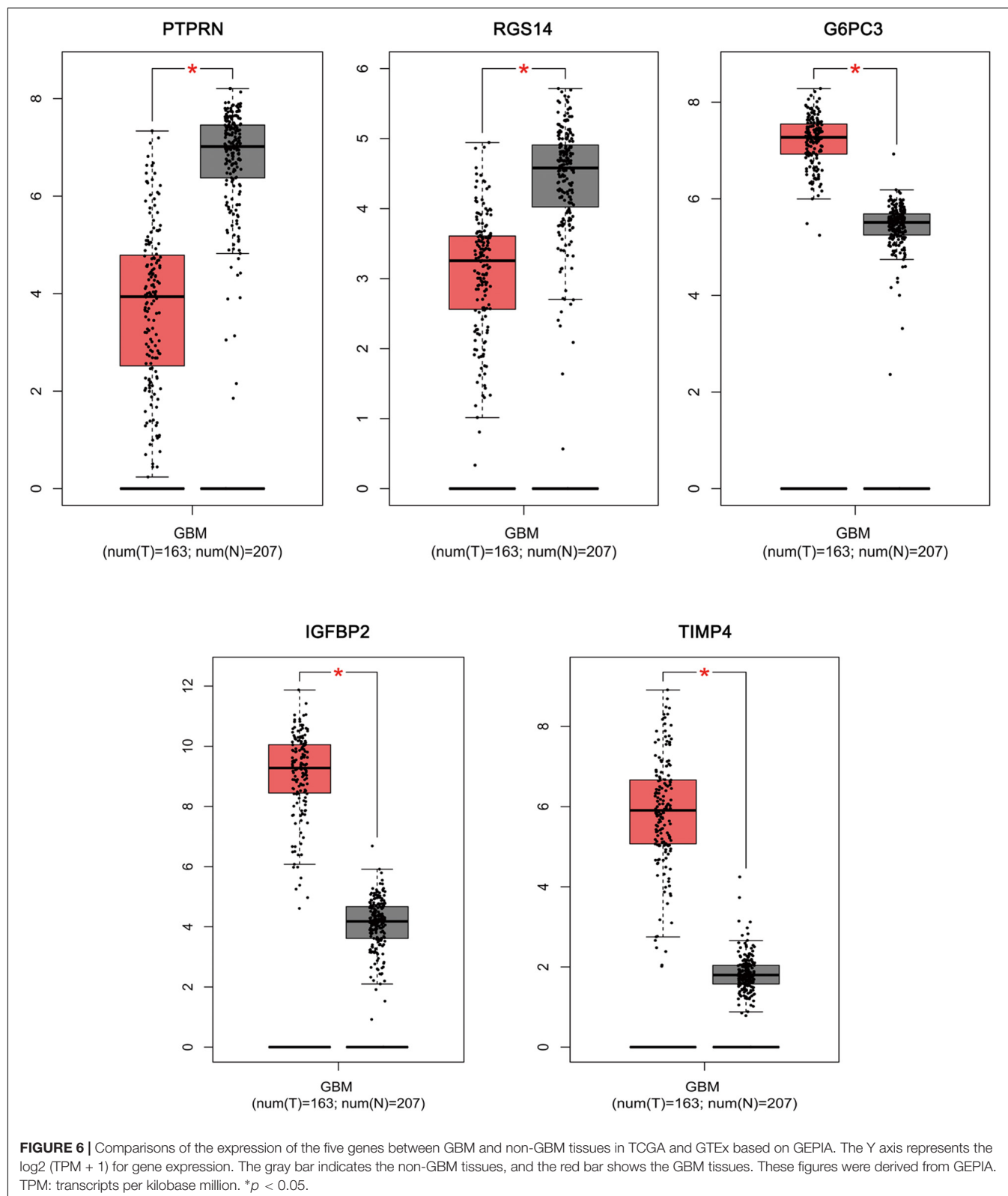
genes and the prognosis of GBM, a five-gene survival risk score system was established based on their Cox coefficients.

$$\text{Risk Score (RS)} = 0.50894 \cdot \text{exp}_{\text{PTPRN}} + 0.54671 \cdot \text{exp}_{\text{PRGS14}} + 1.20753 \cdot \text{exp}_{\text{G6PC3}} + 0.25845 \cdot \text{exp}_{\text{IGFBP2}} - 0.20684 \cdot \text{exp}_{\text{TIMP4}}$$

Then, the risk score for each patient was calculated in TCGA dataset and ranked according to the risk scores. Thus, patients were divided into a high-risk group ( $n = 75$ ) and a low-risk group ( $n = 76$ ). The survival time of GBM patients was adversely

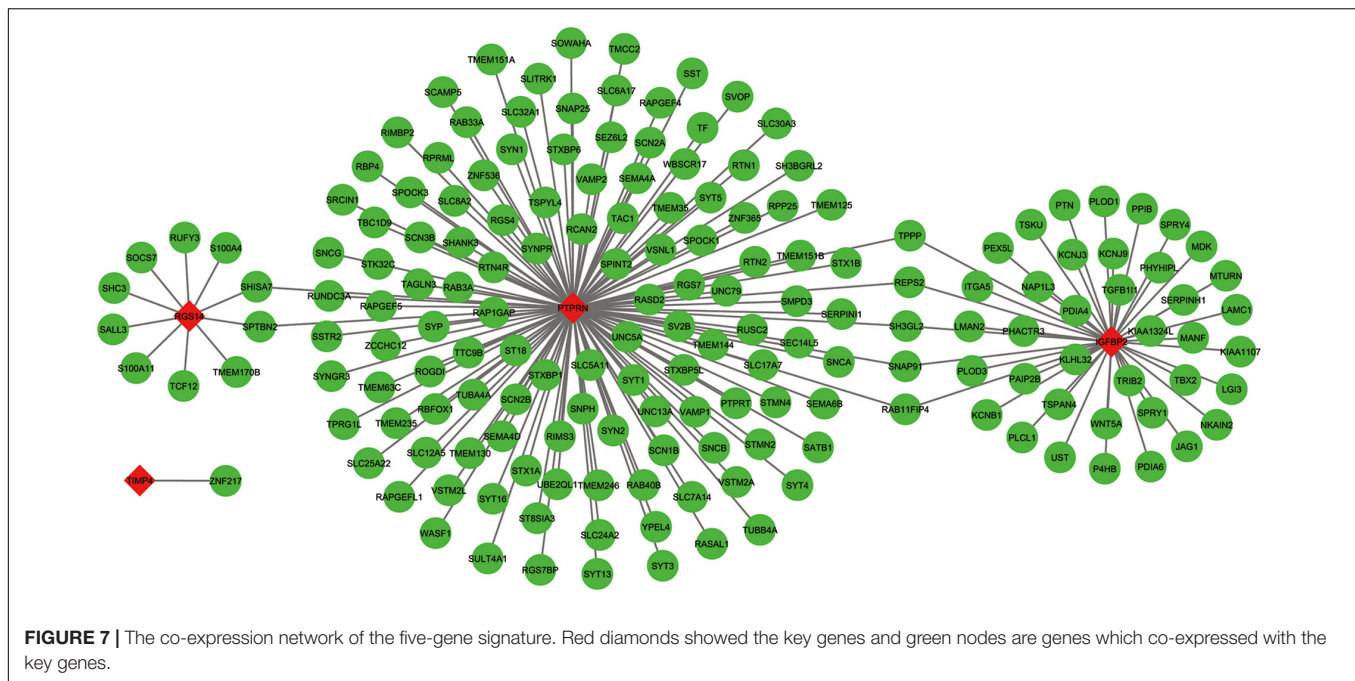
associated with their risk scores (Figure 3A). A remarkably lower expression was noted for TIMP4 in the high-risk groups, while a higher expression was observed for the other genes in the high-risk groups (Figure 3B). The Kaplan–Meier analysis and log-rank test showed that patients in the low-risk group had a significantly positive overall survival time compared to the high-risk group ( $p = 7.055906 \times 10^{-11}$ ) (Figure 3C).

Moreover, ROC analysis was performed for this risk score system. Figure 4A shows that the area under the ROC Curves (AUC) was 0.704. The optimal cutoff point was selected as 8.421. With this cutoff point, the patients were further divided into a



high-risk group and a low-risk group. The Kaplan–Meier analysis and log-rank test further indicated a significant difference in overall survival between the two groups ( $p = 1.075619\text{e-}11$ )

(Figure 4B). Similarly, with different cutoff points, the patients in each subtype were divided into a high-risk group and a low-risk group. The Kaplan–Meier analysis and log-rank test also



indicated a significant difference between the two groups in each subtype (**Figures 5A–D**).

### Analysis in GEPIA and Exploring Co-expression by WGCNA

Based on the results derived from GEPIA, the expression of G6PC3, IGFBP2, and TIMP4 were significantly up-regulated in GBM, while the expression of PTPRN and RGS14 were significantly down-regulated (**Figure 6**). By using GEPIA, the selected five genes were verified as DEGs in GBM with amplified normal sample sizes.

The co-expressed genes of the five genes were determined by WGCNA. Finally, 129 genes were discovered to be co-expressed with PTPRN, 41 genes were co-expressed with IGFBP2, 10 genes with RGS14 and 1 gene with TIMP4. However, no gene was co-expressed with G6PC3. The co-expression network of the four genes is visualized by WGCNA in **Figure 7**.

### Validation of the Five-Gene Prognostic Signature by GEO Dataset and TCGA Microarray Dataset

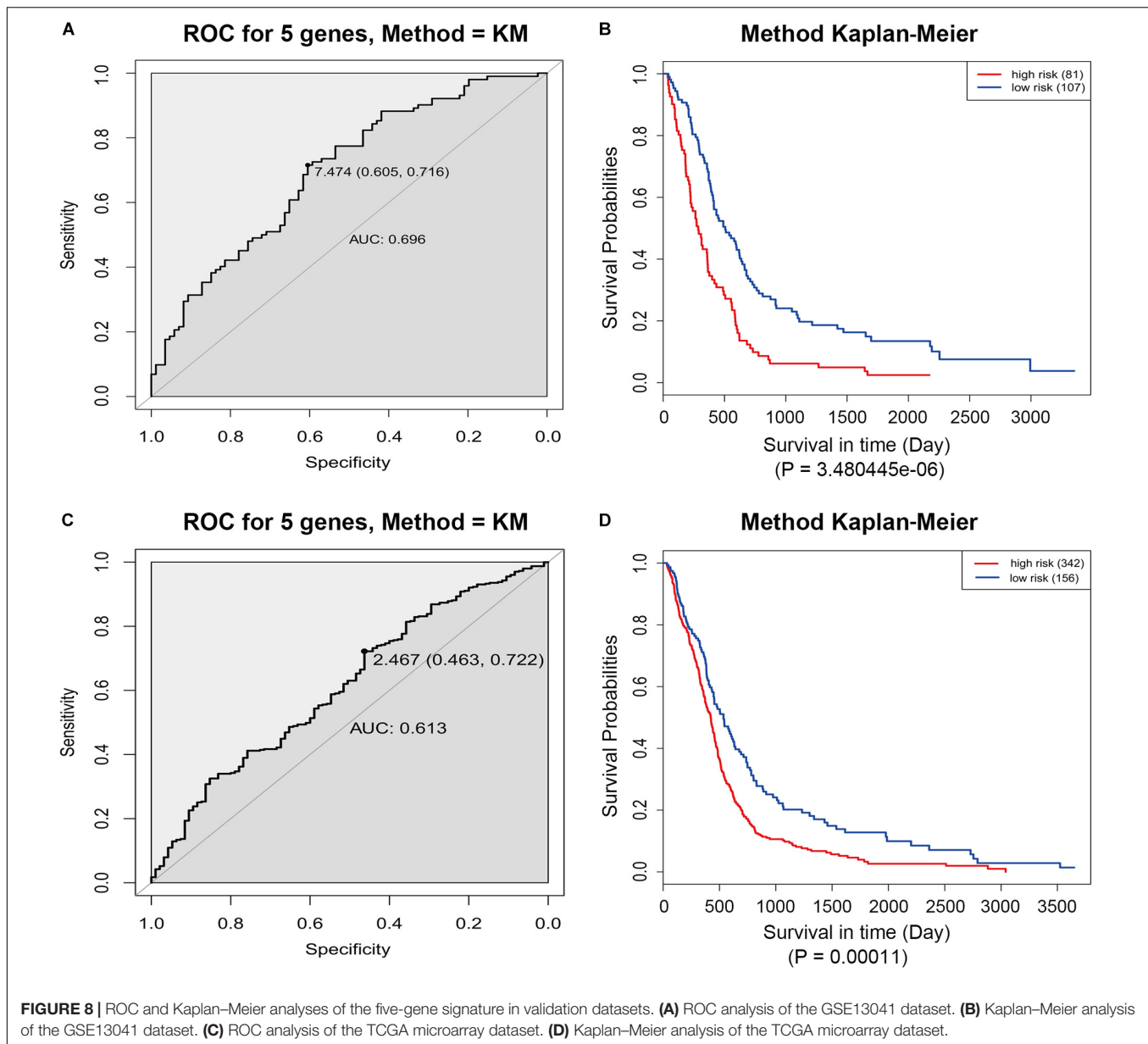
The GSE13041 dataset including 188 GBM patients and the TCGA microarray dataset including 498 GBM patients were used for the validation of the five-gene signature separately. Similarly, the risk score for each patient was calculated. ROC analyses were used to identify the optimal cutoff points (**Figures 8A,C**). Then, we divided the patients into a high-risk group and a low-risk group using the selected optimal cut-off points, respectively. The Kaplan–Meier analyses suggested a significantly prolonged survival time in the low-risk patients compared to that in the high-risk patients ( $p = 3.480445e-06$  and  $p = 0.00011$ ) (**Figures 8B,D**).

## DISCUSSION

GBM is the most aggressive brain tumor associated with poor prognosis. By analyzing TCGA and GSE7696 datasets, we identified 2241 significantly overlapping DEGs. A total of 292 survival-related DEGs were selected from the overlapping DEGs. Functional analyses demonstrated that these genes are mainly associated with following pathways: glycosaminoglycan degradation, proteoglycans in cancer, lysosome, and regulation of the actin cytoskeleton. More importantly, based on multivariate Cox regression analysis of TCGA dataset, five genes which could predict overall survival were screen out, namely PTPRN, RGS14, G6PC3, IGFBP2, and TIMP4. According to their Cox coefficients derived from cox regression, a risk score system based on the five genes was established. Additionally, after identifying the optimal cut-off point by ROC analysis, patients were classified into high-risk and low-risk groups. This five-gene signature was further successfully validated as a prognostic marker in each subtype of GBM, another independent GEO dataset (GSE13041) and TCGA microarray dataset. Furthermore, differential expression analysis of the five genes in GEPIA validated that three genes (G6PC3, IGFBP2, and TIMP4) were significantly up-regulated and two genes (PTPRN and RGS14) were significantly down-regulated in GBM. Co-expression network analysis revealed the regulation network of the five genes. These results suggest that these genes may play an important role in the molecular pathogenesis, progression and prognosis of GBM.

Based on GO and KEGG enrichment analyses of the survival-related DEGs among different studies, “negative regulation of catalytic activity” was the most significant enrichment in BP. This indicated that inhibiting the catalytic activity of some genes may be critical for cancer progression. Coincidentally, Zhao et al. (2009) found that IDH1 mutation could inhibit





IDH1 catalytic activity and contribute to the tumorigenesis of glioma. Other BPs such as regulation of cell shape and negative regulation of monocyte chemotaxis were also enriched. For the CC category, focal adhesion was the most significant enrichment which has been shown to be as a major determinant of cell migration and an essential process in tumor invasion (Garzon-Muvdi et al., 2012). The following three kinds of CCs, extracellular space, synaptic vesicle membrane and extracellular exosome, may also play important roles in tumor development and its micro-environmental manipulation (Wei et al., 2017). Regarding the MF category, calcium ion binding was the most affected MF.  $\text{Ca}^{2+}$ -mediated cell connectivity and plasticity are unique features of the central nervous system, and the  $\text{Ca}^{2+}$ /calmodulin-dependent process is able to regulate cell cycle progression and inhibit proliferation of malignant glioma

(Cheng et al., 1995; Liu et al., 2011). For KEGG pathway enrichment analysis, glycosaminoglycan degradation was the most significant pathway. Extracellular proteoglycans play critical roles in driving oncogenic pathways in tumor cells and promoting critical tumor-microenvironment interactions (Wade et al., 2013). The other KEGG pathways, proteoglycans in cancer, lysosome, and regulation of actin cytoskeleton, were also closely related to oncogenesis (Liu et al., 2012; Terakawa et al., 2013; Wade et al., 2013).

The five-gene signature provides a wealth of potential biological and therapeutic information about GBM. PTPRN (protein tyrosine phosphatase, receptor type N), located on the long arm of human chromosome 2 (2q35) (Lan et al., 1996), is an integral transmembrane protein of dense core vesicles and plays an important role in the secretion of hormones and

neurotransmitters (Xu et al., 2016). PTPRN has been confirmed to be negatively related to the survival of hepatocellular carcinoma patients and closely related to liver tumorigenesis (Zhangyuan et al., 2018). Moreover, the hypermethylation of PTPRN is also associated with shorter survival in ovarian cancer patients (Bauerschlag et al., 2011). A high expression of PTPRN in small cell lung cancer is associated with tumor growth and proliferation. Interestingly, Shergalis et al. also found that a high PTPRN expression is strongly associated with a poor prognosis in GBM patients, which was consistent with our finding (Shergalis et al., 2018). RGS14 is a member of the regulator of the G-protein signaling (RGS) protein family and is highly expressed in the caudate nucleus of the brain, spleen and thymus (Cho et al., 2005; Gerber et al., 2016). Previous study found that RGS14 is important for centrosome function, transcriptional regulation and stress-induced cellular responses (Cho et al., 2005). However, little work has been done to elucidate the role of RGS14 in cancer. Interestingly, PTPRN and RGS14 expressed at low levels in GBM tissue, but their increased expression was associated with poor prognosis. The reason may be that they have different functions in normal and tumor tissues. More work is needed to elucidate their functions in GBM. G6PC3, namely, glucose-6-phosphatase isoform  $\beta$ , is a catalysis subunit of G6PC (Gao et al., 2017). G6PC (glucose-6-phosphatase) is a key enzyme that regulates glucose homeostasis and glycogenolysis, which has been reported as a specific enzyme regulating proliferation and invasiveness in several tumors, such as liver, kidney and ovarian cancer (Gao et al., 2017). Furthermore, a previous study revealed that G6PC is a key enzyme regulating glioblastoma invasion (Abbadi et al., 2014). Our study demonstrated that G6PC3 was significantly up-regulated in GBM samples compared with normal brain tissue, and the high expression of G6PC3 was closely related to a poor prognosis in GBM patients. IGFBP2 (Insulin-like growth factor binding protein 2), an important member of the Insulin-like growth factor binding protein family, modulates cell growth, differentiation, migration, and invasion in neoplasms (Fukushima and Kataoka, 2007). IGFBP2 is involved in immunosuppressive activities and is a potential immunotherapeutic target for GBM (Cai et al., 2018). Our study confirmed that IGFBP2 was significantly up-regulated in GBM and predicted a worse outcome for patients, which was consistent with the previous study (Cai et al., 2018). TIMP4 is a member of tissue inhibitors of matrix metalloproteinases (TIMPs), which are involved in several processes of tumorigenesis including proliferation, migration, and invasion (Boufraqueh et al., 2016). A high-expression of TIMP4 has been found in patients with breast, cervical, and prostate cancers, whereas a low expression has been observed in patients with pancreatic cancer (Boufraqueh et al., 2016).

## REFERENCES

Abbadi, S., Rodarte, J. J., Abutaleb, A., Lavell, E., Smith, C. L., Ruff, W., et al. (2014). Glucose-6-phosphatase is a key metabolic regulator of glioblastoma invasion. *Mol. Cancer Res.* 12, 1547–1559. doi: 10.1158/1541-7786.MCR-14-0106-T

Interestingly, our study found that TIMP4 was high-expressed in GBM patients, however, its high expression was associated with a good prognosis in patients with GBM. More work is also needed to elucidate its functions in GBM. In summary, the five-gene signature not only is robust for predicting the overall survival for GBM, but also has promising practical value in the treatment of GBM.

There are some limitations in our work. First of all, there were only very limited normal samples included in our differential expression analyses, which might neglect some potential mRNAs. Moreover, the efficiency of the five-gene signature should be confirmed in more GBM patients. Furthermore, the molecular mechanisms how the five-gene signature affected the prognosis of GBM patients should be further elucidated by a series of experiments.

## CONCLUSION

In conclusion, our study identified five novel biomarkers that have potential for the prognosis prediction in GBM. Moreover, our findings provide new insights into the pathogenesis and prognosis of GBM.

## AUTHOR CONTRIBUTIONS

WY and XJ conceived and designed the study. GT, QZ, YC, HL, XF, and ZW performed the analysis procedures. GT, WY, and XJ analyzed the results. WY and XJ wrote the manuscript. All authors contributed to the editing of the manuscript.

## FUNDING

This work was supported by the National Natural Science Foundation of China (No. 81472355).

## ACKNOWLEDGMENTS

We sincerely acknowledge the public databases: TCGA, GEO, and GEPIA.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.00419/full#supplementary-material>

Ahn, R., Gupta, R., Lai, K., Chopra, N., Arron, S. T., and Liao, W. (2016). Network analysis of psoriasis reveals biological pathways and roles for coding and long non-coding RNAs. *BMC Genomics* 17:841. doi: 10.1186/s12864-016-3188-y

Alfade, K., Zadeh, G., Mansouri, S., Reifenberger, G., and von Deimling, A. (2015). Glioblastoma: pathology, molecular mechanisms and markers. *Acta Neuropathol.* 129, 829–848. doi: 10.1007/s00401-015-1432-1

- Bao, Z. S., Li, M. Y., Wang, J. Y., Zhang, C. B., Wang, H. J., Yan, W., et al. (2014). Prognostic value of a nine-gene signature in glioma patients based on mRNA expression profiling. *CNS Neurosci. Therapeut.* 20, 112–118. doi: 10.1111/cns.12171
- Bauerschlag, D. O., Ammerpohl, O., Brautigam, K., Schem, C., Lin, Q., Weigel, M. T., et al. (2011). Progression-free survival in ovarian cancer is reflected in epigenetic DNA methylation profiles. *Oncology* 80, 12–20. doi: 10.1159/000327746
- Binabaj, M. M., Bahrami, A., ShahidSales, S., Joodi, M., Hassanian, S. M., Anvari, K., et al. (2018). The prognostic value of MGMT promoter methylation in glioblastoma: a meta-analysis of clinical trials. *J. Cell. Physiol.* 233, 378–386. doi: 10.1002/jcp.25896
- Boufraqech, M., Zhang, L., Nilubol, N., Sadowski, S. M., Kotian, S., Quezado, M., et al. (2016). Lysyl oxidase (LOX) transcriptionally regulates SNAI2 expression and TIMP4 secretion in human cancers. *Clin. Cancer Res.* 22, 4491–4504. doi: 10.1158/1078-0432.CCR-15-2461
- Cai, J., Chen, Q., Cui, Y., Dong, J., Chen, M., Wu, P., et al. (2018). Immune heterogeneity and clinicopathologic characterization of IGF2BP2 in 2447 glioma samples. *Oncoimmunology* 7:e1426516. doi: 10.1080/2162402X.2018.1426516
- Chen, H., and Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* 12:35. doi: 10.1186/1471-2105-12-35
- Chen, J., Wang, X., Hu, B., He, Y., Qian, X., and Wang, W. (2018). Candidate genes in gastric cancer identified by constructing a weighted gene co-expression network. *PeerJ* 6:e4692. doi: 10.7717/peerj.4692
- Cheng, E. H., Gorelick, F. S., Czernik, A. J., Bagaglio, D. M., and Hait, W. N. (1995). Calmodulin-dependent protein kinases in rat glioblastoma. *Cell Growth Differ.* 6, 615–621.
- Cho, H., Kim, D. U., and Kehrl, J. H. (2005). RGS14 is a centrosomal and nuclear cytoplasmic shuttling protein that traffics to promyelocytic leukemia nuclear bodies following heat shock. *J. Biol. Chemistry* 280, 805–814. doi: 10.1074/jbc.m408163200
- Dennis, G. Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et al. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* 4:3.
- Fukushima, T., and Kataoka, H. (2007). Roles of insulin-like growth factor binding protein-2 (IGFBP-2) in glioblastoma. *Anticancer Res.* 27, 3685–3692.
- Gao, Y., Li, L., Xing, X., Lin, M., Zeng, Y., Liu, X., et al. (2017). Coronin 3 negatively regulates G6PC3 in HepG2 cells, as identified by label-free mass spectrometry. *Mol. Med. Rep.* 16, 3407–3414. doi: 10.3892/mmr.2017.7002
- Garzon-Muvdi, T., Schiapparelli, P., Smith, C., Kim, D. H., Kone, L., Farber, H., et al. (2012). Regulation of brain tumor dispersal by NKCC1 through a novel role in focal adhesion regulation. *PLoS Biol.* 10:e1001320. doi: 10.1371/journal.pbio.1001320
- Gerber, K. J., Squires, K. E., and Hepler, J. R. (2016). Roles for regulator of G protein signaling proteins in synaptic signaling and plasticity. *Mol. Pharmacol.* 89, 273–286. doi: 10.1124/mol.115.102210
- Huang, J., Liu, F., Liu, Z., Tang, H., Wu, H., Gong, Q., et al. (2017). Immune checkpoint in glioblastoma: promising and challenging. *Front. Pharmacol.* 8:242. doi: 10.3389/fphar.2017.00242
- Jhanwar-Uniyal, M., Labagnara, M., Friedman, M., Kwasnicki, A., and Murali, R. (2015). Glioblastoma: molecular pathways, stem cells and therapeutic targets. *Cancers* 7, 538–555. doi: 10.3390/cancers7020538
- Lambiv, W. L., Vassallo, I., Delorenzi, M., Shay, T., Diserens, A. C., and Misra, A. (2011). The Wnt inhibitory factor 1 (WIF1) is targeted in glioblastoma and has a tumor suppressing function potentially by induction of senescence. *Neurooncology* 13, 736–747. doi: 10.1093/neuonc/nor036
- Lan, M. S., Modi, W. S., Xie, H., and Notkins, A. L. (1996). Assignment of the IA-2 gene encoding an autoantigen in IDDM to chromosome 2q35. *Diabetologia* 39, 1001–1002. doi: 10.1007/s001250050545
- Lee, Y., Scheck, A. C., Cloughesy, T. F., Lai, A., Dong, J., and Farooqi, H. K. (2008). Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC Med. Genomics* 1:52. doi: 10.1186/1755-8794-1-52
- Li, J., Lan, C. N., Kong, Y., Feng, S. S., and Huang, T. (2018). Identification and analysis of blood gene expression signature for osteoarthritis with advanced feature selection methods. *Front. Genet.* 9:246. doi: 10.3389/fgene.2018.00246
- Liu, S., Han, Y., Zhang, T., and Yang, Z. (2011). Protective effect of trifluoperazine on hydrogen peroxide-induced apoptosis in PC12 cells. *Brain Res. Bull.* 84, 183–188. doi: 10.1016/j.brainresbull.2010.12.008
- Liu, Y., Zhou, Y., and Zhu, K. (2012). Inhibition of glioma cell lysosome exocytosis inhibits glioma invasion. *PLoS One* 7:e45910. doi: 10.1371/journal.pone.0045910
- Luo, D., Deng, B., Weng, M., Luo, Z., and Nie, X. (2018). A prognostic 4-lncRNA expression signature for lung squamous cell carcinoma. *Artif. Cells Nanomed. Biotechnol.* 46, 1207–1214. doi: 10.1080/21691401.2017.1366334
- Ma, Q., Long, W., Xing, C., Chu, J., Luo, M., Wang, H. Y., et al. (2018). Cancer stem cells and immunosuppressive microenvironment in Glioma. *Front. Immunol.* 9:2924. doi: 10.3389/fimmu.2018.02924
- Mao, X., Qin, X., Li, L., Zhou, J., Zhou, M., Li, X., et al. (2018). A 15-long non-coding RNA signature to improve prognosis prediction of cervical squamous cell carcinoma. *Gynecol. Oncol.* 149, 181–187. doi: 10.1016/j.ygyno.2017.12.011
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34.
- O'Quigley, J., and Moreau, T. (1986). Cox's regression model: computing a goodness of fit statistic. *Comput. Methods Programs Biomed.* 22, 253–256. doi: 10.1016/0169-2607(86)90001-5
- Ostrom, Q. T., Gittleman, H., Farah, P., Ondracek, A., Chen, Y., Wolinsky, Y., et al. (2013). CBRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2006–2010. *Neurooncology* 15(Suppl. 2), iii1–ii56. doi: 10.1093/neuonc/no151
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303
- Shergalis, A., Bankhead, A. III, Luesakul, U., Muangsins, N., and Neamati, N. (2018). Current challenges and opportunities in treating glioblastoma. *Pharmacol. Rev.* 70, 412–445. doi: 10.1124/pr.117.014944
- Szopa, W., Burley, T. A., Kramer-Marek, G., and Kaspera, W. (2017). Diagnostic and therapeutic biomarkers in glioblastoma: current status and future perspectives. *BioMed Res. Int.* 2017:8013575. doi: 10.1155/2017/8013575
- Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, W98–W102. doi: 10.1093/nar/gkx247
- Terakawa, Y., Agnihotri, S., Golbourn, B., Nadi, M., Sabha, N., and Smith, C. A. (2013). The role of drebrin in glioma migration and invasion. *Exp. Cell Res.* 319, 517–528. doi: 10.1016/j.yexcr.2012.11.008
- van den Bent, M. J., Weller, M., Wen, P. Y., Kros, J. M., Aldape, K., and Chang, S. (2017). A clinical perspective on the 2016 WHO brain tumor classification and routine molecular diagnostics. *Neurooncology* 19, 614–624. doi: 10.1093/neuonc/now277
- Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020
- Wade, A., Robinson, A. E., Engler, J. R., Petritsch, C., James, C. D., and Phillips, J. J. (2013). Proteoglycans and their roles in brain cancer. *FEBS J.* 280, 2399–2417. doi: 10.1111/febs.12109
- Wang, S., Feng, C., Xie, Y., Ye, L., Wang, F., and Li, X. (2016). Sample level enrichment analysis of KEGG pathways identifies clinically relevant subtypes of Glioblastoma. *J. Cancer* 7, 1701–1710. doi: 10.7150/jca.15486
- Wei, Z., Batagov, A. O., Schinelli, S., Wang, J., Wang, Y., El Fatimy, R., et al. (2017). Coding and noncoding landscape of extracellular RNA released by human glioma stem cells. *Nat. Commun.* 8:1145. doi: 10.1038/s41467-017-01196-x
- Wen, P. Y., and Kesari, S. (2008). Malignant gliomas in adults. *New Engl. J. Med.* 359, 492–507. doi: 10.1056/nejmra0708126
- Xu, H., Cai, T., Carmona, G. N., Abuhatzira, L., and Notkins, A. L. (2016). Small cell lung cancer growth is inhibited by miR-342 through its effect of the target gene IA-2. *J. Transl. Med.* 14:278.

- Yang, R., Xiong, J., Deng, D., Wang, Y., Liu, H., Jiang, G., et al. (2016). An integrated model of clinical information and gene expression for prediction of survival in ovarian cancer patients. *Transl. Res.* 172, 84.e11–95.e11.
- Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118
- Zeng, W. J., Yang, Y. L., Liu, Z. Z., Wen, Z. P., Chen, Y. H., Hu, X. L., et al. (2018). Integrative analysis of DNA methylation and gene expression identify a three-gene signature for predicting prognosis in lower-grade gliomas. *Cell. Physiol. Biochem.* 47, 428–439. doi: 10.1159/000489954
- Zhangyuan, G., Yin, Y., Zhang, W., Yu, W., Jin, K., Wang, F., et al. (2018). Prognostic value of phosphotyrosine phosphatases in hepatocellular carcinoma. *Cell. Physiol. Biochem.* 46, 2335–2346. doi: 10.1159/000489625
- Zhao, S., Lin, Y., Xu, W., Jiang, W., Zha, Z., Wang, P., et al. (2009). Glioma-derived mutations in IDH1 dominantly inhibit IDH1 catalytic activity and induce HIF-1alpha. *Science* 324, 261–265. doi: 10.1126/science.1170944

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Yin, Tang, Zhou, Cao, Li, Fu, Wu and Jiang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# Co-expression Network Analysis Identifies Four Hub Genes Associated With Prognosis in Soft Tissue Sarcoma

Zhenhua Zhu<sup>1</sup>, Zheng Jin<sup>2</sup>, Yuyou Deng<sup>3</sup>, Lai Wei<sup>4</sup>, Xiaowei Yuan<sup>1</sup>, Mei Zhang<sup>5\*</sup> and Dahui Sun<sup>1\*</sup>

<sup>1</sup> Department of Orthopaedic Trauma, The First Hospital of Jilin University, Changchun, China, <sup>2</sup> Department of Immunology, College of Basic Medical Sciences, Jilin University, Changchun, China, <sup>3</sup> Department of Urology, The First Hospital of Jilin University, Changchun, China, <sup>4</sup> College of Computer and Control Engineering, Nankai University, Tianjin, China, <sup>5</sup> College of Chemistry, Jilin University, Changchun, China

## OPEN ACCESS

### Edited by:

Monica Bianchini,  
Università degli Studi di Siena, Italy

### Reviewed by:

Haibo Liu,  
Iowa State University, United States  
Rahul Kumar,  
Columbia University Irving Medical  
Center, United States

### \*Correspondence:

Mei Zhang  
zhangmei@jlu.edu.cn  
Dahui Sun  
sundahui1971@sina.com

### Specialty section:

This article was submitted to  
Bioinformatics and Computational  
Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 07 August 2018

**Accepted:** 18 January 2019

**Published:** 04 February 2019

### Citation:

Zhu Z, Jin Z, Deng Y, Wei L,  
Yuan X, Zhang M and Sun D (2019)  
Co-expression Network Analysis  
Identifies Four Hub Genes Associated  
With Prognosis in Soft Tissue  
Sarcoma. *Front. Genet.* 10:37.  
doi: 10.3389/fgene.2019.00037

**Background:** Soft tissue sarcomas (STS) are heterogeneous tumors derived from mesenchymal cells that differentiate into soft tissues. The prognosis of patients who present with an STS is influenced by the regulation of a complex gene network.

**Methods:** Weighted gene co-expression network analysis (WGCNA) was performed to identify gene modules associated with STS (Samples = 156).

**Results:** Among the 11 modules identified, the black and blue modules were highly correlated with STS. However, using preservation analysis, the black module demonstrated low preservation, therefore the blue module was chosen as the module of interest. Furthermore, a total of 20 network hub genes were identified in the blue module, 12 of which were also hub nodes in the protein-protein interaction network of the module genes. Following additional verification, 4 of 12 genes (*RRM2*, *BUB1B*, *CENPF*, and *KIF20A*) demonstrated poorer overall survival and disease-free survival rate in the test datasets. In addition, gene set enrichment analysis (GSEA) demonstrated that samples with a high level of blue module eigengene (ME) were enriched in cell cycle and metabolism associated signaling pathways.

**Conclusion:** In summary, co-expression network analysis identified four hub genes associated with prognosis for STS, which may diminish the prognosis by influencing cell cycle and metabolism associated signaling pathways.

**Keywords:** soft tissue sarcoma, weighted gene co-expression analysis, *RRM2*, *BUB1B*, *CENPF*, *KIF20A*

## INTRODUCTION

Soft tissue sarcoma (STS) is a rare group of tumors that accounts for approximately 1% of adult cancers. In 2009, it was estimated that 3,300 new cases were diagnosed in Britain and 10,000 in the United States (Linch et al., 2014). There are approximately 50 STS subtypes, which differ significantly in their disease presentation, response to currently available treatments and risk of tumor progression (Casali et al., 2018). Multiple factors have been reported to be related to the

progression of STS, including capillary morphogenesis gene 2 (CMG2) (Greither et al., 2017), HIF-2 $\alpha$  protein (Nakazawa et al., 2016), epidermal growth factor receptor (EGFR) protein (Yang et al., 2017) and microRNAs (Smolle et al., 2017). However, no molecular biomarkers have been defined for predicting the prognosis of the disease in clinical. Therefore, a better understanding of the molecular pathogenesis is required.

To date, microarray-based expression data have been used to identify genes related to tumor progression and prognosis. Takahashi et al. (2014) identified 25 survival-associated genes using a knowledge-based filtering and multiple testing approach. Beck et al. (2010) has reviewed the manner in which gene expression profiling has been used to understand sarcoma pathobiology and identify clinically useful biomarkers. However, most studies have focused on screening genes that have different patterns of expression with explanations gained from gene ontology (GO) analysis. Such approaches, however, have failed to address the large number of interconnections between genes, because genes with similar expression profiles are most likely to function closely together. Therefore, weighted gene co-expression network analysis (WGCNA) clusters genes co-expressed in a network, based on similarities in expression profiles among samples and in clinical traits, to define sub-network regions (known as modules) (Langfelder and Horvath, 2008).

In this study, we utilized WGCNA to identify the most relevant module in STS. Key genes in the module were identified and validated using survival and protein-protein interaction (PPI) analyses. These key genes may shed new light on the biological mechanisms underlying STS progression and could potentially be used as prognostic biomarkers or therapeutic targets.

## MATERIALS AND METHODS

### Study Design and Data Collection

Study design, data preparation, preprocessing, analysis and validation are described in a flowchart (Figure 1). Core codes used to reproduce the results were provided in **Supplementary Table S1**. Firstly, normalized RNAseq data and associated clinical data were downloaded from the NCBI Gene Expression Omnibus (GEO). Dataset GSE21122 (Barretina et al., 2010), which was generated using an Affymetrix human genome U133A microarray (HG-U133A), was used as a training set to construct the co-expression network and identify key modules in this study. This dataset included 149 STS samples and 9 normal fat tissue samples. The STS samples contained 116 different types of liposarcoma and 34 malignant fibrous histiocytomas (MFHs). Most STSs (68.8%) were primary tumors at the time of sample procurement from patients whose mean age was 56 years. In addition, two test datasets were used to test the preservation of identified modules and survival significance of hub genes. The first one, which included RNA sequencing data and associated clinical information of 265 STS samples, were downloaded from The Cancer Genome Atlas (TCGA) database<sup>1</sup>. The other one,

GSE21050 dataset (Chibon et al., 2010), which included RNA sequencing data and associated clinical information of 310 STS samples were downloaded from the NCBI GEO.

### Data Preprocessing

Firstly, we extracted training expression data from the GSE21122 MINiML file. The expression data was background corrected using the Robust Multi-array Average (RMA) algorithm and log base 2 normalized. The data were then checked to ascertain whether there was a batch effect. No apparent batch effect was observed after analysis of expression clusters, box plots and principal components analysis (PCA) (**Supplementary Figure S1**). In order to detect outliers for WGCNA analysis, sample network was calculated based on squared Euclidean distance. The connectivity of each sample was defined as the sum of the connectivity of that sample with all other samples. Outliers were identified after normalization of the connectivity of each sample, by use of the threshold  $z.k < 0.6$ . Generally, genes whose expression varies greatly are more biologically relevant. To reduce background noise, we selected genes that were varied expressed across samples and removed those whose expression was the same across samples. The median absolute deviation (MAD) was calculated for each gene as a robust measure of variability. Then, genes were sorted based on the MAD value and the top 3,000 ranked genes were used for the subsequent WGCNA analysis.

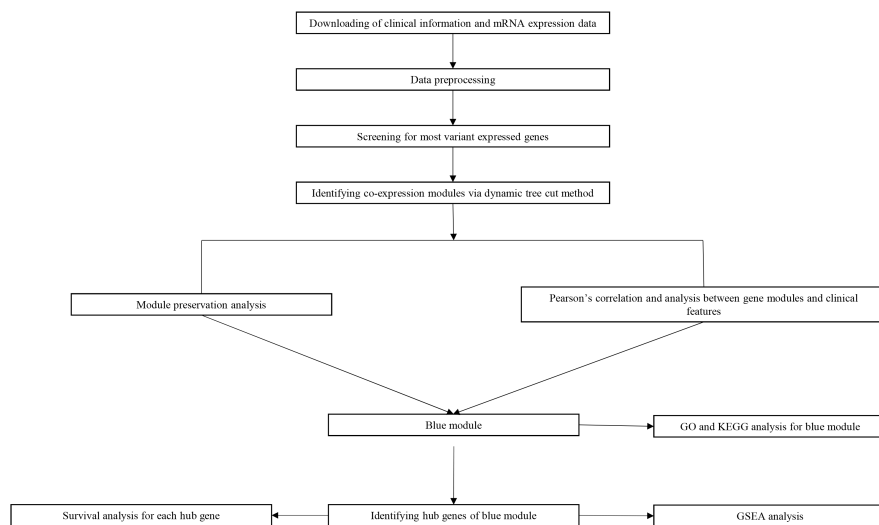
### Co-expression Network Construction and Module Preservation Analysis

The WGCNA package (Langfelder and Horvath, 2008) was used to construct the co-expression network. The concordance of genes in the expression dataset was measured with Pearson correlation, then the Pearson correlation matrix was transformed to weighted network with the power adjacency function. The first step in this process was selection of an appropriate soft power, in which strong connections between genes are promoted and weak connections penalized, so as to transform the network into one meeting the requirements of a scale-free network. Modules were identified using the dynamic tree-cutting function with a deepSplit argument value of 2 and a minimum size cutoff of 30. To test whether the identified modules were stable in the test TCGA dataset, the downloaded fragments per million (FPKM) expression data of 265 samples were transformed to the transcripts per million (TPM). A total of 2704 common genes in the training and TCGA datasets were used for preservation analysis. The module Preservation function (nPermutations = 200) of the WGCNA package (Langfelder et al., 2011) was utilized, in which the preservation statistic Zsummary was used to quantify the preservation of gene modules between datasets.

### Finding Modules of Interest and Functional Annotation

Because the module eigengene (ME) provides the most appropriate synopsis of gene expression profiles of any given module, we correlated MEs with clinical traits. In this study,

<sup>1</sup><https://genome-cancer.ucsc.edu/>



**FIGURE 1 |** Flow diagram of strategy for data preparation, preprocessing and analysis used in this study.

clinical traits refer to whether the sample was a STS or normal fat tissue. Correlations were then calculated using linear regression model. The modules for which the eigengenes showed high correlation were chosen as the modules of interest. In an attempt to ascertain possible mechanisms of genes within a module

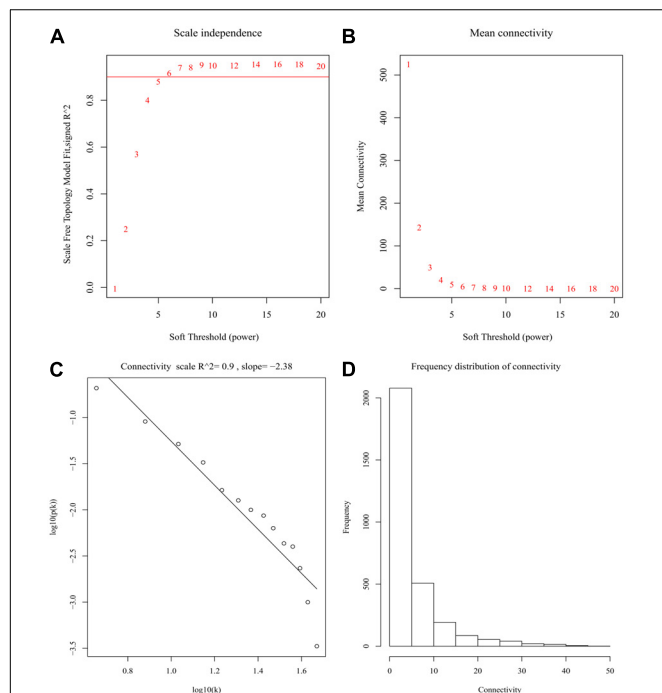
affecting STS progression, functional enrichment analyses using the KEGG and GO databases of the hub module was performed with the “clusterProfile” package in R (Yu et al., 2012).

## Identification of Hub Genes and Correlation Analysis

Hub genes are those that have a high degree of intra-module connectivity. In this study, hub genes were defined as the 20 module genes with highest connectivity in the interested module. A PPI network was constructed in order to identify hub nodes by uploading all genes in the hub module to the Search Tool for the Retrieval of Interacting Gene (STRING) database<sup>2</sup>. The PPI network was then imported into the Cytoscape software platform and a comprehensive analysis of the relationship between nodes was performed using the Maximal Clique Centrality (MCC) function, reported to be the most effective method of finding hub nodes in a co-expression network (Chin et al., 2014), within the “cytoHubba” application. In this way, the most cohesive genes were marked as “first stage nodes.” In the PPI network of blue module genes, the 30 most highly ranked nodes were identified as “first stage nodes.” Genes that were defined as both hub genes in the module and “first stage nodes” in the PPI network were chosen as primary hub genes.

## Survival Analysis and Efficacy Evaluation

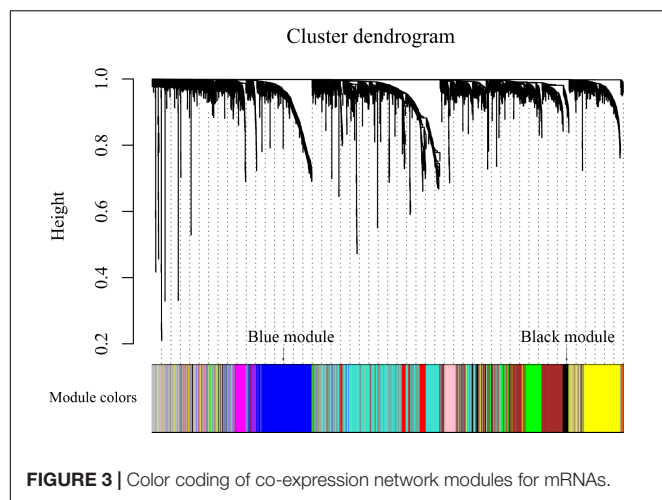
The internet tool, Gene Expression Profiling Interactive Analysis (GEPIA)<sup>3</sup>, was used to perform overall survival and disease-free survival analyses for all hub genes. The platform utilizes all expression data and survival information of the TCGA database. Users are able to accomplish survival analysis by simply submitting a gene name and selecting a tumor type. Patients were divided into two groups (high vs. low) based



**FIGURE 2 |** Determination of soft-thresholding power in the weighted gene co-expression network analysis (WGCNA). **(A)** Analysis of scale-free fit index for various soft-thresholding powers ( $\beta$ ). **(B)** Analysis of mean connectivity for various soft-thresholding powers. **(C)** Linear model fitting of  $R^2$  index showed good quality of fit. **(D)** Frequency distribution of connectivity.

<sup>2</sup><http://www.string-db.org>

<sup>3</sup><http://gepia.cancer-pku.cn>



on the hub gene expression level in comparison to the mean expression level of that hub gene. Furthermore, dataset GSE21050, which includes 310 STS samples in which metastasis status and survival time were provided, was used to test the significance of hub genes for metastasis survival. A Kaplan-Meier survival plot was constructed using the “survival” package in R (Li, 2003). Differential expression between STS and normal tissue in the training set was plotted as a box plot graph.

### Gene Set Enrichment Analysis (GSEA)

In the training data set, 156 samples were dichotomized into two groups (High vs. Low) based on the ME value of blue module in comparison to the mean ME level of blue module of all

samples. GSEA was then performed between the two groups. The 3,000 most variable genes from the WGCNA were imported for enrichment. In this way, GSEA was used to validate the results of GO and KEGG analysis of the blue module. The cut-off criterion for GSEA was  $FDR < 0.05$ .

## RESULTS

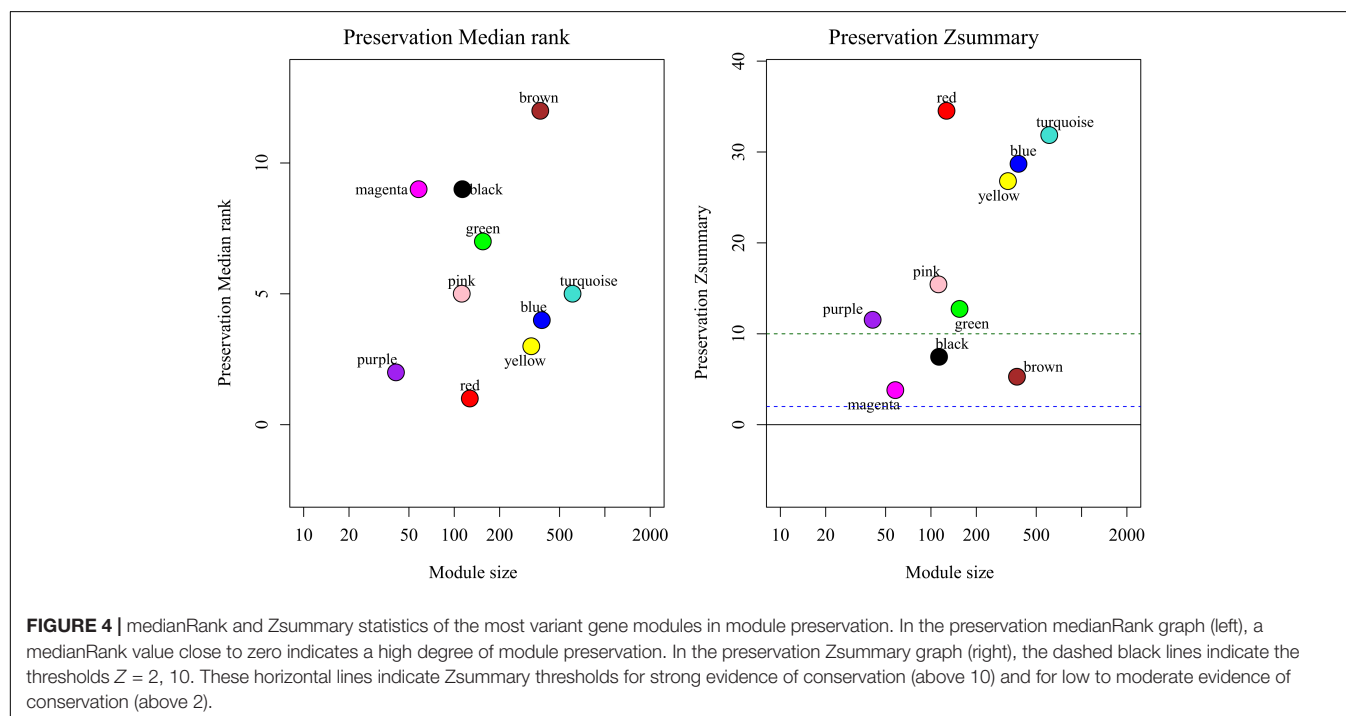
### Co-expression Network Construction and Module Preservation Analysis

After discarding two outlier samples (GSM528297 and GSM528333), WGCNA was performed on the 3,000 most variable genes of 156 samples. Soft threshold power was set to 6, in which  $R^2$  was 0.916, ensured a scale-free network (Figure 2). Following this, 11 co-expression modules were identified, ranging in size from 43 to 669 genes (with each module assigned a color) (Figure 3).

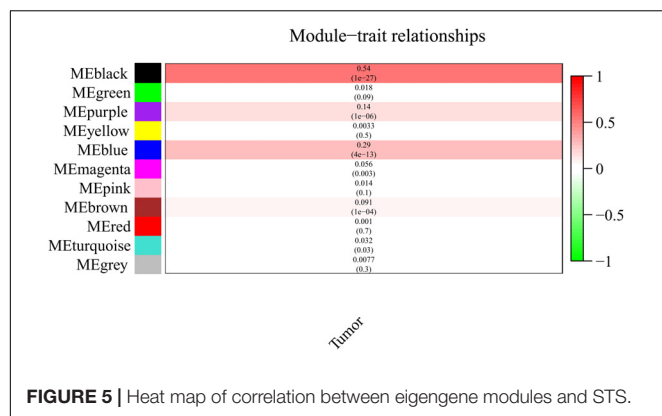
By comparing the training dataset GSE21122 with the TCGA test dataset, we were able to establish whether the co-expression modules produced in the training dataset could be reproduced in the test dataset through summary preservation statistics. Three modules (black, brown, and magenta) demonstrated poor preservation with each Zsummary statistic  $< 10$ . The remaining modules, including the blue module were stable enough, suggesting they were preserved between the training data set and the test data set (Figure 4).

### Finding Modules of Interest and Functional Annotation

It is important to identify the most significant modules related to STS. Both black and blue modules showed a significantly high







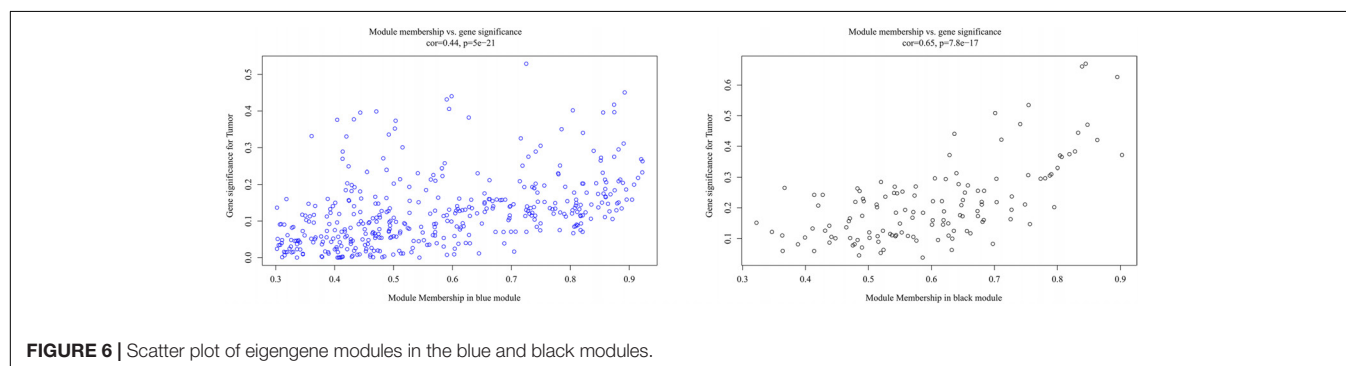
**FIGURE 5 |** Heat map of correlation between eigengene modules and STS.

correlation with sarcomas (Figures 5, 6). However, due to the lack of stability of the statistical data ( $Z_{\text{summary}} < 10$ ), the black module was not further analyzed. Therefore, the blue module was defined as an important module of clinical significance and extracted for further analysis.

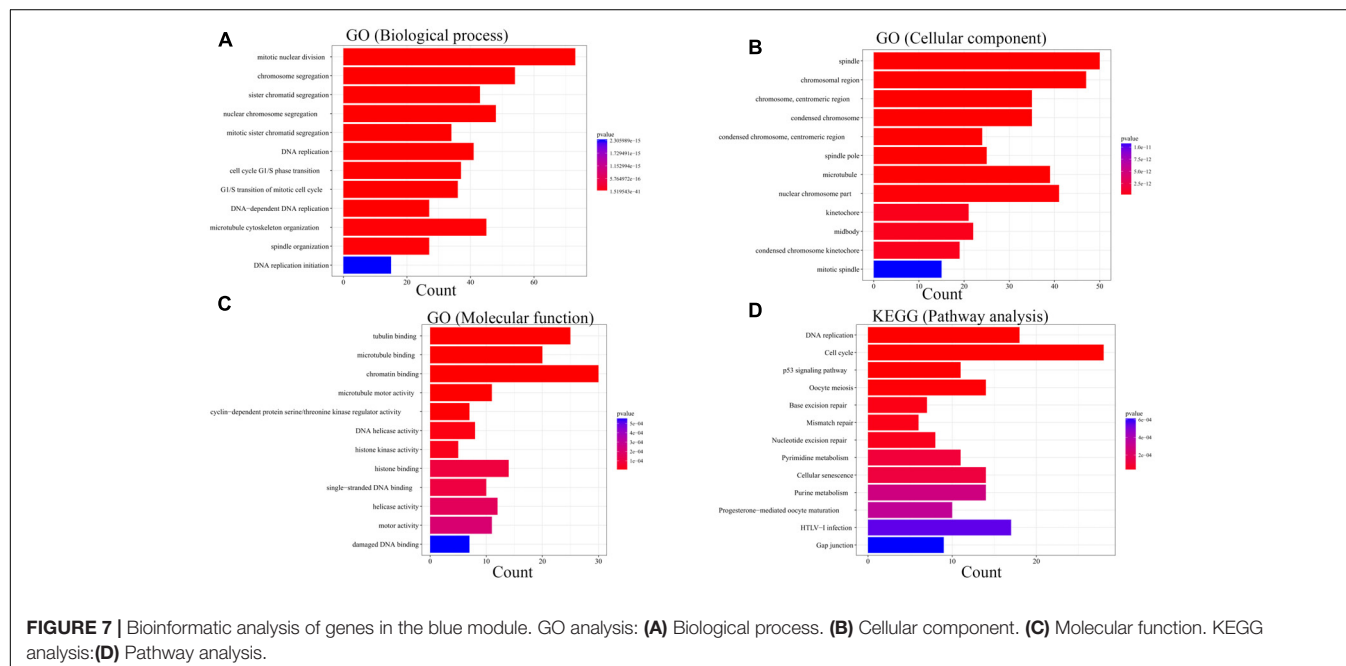
For the sake of exploration of the biological relevance of the blue module, GO functional and KEGG pathway enrichment analyses were performed on 414 genes in the blue module. The biological processes of the genes in the blue module were found to associate with the cell cycle, such as mitotic nuclear division, chromosome segregation and sister chromatid segregation. In the KEGG pathway analysis, cell cycle associated signaling pathways such as DNA replication, cell cycle, p53 signaling pathway, oocyte meiosis, mismatch repair and metabolism associated pathways such as pyrimidine metabolism and purine metabolism were enriched (Figure 7).

## Identification of Sarcoma Hub Genes in the Blue Module

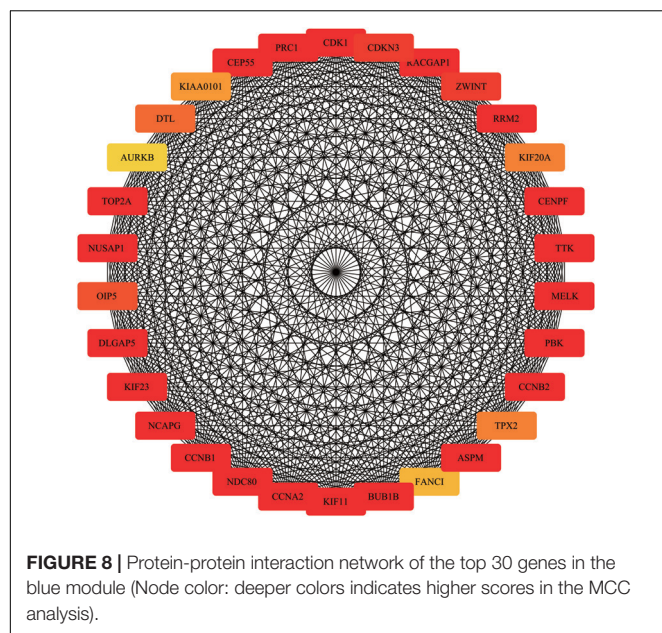
Highly connected hub genes within a module perform important roles in tumor biological processes. Therefore, the 20 genes with greatest module relevance in the blue module were selected as candidate hub genes for STS (Supplementary Data Sheet S1). In addition, a PPI network in the blue module was constructed in accordance with the STRING database (Figure 8). Twelve of



**FIGURE 6 |** Scatter plot of eigengene modules in the blue and black modules.



**FIGURE 7 |** Bioinformatic analysis of genes in the blue module. GO analysis: (A) Biological process. (B) Cellular component. (C) Molecular function. KEGG analysis: (D) Pathway analysis.



the 20 candidate genes in the co-expression network were also identified as hub nodes of the PPI network. Finally, these 12 genes were considered “primary” hub genes associated with STS and therefore selected for additional analyses.

## Survival Analysis and Efficacy Evaluation

While testing the TCGA dataset, four out of 12 hub genes demonstrated significant connectivity with overall and disease-free survival (Figure 9). When testing the GSE21050 dataset, these four hub genes showed significant correlation with

metastasis free survival (Figure 10). Furthermore, they were significantly highly expressed in STS tissue compared to normal fat tissue (Figure 11).

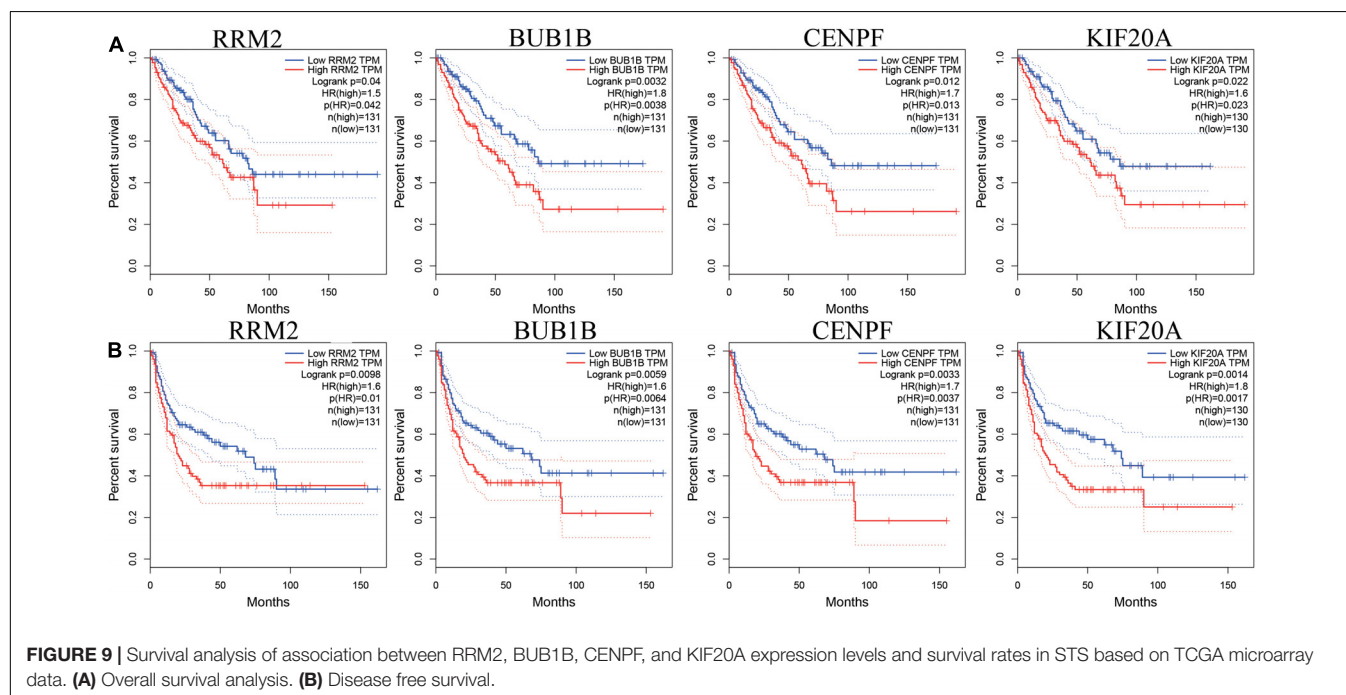
## Gene Set Enrichment Analysis

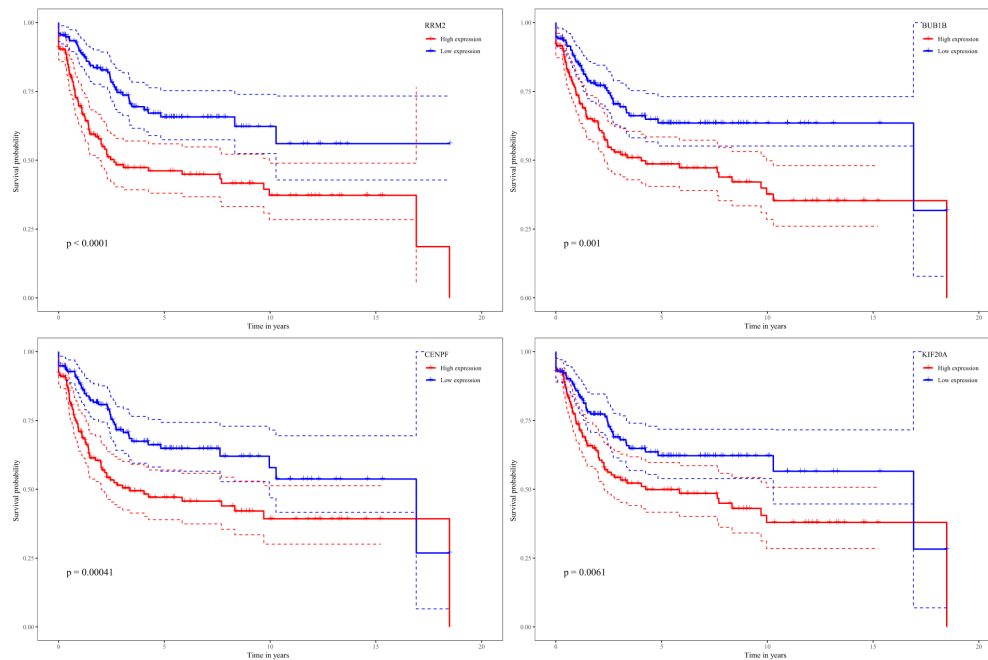
In order to find out the potential function of both blue module and hub genes, GSEA was performed to identify KEGG pathways enriched in samples with higher level of ME of blue module. In GSEA analysis, five signaling pathways were significantly enriched, including ubiquitin mediated proteolysis (FDR = 0.01), pyrimidine metabolism (FDR = 0.03), oocyte meiosis (FDR = 0.02), cell cycle (FDR = 0.04) and DNA replication (FDR = 0.04) (Figure 12). Moreover, the last four pathways were consistent with the results of KEGG pathway analysis (Figure 7D).

## DISCUSSION

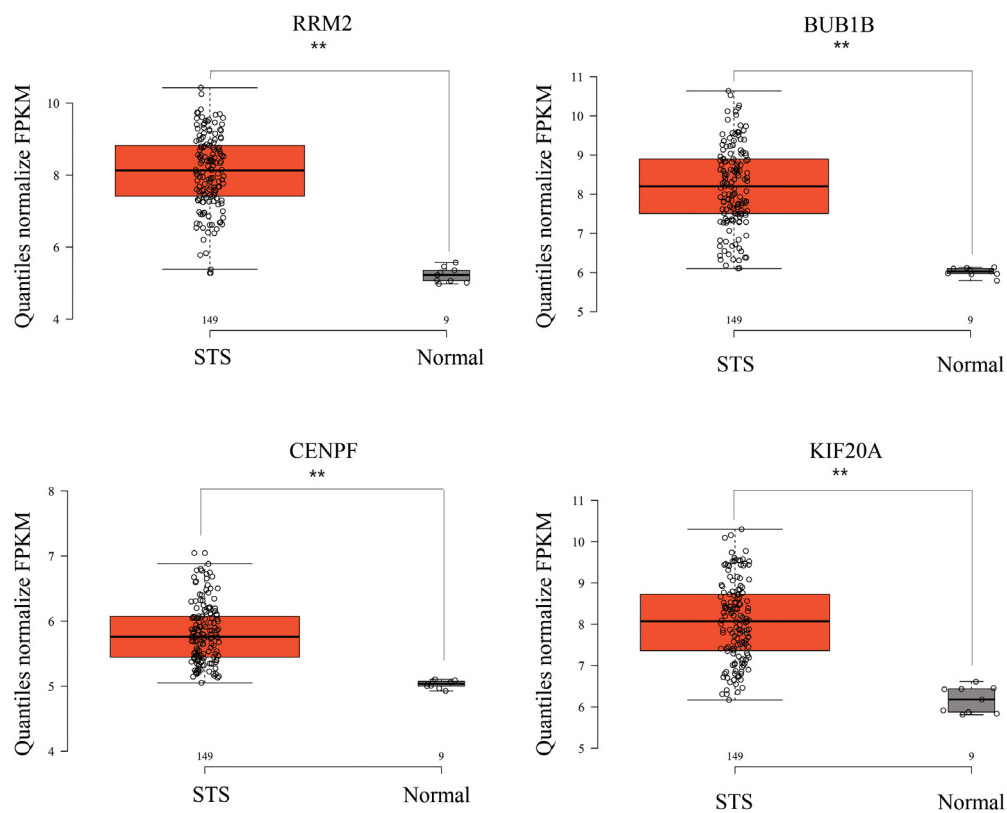
Soft tissue sarcomas remain among the most challenging diseases for medical oncologists to treat. STSs are mesenchymal neoplasms that can arise from any site within the body, including extremities, the trunk, retroperitoneum, head, and neck. These are biologically heterogeneous diseases of which greater than 50 subtypes exist, varying by molecular, histological and clinical characteristics.

In this study, WGCNA was utilized to construct a co-expression network for identification of gene co-expression modules associated with STS. The blue module was positively identified and 20 hub genes selected from this module. In addition, as a result of the PPI network, 12 genes were identified as hub nodes of the co-expression module and PPI network, indicating that these 12 hub genes were closely

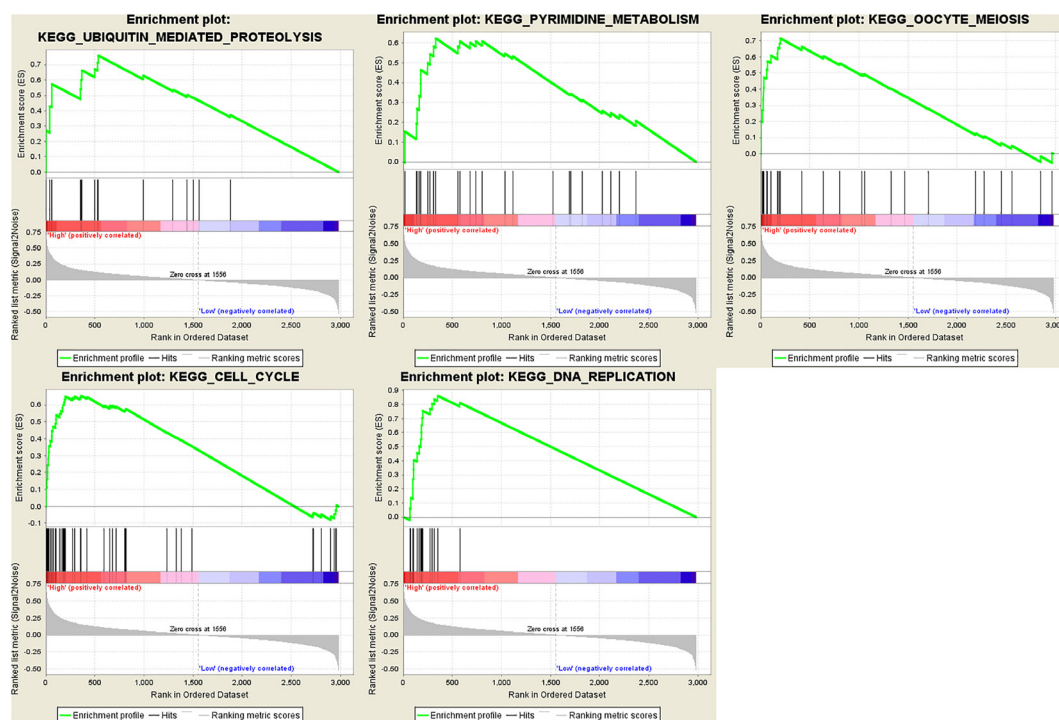




**FIGURE 10 |** Survival analysis of association between RRM2, BUB1B, CENPF, and KIF20A expression levels and metastasis-free survival rates in STS based on GSE21050 microarray data.



**FIGURE 11 |** RRM2, BUB1B, CENPF, and KIF20A were strongly upregulated in STS tissues compared to normal fat tissue, based on GSE21122 microarray data.  $**p < 0.01$ .



**FIGURE 12 |** Gene set enrichment analysis (GSEA). Cell cycle and metabolism associated pathways were enriched.

related to STS and had important biological significance. Subsequent survival analysis established that four of the 12 hub genes (*RRM2*, *BUB1B*, *CENPF*, and *KIF20A*) were significantly associated with survival. We, therefore, focused on these four genes.

The ribonucleotide reductase regulatory subunit M2 (*RRM2*) is one of two subunits that constitute ribonucleotide reductase, the enzyme responsible for catalyzing the conversion of ribonucleotides into deoxyribonucleotides, and thus performing an important role in DNA synthesis. *RRM2* is important in controlling cellular function in a number of human malignant tumors, including DNA repair, cell proliferation and senescence. Importantly, *RRM2* functions as a driver in a variety of tumors, with *in vivo* and *in vitro* experiments confirming that knocking down expression using siRNA significantly inhibits tumor cell proliferation (Fang et al., 2016).

The *BUB1* mitotic checkpoint serine/threonine kinase B (*BUB1B*) is a member of the spindle assembly checkpoint protein family, crucial for ensuring correct chromosome separation during cell division (Fu et al., 2016). *BUB1B* performs a role in the inhibition of APC expression, established as a tumor suppressor gene in most colorectal cancers. Accordingly, many reports have shown that upregulation of *BUB1B* is related to the recurrence and progression of bladder cancer (Yamamoto et al., 2007), gastric cancer (Ando et al., 2010), esophageal squamous cell carcinoma (Tanaka et al., 2008), breast cancer (Yuan et al., 2006), hepatocellular carcinoma (Liu et al., 2009) and others.

Centromere protein F (*CENPF*) is another important protein involved in chromosome segregation during mitosis.

Upregulation of *CENPF* protein expression, especially through a gene amplification effect, suggests that high levels of *CENPF* protein may affect the occurrence of tumors, especially in the early stages of tumor development (Varis et al., 2006). Clinical research has demonstrated that high expression levels of *CENPF* results in poor prognosis in nasopharyngeal carcinoma (Cao et al., 2010), colorectal gastrointestinal stromal tumors (Chen et al., 2011), esophageal squamous cell carcinoma (Mi et al., 2013) and prostate cancer (Zhuo et al., 2015). It has also been shown to play an important role in driving hepatocellular carcinoma (Dai et al., 2013).

Kinesin family member 20A (*KIF20A*, also known as *RAB6KIFL*) belongs to the kinesin superfamily-6, located in the Golgi apparatus and contributes to intracellular organelle transport and cell division (Echard et al., 1998). Recently, it has been reported that *KIF20A* is associated with mitosis, cell adhesion, migration and proliferation. Furthermore, recent studies have demonstrated that *KIF20A* is involved in tumor progression and angiogenesis. High expression of *KIF20A* results poor prognosis in glioma patients (Duan et al., 2016; Saito et al., 2017), nasopharyngeal cancer (Liu et al., 2017), hepatocellular carcinoma (Shi et al., 2016), melanoma (Yamashita et al., 2012) and early-stage cervical squamous cell carcinoma (Zhang et al., 2016).

Regarding GSEA, it was found that cell cycle and metabolism associated pathways were significant enriched in samples with higher level of ME of blue module. This is consistent with the initial GO and KEGG analysis results of the blue module



and are related to the physiological function of these four hub genes.

In summary, through WGCNA and other related analysis methods, we identified four genes (*RRM2*, *BUB1B*, *CENPF*, and *KIF20A*) related to the progression and prognosis of STS. These genes may play a role by regulating the cell cycle and metabolism associated signaling pathways.

## AUTHOR CONTRIBUTIONS

ZZ and DS designed the study. ZZ and ZJ performed the data collection. ZJ and LW performed the data analysis. ZZ and MZ drafted the manuscript. All authors read and approved the final version of the manuscript.

## REFERENCES

- Ando, K., Kakeji, Y., Kitao, H., Iimori, M., Zhao, Y., Yoshida, R., et al. (2010). High expression of *BUBR1* is one of the factors for inducing DNA aneuploidy and progression in gastric cancer. *Cancer Science*. 101, 639–645. doi: 10.1111/j.1349-7006.2009.01457.x
- Barretina, J., Taylor, B. S., Banerji, S., Ramos, A. H., Lagos-Quintana, M., Decarolis, P. L., et al. (2010). Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nat. Genet.* 42, 715–721. doi: 10.1038/ng.619
- Beck, A. H., West, R. B., and van de Rijn, M. (2010). Gene expression profiling for the investigation of soft tissue sarcoma pathogenesis and the identification of diagnostic, prognostic, and predictive biomarkers. *Virchows Archiv.* 456, 141–151. doi: 10.1007/s00428-009-0774-2
- Cao, J. Y., Liu, L., Chen, S. P., Zhang, X., Mi, Y. J., Liu, Z. G., et al. (2010). Prognostic significance and therapeutic implications of centromere protein F expression in human nasopharyngeal carcinoma. *Mol. Cancer* 9, :237. doi: 10.1186/1476-4598-9-237
- Casali, P. G., Abecassis, N., Bauer, S., Biagini, R., Bielack, S., Bonvalot, S., et al. (2018). Soft tissue and visceral sarcomas: ESMO-EURACAN clinical practice guidelines for diagnosis, treatment and follow-up. *Annals. of Oncology*. 29, 51–67. doi: 10.1093/annonc/mdy096
- Chen, W. B., Cheng, X. B., Ding, W., Wang, Y. J., Chen, D., Wang, J. H., et al. (2011). Centromere protein F and survivin are associated with high risk and a poor prognosis in colorectal gastrointestinal stromal tumours. *J. Clin. Pathol.* 64, 751–755. doi: 10.1136/jcp.2011.089631
- Chibon, F., Lagarde, P., Salas, S., Perot, G., Brouste, V., Tirode, F., et al. (2010). Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. *Nat. Med.* 16, 781–787. doi: 10.1038/nm.2174
- Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., and Lin, C. Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8(Suppl. 4):S11. doi: 10.1186/1752-0509-8-S4-S11
- Dai, Y., Liu, L., Zeng, T., Zhu, Y. H., Li, J., Chen, L., et al. (2013). Characterization of the oncogenic function of centromere protein F in hepatocellular carcinoma. *Biochem. Biophys. Res. Commun.* 436, 711–718. doi: 10.1016/j.bbrc.2013.06.021
- Duan, J., Huang, W., and Shi, H. (2016). Positive expression of *KIF20A* indicates poor prognosis of glioma patients. *Onco Targets Ther.* 9, 6741–6749. doi: 10.2147/OTT.S115974
- Echard, A., Jollivet, F., Martinez, O., Lacapere, J. J., Rousselet, A., Janoueix-Lerosey, I., et al. (1998). Interaction of a Golgi-associated kinesin-like protein with *Rab6*. *Science* 279, 580–585. doi: 10.1126/science.279.5350.580
- Fang, Z., Lin, A., Chen, J., Zhang, X., Liu, H., Li, H., et al. (2016). *CREB1* directly activates the transcription of ribonucleotide reductase small subunit *M2* and promotes the aggressiveness of human colorectal cancer. *Oncotarget* 7, 78055–78068. doi: 10.18632/oncotarget.12938
- Fu, X., Chen, G., Cai, Z. D., Wang, C., Liu, Z. Z., Lin, Z. Y., et al. (2016). Overexpression of *BUB1B* contributes to progression of prostate cancer and

## FUNDING

This study was supported by the Special Projects of Health in Jilin Province (3D5148273428).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.00037/full#supplementary-material>

**FIGURE S1** | Data quality examination.

**TABLE S1** | Code for WGCNA.

**DATA SHEET S1** | Hub genes of blue module.

- predicts poor outcome in patients with prostate cancer. *Onco Targets Ther.* 9, 2211–2220. doi: 10.2147/OTT.S101994
- Greither, T., Wedler, A., Rot, S., Kessler, J., Kehlen, A., Holzhausen, H. J., et al. (2017). *CMG2* expression is an independent prognostic factor for soft tissue sarcoma patients. *International Journal. of Molecular. Sciences*. 18, :E2648. doi: 10.3390/ijms18122648
- Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559
- Langfelder, P., Luo, R., Oldham, M. C., and Horvath, S. (2011). Is my network module preserved and reproducible? *PLoS Comput. Biol.* 7:e1001057. doi: 10.1371/journal.pcbi.1001057
- Li, J. C. A. (2003). Modeling survival data: extending the Cox model. *Sociological Methods & Research*. 32, 117–120. doi: 10.1177/0049124103031004005
- Linch, M., Miah, A. B., Thway, K., Judson, I. R., and Benson, C. (2014). Systemic treatment of soft-tissue sarcoma-gold standard and novel therapies. *Nat. Rev. Clin. Oncol.* 11, 187–202. doi: 10.1038/nrclinonc.2014.26
- Liu, A. W., Cai, J., Zhao, X. L., Xu, A. M., Fu, H. Q., Nian, H., et al. (2009). The clinicopathological significance of *BUBR1* overexpression in hepatocellular carcinoma. *J. Clin. Pathol.* 62, 1003–1008. doi: 10.1136/jcp.2009.066944
- Liu, S. L., Lin, H. X., Qiu, F., Zhang, W. J., Niu, C. H., Wen, W., et al. (2017). Overexpression of kinesin family member 20A correlates with disease progression and poor prognosis in human nasopharyngeal cancer: a retrospective analysis of 105 patients. *PLoS One* 12:e0169280. doi: 10.1371/journal.pone.0169280
- Mi, Y. J., Gao, J., Xie, J. D., Cao, J. Y., Cui, S. X., Gao, H. J., et al. (2013). Prognostic relevance and therapeutic implications of centromere protein F expression in patients with esophageal squamous cell carcinoma. *Dis. Esophagus* 26, 636–643. doi: 10.1111/dote.12002
- Nakazawa, M. S., Eisinger-Mathason, T. S., Sadri, N., Ochocki, J. D., Gade, T. P., Amin, R. K., et al. (2016). Epigenetic re-expression of *HIF-2alpha* suppresses soft tissue sarcoma growth. *Nat. Commun.* 7, :10539. doi: 10.1038/ncomms10539
- Saito, K., Ohta, S., Kawakami, Y., Yoshida, K., and Toda, M. (2017). Functional analysis of *KIF20A*, a potential immunotherapeutic target for glioma. *J. Neurooncol.* 132, 63–74. doi: 10.1007/s11060-016-2360-1
- Shi, C., Huang, D. L., Lu, N. H., Chen, D., Zhang, M. H., Yan, Y. H., et al. (2016). Aberrantly activated *Gli2-KIF20A* axis is crucial for growth of hepatocellular carcinoma and predicts poor prognosis. *Oncotarget* 7, 26206–26219. doi: 10.18632/oncotarget.8441
- Smolle, M. A., Leithner, A., Posch, F., Szkandera, J., Liegl-Atzwanger, B., and Pichler, M. (2017). MicroRNAs in different histologies of soft tissue sarcoma: a comprehensive review. *International Journal. of Molecular. Sciences*. 18, :E1960. doi: 10.3390/ijms18091960
- Takahashi, A., Nakayama, R., Ishibashi, N., Doi, A., Ichinohe, R., Ikuyo, Y., et al. (2014). Analysis of gene expression profiles of soft tissue sarcoma using a combination of knowledge-based filtering with integration of multiple statistics. *PLoS One* 9:e106801. doi: 10.1371/journal.pone.0106801

- Tanaka, K., Mohri, Y., Ohi, M., Yokoe, T., Koike, Y., Morimoto, Y., et al. (2008). Mitotic checkpoint genes, hSMAD2 and BubR1, in oesophageal squamous cancer cells and their association with 5-fluorouracil and cisplatin-based radiochemotherapy. *Clinical. Oncology*. 20, 639–646. doi: 10.1016/j.clon.2008.06.010
- Varis, A., Salmela, A. L., and Kallio, M. J. (2006). Cenp-F (mitosin) is more than a mitotic marker. *Chromosoma* 115, 288–295. doi: 10.1007/s00412-005-0046-0
- Yamamoto, Y., Matsuyama, H., Chochi, Y., Okuda, M., Kawauchi, S., Inoue, R., et al. (2007). Overexpression of BUBR1 is associated with chromosomal instability in bladder cancer. *Cancer Genetics. and Cytogenetics*. 174, 42–47. doi: 10.1016/j.cancergencyto.2006.11.012
- Yamashita, J., Fukushima, S., Jinnin, M., Honda, N., Makino, K., Sakai, K., et al. (2012). Kinesin family member 20A is a novel melanoma-associated antigen. *Acta Dermato-venereologica*. 92, 593–597. doi: 10.2340/00015555-1416
- Yang, J. L., Das Gupta, R., Goldstein, D., and Crowe, P. J. (2017). Significance of phosphorylated epidermal growth factor receptor and its signal transducers in human soft tissue sarcoma. *International. Journal. of Molecular. Sciences*. 18, :E1159. doi: 10.3390/ijms18061159
- Yu, G. C., Wang, L. G., Han, Y. Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics-a Journal. of Integrative. Biology*. 16, 284–287. doi: 10.1089/omi.2011.0118
- Yuan, B. B., Xu, Y., Woo, J. H., Wang, Y. Y., Bae, Y. K., Yoon, D. S., et al. (2006). Increased expression of mitotic checkpoint genes in breast cancer cells with chromosomal instability. *Clinical. Cancer Research*. 12, 405–410. doi: 10.1158/1078-0432.Ccr-05-0903
- Zhang, W. J., He, W. L., Shi, Y. J., Gu, H. F., Li, M., Liu, Z. M., et al. (2016). High expression of KIF20A is associated with poor overall survival and tumor progression in early-stage cervical squamous cell carcinoma. *PLoS One* 11:e0167449. doi: 10.1371/journal.pone.0167449
- Zhuo, Y. J., Xi, M., Wan, Y. P., Hua, W., Liu, Y. L., Wan, S., et al. (2015). Enhanced expression of centromere protein F predicts clinical progression and prognosis in patients with prostate cancer. *International. Journal. of Molecular. Medicine*. 35, 966–972. doi: 10.3892/ijmm.2015.2086

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Zhu, Jin, Deng, Wei, Yuan, Zhang and Sun. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Long Noncoding RNA RAET1K Enhances CCNE1 Expression and Cell Cycle Arrest of Lung Adenocarcinoma Cell by Sponging miRNA-135a-5p

Chang Zheng<sup>1,2</sup>, Xuelian Li<sup>2,3</sup>, Yangwu Ren<sup>2,3</sup>, Zhihua Yin<sup>2,3</sup> and Baosen Zhou<sup>1,2\*</sup>

<sup>1</sup> Department of Clinical Epidemiology, First Affiliated Hospital of China Medical University, Shenyang, China, <sup>2</sup> Department of Epidemiology, School of Public Health, China Medical University, Shenyang, China, <sup>3</sup> Key Laboratory of Cancer Etiology and Intervention, University of Liaoning Province, Shenyang, China

## OPEN ACCESS

### Edited by:

Monica Bianchini,  
University of Siena, Italy

### Reviewed by:

Shaoli Das,  
National Institutes of Health (NIH),  
United States  
Kashmir Singh,  
Panjab University, India

### \*Correspondence:

Baosen Zhou  
bszhou@cmu.edu.cn

### Specialty section:

This article was submitted to  
Bioinformatics and  
Computational Biology,  
a section of the journal  
Frontiers in Genetics

Received: 21 August 2019

Accepted: 10 December 2019

Published: 17 January 2020

### Citation:

Zheng C, Li X, Ren Y, Yin Z and Zhou B  
(2020) Long Noncoding RNA RAET1K  
Enhances CCNE1 Expression and Cell  
Cycle Arrest of Lung Adenocarcinoma  
Cell by Sponging miRNA-135a-5p.  
Front. Genet. 10:1348.  
doi: 10.3389/fgene.2019.01348

Molecular dysregulation is believed to participate in the onset and progression of lung adenocarcinoma (LUAD). This study aimed to identify and evaluate the potential key long noncoding RNAs (lncRNAs) involved in the significant dysfunctional process of LUAD. We found that lncRNA retinoic acid early transcript 1K (RAET1K) was upregulated in tumor tissues and were correlated with a poor prognosis of patients with LUAD; further, for the first time, we detected the biological roles of RAET1K. Weighted gene correlation network and gene set enrichment analysis revealed that high RAET1K expression is related to cell cycle dysfunction through upregulated cyclin E1 (CCNE1) by targeting miR-135. The dual-luciferase reporter gene assay was performed to clarify the binding relationship between RAET1K and miR-135a-5p in transgenic A549 and H1299 cells. Real-time PCR and Western blot analyses showed that RAET1K overexpression and miR-135a-5p inhibition exerted a strong synergistic effect on CCNE1 expression, and cell cycle flow cytometry analysis was used to confirm the arrest of A549 and H1299 cells at the G1/S phase. The lncRNA RAET1K/miR-135a-5p axis might participate in the regulation of LUAD progression by influencing CCNE1 expression and the accumulation of cells arrested at the G1/S phase boundary.

**Keywords:** RAET1K, cell cycle, lung adenocarcinoma, long noncoding RNA, gene regulatory networks

## INTRODUCTION

The latest report released by the International Agency for Research on Cancer has stated that lung cancer (LC) remains the most common and deadly form of malignancy (Siegel et al., 2017; Bray et al., 2018). In general, surgery is the best option for treating patients with early stage disease because the five-year survival rate of pathological stage I non-small cell LC (NSCLC) after lobectomy is 45%–65% (Ettinger et al., 2015). However, approximately 70% of patients are diagnosed in the late stage of the disease; therefore, the five-year survival rate of these patients is only 16.38% (Ettinger et al., 2015). Lung adenocarcinoma (LUAD) is the most common type of

NSCLC, accounting for approximately 40% of cases (Ferlay et al., 2010). Therefore, the focus of the present study was limited to the complex molecular mechanisms leading to the onset and poor prognosis of LUAD.

Dysregulation of the cell cycle result in increased cell proliferation, and the abnormal expression of cell cycle regulators can lead to tumor formation (Otto and Sicinski, 2017). Various chemotherapeutic agents have been developed to target the cell cycle (Ingham and Schwartz, 2017). For example, cisplatin is one of the most successful anticancer drugs used to nonspecifically block the cell cycle (Besse and Le Chevalier, 2012). By focusing on the complex gene networks that cause dysregulation of cell cycle regulators, a potential strategy for the treatment of LC could be developed.

Previous studies have reported that noncoding RNAs, such as long noncoding RNAs (lncRNAs) and microRNAs (miRNAs) are involved in cell cycle processes (Djebali et al., 2012). Furthermore, it has been widely reported that lncRNAs functioning as the competing endogenous RNAs (ceRNAs) could regulate cancer by sponging miRNAs (Salmena et al., 2011; Dong et al., 2018; Dong et al., 2019). Despite the rapid evolution of genomic technologies and analytical tools, the identification of novel lncRNA-related ceRNA networks affecting the cell cycle and ultimately influencing LUAD remains challenging. Therefore, the present study aimed to investigate lncRNA expression profiles of The Cancer Genome Atlas (TCGA) database *via* complex bioinformatics analysis to identify novel lncRNAs and related biological functions, which initially identified that lncRNA retinoic acid early transcript 1K (RAET1K) was significantly upregulated. Furthermore, we revealed that the upregulated expression of lncRNA RAET1K was correlated with poor prognosis in LUAD patients and facilitated cell cycle arrest at the G1 phase by functioning as a ceRNA to upregulate cyclin E1 (CCNE1).

## MATERIAL AND METHODS

### Data Sets and Preprocess

The RNA and miRNA sequence data of LUAD and corresponding clinical information were downloaded from the TCGA database (<https://cancergenome.nih.gov>). The study cohort consisted of 564 LUAD patients with level 3 Illumina HiSeq RNA sequencing (RNA-seq) data and 505 patients with level 3 miRNA sequencing (miRNA-seq) data. On the basis of the clinical traits of the patients, the samples were classified into two groups: early stage (stages I and II) and advanced stage (stages III and IV). The gene symbol and type were converted from transcript IDs of RNA-seq data with the use of Genome Reference Consortium Human Build 38 patch release 12 (GRCh38.p12) of the Ensembl genome browser (<http://asia.ensembl.org/biomart>). The DESeq2 package (Love et al., 2014) was used to normalize raw data sets and identify differentially expressed genes (DIFF-genes). The cutoff values were an absolute value of log2 fold change of  $\geq 2$  and an adjusted probability (*P*) value of  $\leq 0.01$ .

## Construction of Co-Expression Networks

The R package for weighted correlation network analysis (WGCNA) was used to build co-expression networks (Langfelder and Horvath, 2008). Significant DIFF-genes were selected to generate co-expression networks for both the early and advanced stages of NSCLC. Briefly, a connection-weighted adjacency matrix of pair-wise genes was initially built according to unsupervised classifications. In accordance with the scale-independent topological criterion, the acceptable soft threshold value was set to 5 on the basis of a correlation coefficient threshold of 0.85 (Zhang and Horvath, 2005). Thereafter, a topological overlap matrix (TOM) was initially built on the adjacency matrix. The dynamic tree cutting method was performed to cluster DIFF-genes into modules with 30 as the minimum module sizes of the genes and 0.25 as the cluster merge height, respectively. Each module contained genes with similar expression patterns. The gray module consisted of a cluster of unclassified genes. After defining the modules, the module eigengene (ME) values were calculated for all genes in each module. The correlations between the ME values and the LUAD patient clinical traits were calculated (Langfelder and Horvath, 2007). Several significantly associated gene sets were chosen for functional enrichment analysis.

## Prognostic Analysis

Survival analysis was performed with SPSS Statistics for Windows, version 17.0. (SPSS, Inc., Chicago, IL, USA). On the basis of the gene expression value of the lower or upper quartile, samples were categorized into two groups: low-exp and high-exp. The hazard ratio (HR) and estimated 95% confidence interval (CI) were calculated using the Cox proportional hazard regression model. Kaplan-Meier curves were plotted to estimate the overall survival (OS), and the log rank test was used for univariate comparisons. A *P* value  $< 0.05$  was considered statistically significant. Furthermore, a nomogram was generated using a multivariate Cox regression model to evaluate the potential prognostic signature of lncRNA RAET1K for OS of LUAD patients.

## Function Annotation and Gene Set Enrichment Analysis (GSEA)

Gene ontology (GO) enrichment analysis was performed to identify the biological processes (BPs) of the module. Relevant genes in the Database for Annotation, Visualization, and Integration Discovery (DAVID) were visualized using bubble plots. The DIFF-genes in specific modules were clustered into various Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway ontologies using the ClueGO plug-in for the visualization of nonredundant biological terms for large clusters of genes in a functionally grouped network (Bindea et al., 2009). According to the gene expression level, GSEA was performed to identify the BPs and biological functions of hub genes clustered into the modules (Subramanian et al., 2005). For miRNAs, the miRcode (Jeggari et al., 2012) database was used to identify target genes and binding sites based on seed complementarity and evolutionary conservation of the seed region of the miRNAs.



## Cell Lines and Culture Conditions

Human LUAD A549 and H1299 cell lines were routinely cultured in a Roswell Park Memorial Institute 1640 medium (Gibco, Carlsbad, CA, USA) supplemented with 10% fetal bovine serum and 100 U/ml of penicillin/streptomycin (Beijing Solarbio Science & Technology Co., Ltd., Beijing, China) in an incubator (Thermo Fisher Scientific, Waltham, MA, USA) at 37°C under an atmosphere of 5% CO<sub>2</sub>/95% air, as previously described (Zheng et al., 2018).

## Cell Transfection

Cells were inoculated into the wells of a six-well plate before transfection. The RAET1K overexpression lentivirus and a negative control (NC) lentivirus were purchased from GenePharma Co., Ltd. (Shanghai, China). The cells in each well were transfected with 10<sup>6</sup> lentiviruses. Four days later, the transfection efficiency was evaluated by determining the proportion of green fluorescent protein-positive cells. A medium supplemented with 2 µg/ml of puromycin was used to screen out the A549 and H1299 cells that were unsuccessfully transfected with the RAET1K and NC lentiviruses.

Cells were transiently transfected with a group of miR-135a-5p mimics and inhibitors (GenePharma Co., Ltd.) by using jetPRIME<sup>®</sup> transfection reagent (Polyplus-transfection S.A., Illkirch-Grattenstaden, France), as previously described (Zheng et al., 2018). The cells were harvested at 24 h after transfection for further use.

## RNA Isolation and Real-Time Polymerase Chain Reaction (RT-PCR) Analysis

Total RNA was extracted using the NucleoSpin RNA Plus kit (TaKaRa Biotechnology [Dalian] Co., Ltd., Dalian, China) in accordance with the manufacturer's protocol. RNA was reverse-transcribed to complementary DNA (cDNA) using the PrimeScript RT Reagent Kit (TaKaRa Biotechnology [Dalian] Co., Ltd.). RT-PCR analysis was performed using SYBR Green Master Mixture reagent (Takara Bio, Inc., Kusatsu, Shiga, Japan) and an ABI 7500-Fast Real-Time PCR system (Applied Biosystems, Carlsbad, CA, USA). The cycling conditions for cDNA amplification are described elsewhere (Zheng et al., 2018). The fold change in relative gene expression was calculated using the 2<sup>-ΔΔCt</sup> method with glyceraldehyde 3-phosphate dehydrogenase (GAPDH) as an internal reference. The primers used for RT-PCR are listed in **Supplementary Table S1**.

## Western Blot Analysis

Total protein isolated from cells was sonicated in ice-cold radio immunoprecipitation assay lysis buffer (Pierce Biotechnology, Waltham, MA, USA). Denatured proteins were separated by sodium dodecyl sulfate polyacrylamide gel electrophoresis and then transferred to a polyvinylidene fluoride membrane (EMD Millipore Corporation, Billerica, MA, USA), which was blocked with Tris-buffered saline and 5% skim milk for 2 h. Samples were incubated with primary antibodies against the cyclin E1 (CCNE1) gene (catalog no. 20808; dilution, 1:1000; Cell Signaling Technology, Inc., Danvers, MA, USA) at 4°C

overnight. After rinsing, the membrane was incubated with horseradish peroxidase-conjugated anti-rabbit secondary antibody (#7074; dilution, 1:1000; Cell Signaling Technology, Inc.). The protein bands were visualized using an enhanced chemiluminescence kit (Wanleibio Co., Ltd., Shenyang, China) and the ChemiDoc<sup>™</sup> Touch Imaging System (Bio-Rad Laboratories, Hercules, CA, USA). The degree of gray intensity was determined using ImageJ software (<https://imagej.nih.gov/ij/>) and normalized to that of GAPDH (#2118; dilution, 1:5000; Cell Signaling Technology, Inc.).

## Flow Cytometry Analysis

The cells were fixed with ice-cold 70% ethanol overnight and then resuspended in staining solution included with the cell cycle detection kit (Nanjing KeyGen Biotech. Co. Ltd., Nanjing, China). After incubation for 1 h at 37°C in the dark, the stained cells were subsequently analyzed by flow cytometer fluorescence-activated cell sorting (FACS) using the BD FACSCalibur<sup>™</sup> Cell Analyzer system (BD Biosciences, San Jose, CA, USA).

## Dual-Luciferase Reporter Assay

A fragment of the wild-type (WT) RAET1K 3'-untranslated region (RAET1K-3'UTR-wt) contained a binding site downstream of the luciferase reporter gene, whereas the mutant-type RAET1K (RAET1K-3'UTR-mut) contained mutated binding sites (GenePharma Co., Ltd.). A549 and H1299 cells were transfected in the wells of 24-well plates, cultured until attachment, and co-transfected with miR-135a-5p mimics, miR-135a-5p inhibitors or the miR-NC encoded by the luciferase plasmids (RAET1K-3'UTR-wt or RAET1K-3'UTR-mut). Luciferase gene expression was monitored using the Dual-Luciferase<sup>®</sup> Reporter Assay System (Promega Corporation, Madison, WI, USA), as described previously (Zheng et al., 2018). The results of experiments performed in triplicate were normalized to Renilla luciferase activity values.

## Statistical Analysis

Data are presented as the mean ± standard deviation. All statistical analyses were performed using Prism 8.0 software (GraphPad Software, Inc., La Jolla, CA, USA). Student's *t*-test and one-way analysis of variance were used to analyze two groups and more than two groups, respectively. The Pearson's correlation coefficient was used to identify correlations. Analysis of each sample was performed in triplicate. *P* < 0.05 was considered statistically significant.

# RESULTS

## Significant Genes and Clusters With Functions Related to LUAD

### DIFF-Genes in Early and Advanced Stages of LUAD

The LUAD database included 24,989 genes from 564 tissue samples, which included 59 adjacent noncancerous tissues, 395 early stage LUAD tissues (274 stage I and 121 stage II), and 110 advanced stage LUAD tissues (84 stage III and 26 stage IV). In

total, 1,069 and 425 DIFF-genes were upregulated and downregulated in early stage LUAD (**Figure 1A**), respectively, whereas 888 and 516 were upregulated and downregulated in advanced stage LUAD, respectively (**Figure 1B**). In total, 991 DIFF-genes in both early and advanced stages were used to construct the weighted correlation network.

### Construction of the Gene Co-Expression Network in LUAD

WGCNA was performed for 991 DIFF-genes. First, potential hub genes in each module were investigated to identify correlations with the clinical features of LUAD patients. The generalized TOM defined the relationships of each pair of DIFF-genes from the adjacency matrix. The hierarchical clustering tree method detected that four modules contained DIFF-genes that highly correlated with LUAD, as depicted in turquoise, brown, blue and green color (**Figure 1C**). In the middle of the TOM network, a heatmap of the independent genes in different modules was constructed. The genes clustered into the blue and turquoise modules were significantly co-expressed with each other.

The DIFF-genes in each module were spontaneously clustered according to the following clinical features: early stage, advanced stage, tumor size (T), lymph node involvement (N), and presence of metastasis (M). Module trait relationships were calculated by correlating the ME values with the clinical features (**Figure 1D**). There were no significantly positive modules related to early stage disease or other clinical traits. However, the genes in the blue and brown modules were significantly and positively correlated with advanced stage disease, whereas the genes in the blue module showed strong associations (correlation rate = 0.8, **Figure 1E**) and were chosen for subsequent analyses.

### Functional Enrichment Analysis of Selected Modules

To describe the BPs and mechanisms of hub genes, the GO functional enrichment analysis of 203 DIFF-genes in the blue module were performed using DAVID as a reference. The top 10 BPs were visualized using a bubble plot (**Figure 1F**), which showed that most of the DIFF-genes were involved in the cell cycle (**Supplementary Table S2**). Furthermore, ClueGO was performed to enrich the KEGG pathways of the DIFF-genes in the blue module (**Figure 1G**). In total, 168 protein-coding RNAs in the blue module were grouped into six significant KEGG pathways ( $P \leq 0.05$ ). The red nodes contained 23 genes enriched in the cell cycle pathway (**Supplementary Table S3**).

### Function of RAET1K as a Key Gene in LUAD

#### Detection of Significant Genes in the Blue Module

According to GRCh38.p12, 12 lncRNAs and 191 mRNAs were assigned to the blue module. To further validate the hub genes and identify potential biomarkers for LUAD, Cox proportional hazard and Kaplan-Meier analyses of the genes in the blue module were performed. In total, 141 highly expressed hub genes were significantly associated with poor prognosis. Because there was only one lncRNA out of 141 significant

genes in the blue module, and then we focused on this lncRNA RAET1K for further biological study.

### RAET1K Is Highly Expressed in LUAD and Positively Correlated With the Prognosis of LUAD

RAET1K (HR = 1.428; 95% CI = 1.052–1.939;  $P = 0.022$ , **Figure 2A**) was the only lncRNA among the 141 hub genes that was significantly upregulated in tumor tissue compared with normal tissue (**Figure 2H**). Furthermore, a nomogram was constructed to predict 1- and 3-year survival rates in patients with LUAD by showing the risk score of clinical stage, age, sex, and RAET1K expression level (**Figure 2B**). The concordance index, which was evaluated using the calibration plot of this nomogram model, further supported the predictive prognostic signature of lncRNA RAET1K in LUAD OS (**Figure 2C**).

### RAET1K May Regulate the Cell Cycle Phase in LUAD

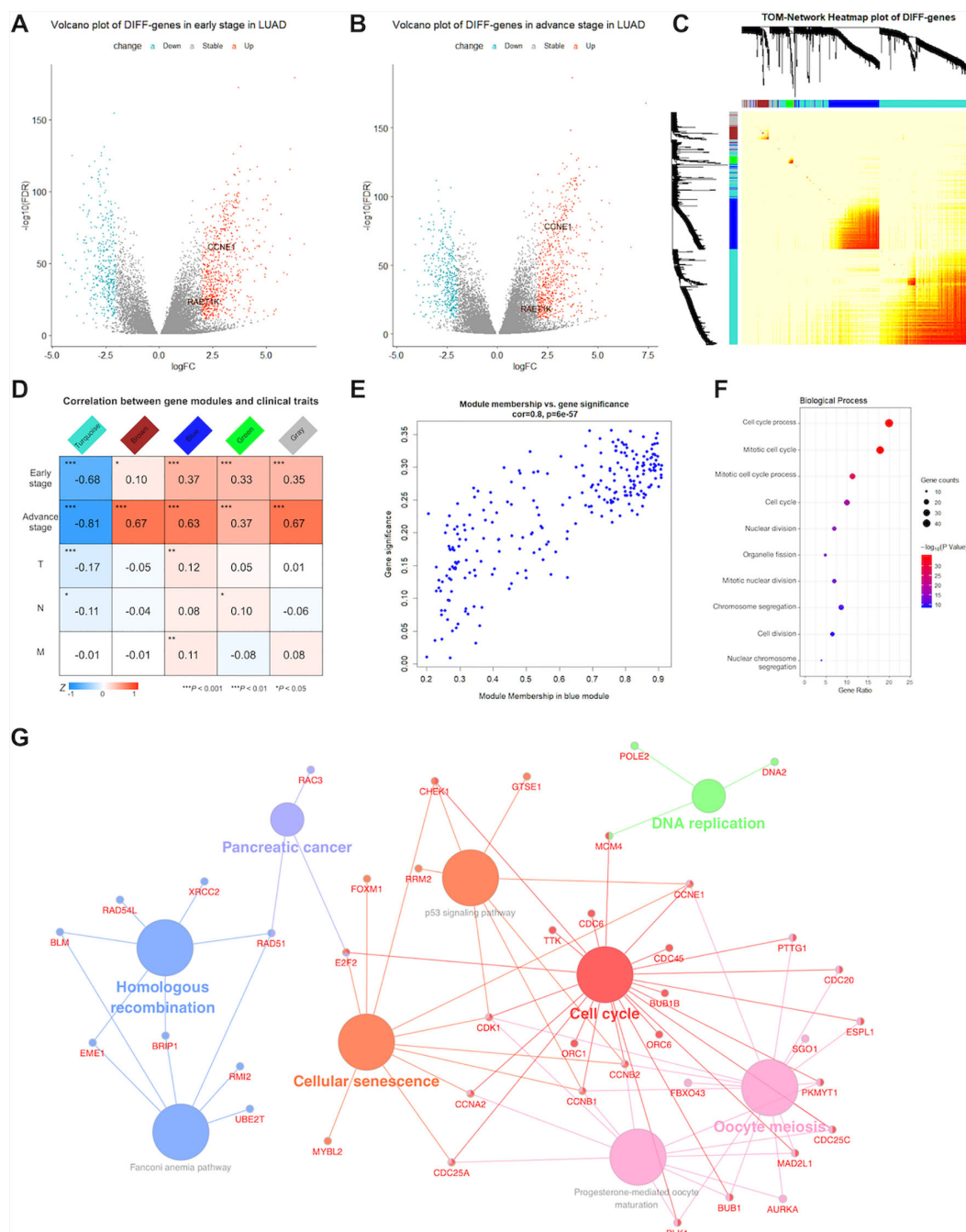
To further explore the biological functions of RAET1K, GO enrichment for GSEA was performed. The LUAD samples with higher expression levels of RAET1K were enriched in genes correlated with cell cycle biological behavior. The GSEA results also indicated that among the genes in the blue module, lncRNA RAET1K expression was enriched in the cell cycle (**Figure 2D**).

### The RAET1K/miR-135a-5p Axis May Influence the Cell Cycle via CCNE1 in LUAD Patients

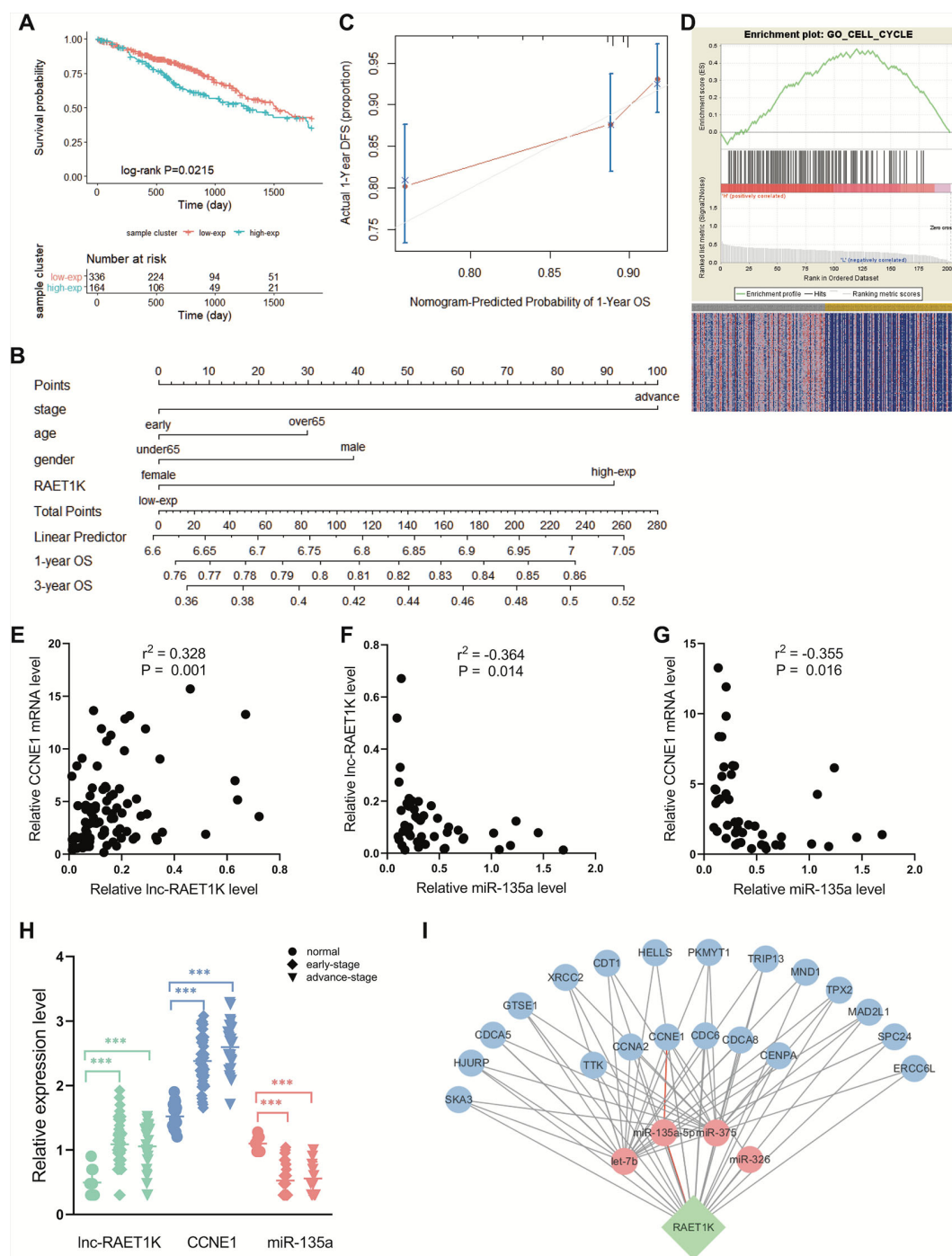
lncRNAs can regulate mRNA expression via miRNA-mediated ceRNAs (Salmena et al., 2011). The expression of ceRNA transcripts that harbor the same miRNA binding sites should be parallel based on the ceRNA hypothesis. The interaction of the ceRNA network and RAET1K is described in **Figure 2I**, which was combined with the expressional correlation and target sites. Among the genes influencing OS, according to the Pearson's correlation coefficient, mRNAs that were positively correlated with RAET1K ( $r > 0.3$  and  $P < 0.05$ , **Figure 2E**) and miRNAs that were negatively correlated with RAET1K and mRNAs ( $r < -0.3$  and  $P < 0.05$ , **Figures 2F, G**) were selected, and then combined with the miRcode database, which was used to predict miRNA-interacting targets. As shown in **Figure 2I**, RAET1K may function as a sponge to absorb miR-135a-5p to modulate CCNE1 expression.

### The RAET1K/miR-135a-5p Axis Arrested LUAD Cells in the G1 Phase by Upregulating CCNE1

**RAET1K Regulated CCNE1 by Sponging miR-135a-5p**  
Subsequently, to investigate the validity and potential biological mechanisms of the effects of the RAET1K/miR-135a-5p axis on CCNE1 expression, *in vitro* experiments with A549 and H1299 cells were performed. The efficiency of RAET1K overexpression lentivirus interference was confirmed by RT-PCR (**Figure 3A**). To further investigate the synergistic effect of the RAET1K/miR-135a-5p axis on CCNE1 expression, A549 and H1299 cells were transfected with lentiviral vectors stably overexpressing RAET1K and an empty control (hereafter referred to as A549<sup>RAET1K</sup>, A549<sup>Con</sup>, H1299<sup>RAET1K</sup>, and H1299<sup>Con</sup> cells, respectively).

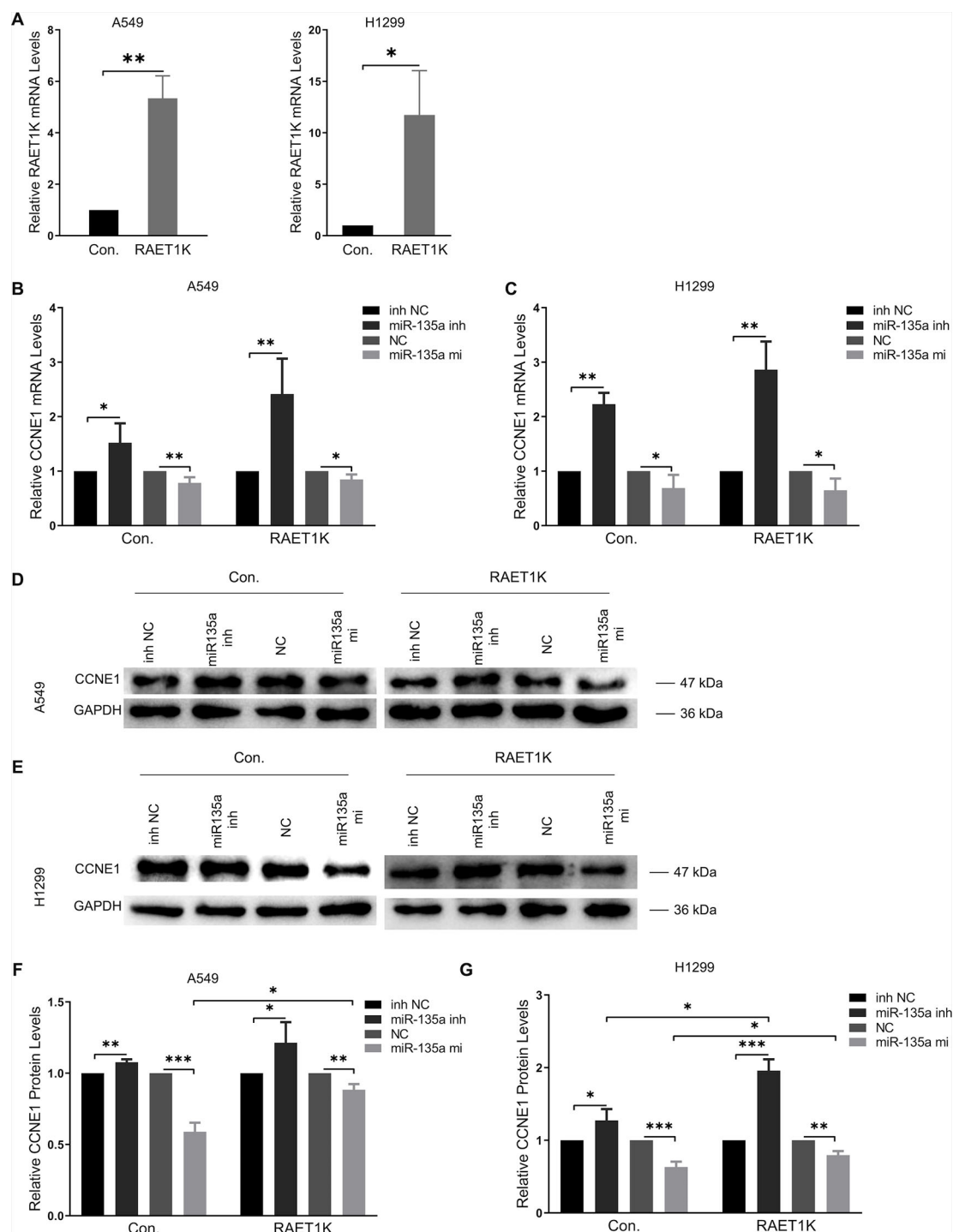


**FIGURE 1 |** Detection of significant genes and their function related to lung adenocarcinoma (LUAD). Volcano plots showed fold change (FC) and  $P$ -values of differentially expressed genes in early (A) and advance (B) stage LUAD versus normal samples. Blue nodes present significantly down-regulated, and red nodes are up-regulated expressed genes. Grey nodes are not differentially expressed. RAET1K and CCNE1 expression are annotated. (C) In middle topological overlap matrix (TOM) heatmap, every row and column present one gene, light color presents low, while darker red presents higher weighted correlation. The dynamic tree cluster dendrogram of DIFF-genes are showed in the left and top, gray square indicates genes that are involved in any known module. (D) LUAD module-clinical feature relationships. The row matches a clinical trait (early stage, advance stage, T for tumor size, N for lymph node and M for metastasis) and the column matches a genes module. Correlation of module and clinical trait is showed in each cell. The darker the color is, the higher the degree of correlation is. Red presents positive, while blue presents negative correlation. (E) Scatterplot of gene significance and module membership in the blue module. Correlation coefficients and  $P$ -values are at the top. (F) Bubble plots showed top 10 terms of gene ontology (GO) enrichment analysis in biological process for blue module. The Y-axis correspond to the GO terms. The gene counts and  $-\log$  (enrichment  $P$ -value) in every GO term were proportional to the area and color of the bubble, respectively. (G) Genes Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis in blue module. The small size nodes in the network represent the genes enriched in the specific pathway, the big size nodes represent pathway term. The node colors correspond to the ClueGO-determined KEGG pathway clusters.



**FIGURE 2 |** Identification of lncRNA RAET1K function and biological mechanism. **(A)** The Kaplan-Meier curve of the risk score for the overall survival of RAET1K in lung adenocarcinoma (LUAD). The blue line presents the lower expression level group of RAET1K, and the red presents the higher ones. Gene enrichment plots showed gene set enrichment analysis (GSEA) between high- and low-expressed RAET1K. **(B)** The nomogram of clinical features and RAET1K expression level for predicting the 1- and 3-year survival with risk score. **(C)** calibration plot indicated this nomogram model had a predictive power for overall survival. **(D)** The upper enrichment plots contain value of the genes' enrichment scores and the corresponding barcode plot shows the genes position. In the bottom heatmap red represents Spearman correlations with higher expression level of RAET1K, blue represents Spearman correlations with lower expression level of RAET1K. Expression of RAET1K and CCNE1 expression level were positive related with each other **(E)**, while RAET1K **(F)** and CCNE1 **(G)** were negatively correlated with miR-135a. **(H)** RAET1K and CCNE1 expression were upregulated in both early and advance stage of LUAD, while miR-135a was downregulated,  $***P < 0.001$ . **(I)** Construction of ceRNA network of lncRNA-miRNA-mRNA in blue module. The green node in diamond was lncRNA RAET1K, the blue circle nodes were mRNAs, and the pink circle nodes were miRNA. The line between nodes present their relation and the red lines shown RAET1K targeted miR-135a-5p and CCNE1.





**FIGURE 3 |** Overexpression RAET1K upregulated CCNE1 by sponging miR-135a-5p. **(A)** The interference efficiency of RAET1K overexpression lentivirus was detected by real-time PCR in A549 and H1299. Relative CCNE1 mRNA expression level after co-transfected with miR-135a-5p (or inhibitor) and RAET1K in A549 **(B)** and H1299 **(C)** cell lines, while the cyclin E1 protein levels was measured by Western blot in A549 **(D and F)** and H1299 **(E and G)**. Bands were quantitatively compared with relative negative control groups. Data are represented as means  $\pm$  S.D. from three independent experiments, \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ . Con., control; inh NC, miRNA-135a-5p inhibitor negative control; inh, inhibitor; NC, negative control; mi, mimics.

Thereafter, A549<sup>RAET1K</sup>, A549<sup>Con</sup>, H1299<sup>RAET1K</sup>, and H1299<sup>Con</sup> cells were transfected with miR-135a-5p mimics, an inhibitor, an NC, or an NC inhibitor.

RT-PCR analyses of A549<sup>Con</sup> and H1299<sup>Con</sup> cells showed that miR-135a-5p inhibition resulted in a 1.5- and 2.2-fold increase, respectively, in CCNE1 mRNA expression relative to the NCs (**Figures 3B, C**, left panel). We observed that overexpression of RAET1K increased miR-135a-5p inhibition, as compared with NC (2.4- and 2.9-fold increases in A549<sup>RAET1K</sup> and H1299<sup>RAET1K</sup>, respectively, **Figures 3B, C**, right panel).

Western blot analysis showed that cyclin E1 protein levels were similar (**Figures 3D–G**). We observed that A549<sup>Con</sup> and A549<sup>RAET1K</sup> cells transfected with miR-135a-5p mimics reduced cyclin E1 protein expression levels, whereas miR-135a-5p inhibitors had an opposite effect (**Figure 3D**). Consistently, cyclin E1 protein expression showed similar tendencies with higher fold changes in H1299<sup>Con</sup> and H1299<sup>RAET1K</sup> cells co-transfected with miR-135a-5p inhibitor compared with those with NC inhibitor (**Figure 3E**). Additionally, although miR-135a-5p mimics significantly decreased cyclin E1 protein expression, this change was salvaged by RAET1K overexpression, thereby indicating that the change in cyclin E1 protein expression in response to RAET1K and miR-135a-5p was due to posttranscriptional modulation in both A549 and H1299 cells. Considering these results, lncRNA RAET1K inhibited CCNE1 mRNA expression probably *via* the downregulation of miR-135a-5p expression.

### RAET1K as a Target of miR-135a-5p

The expression levels of miR-135a-5p and RAET1K were inversely correlated in LUAD tissues and cell lines. Bioinformatics analysis predicted that RAET1K was a potential target of miR-135a-5p. **Figure 4A** describes a putative interaction of RAET1K-3'UTR and modified RAET1K-3'UTR-mut with the miR-135a-5p binding sequence. The luciferase reporter assay was performed to validate the interactions between miR-135a-5p and RAET1K in A549 and H1299 cells. Relative luciferase activity was inhibited by co-transfection with the miR-135a-5p mimics and the luciferase reporters containing RAET1K-3'UTR. However, inhibition was relatively weak in the RAET1K-3'UTR-mut group (**Figures 4B, C**). Luciferase activity was enhanced with the use of the miR-135a-5p inhibitor (**Figures 4B, C**).

### The RAET1K/miR-135a-5p Axis Arrested LUAD Cells in the G1 Phase

To determine whether the RAET1K/miR-135a-5p axis exerted synergistic effects on cell cycle progression, cell cycle distributions were investigated following the co-transfection of RAET1K and miR-135a-5p mimics or an inhibitor in A549 and H1299 cells. Although the proportions of A549<sup>Con</sup> cells in the various cell cycle phases were not significantly altered by miR-135a-5p expression levels, a tendency for such alterations was observed (**Figure 4D**). In comparison with the NC group, transfection with the miR-135a-5p inhibitor decreased the

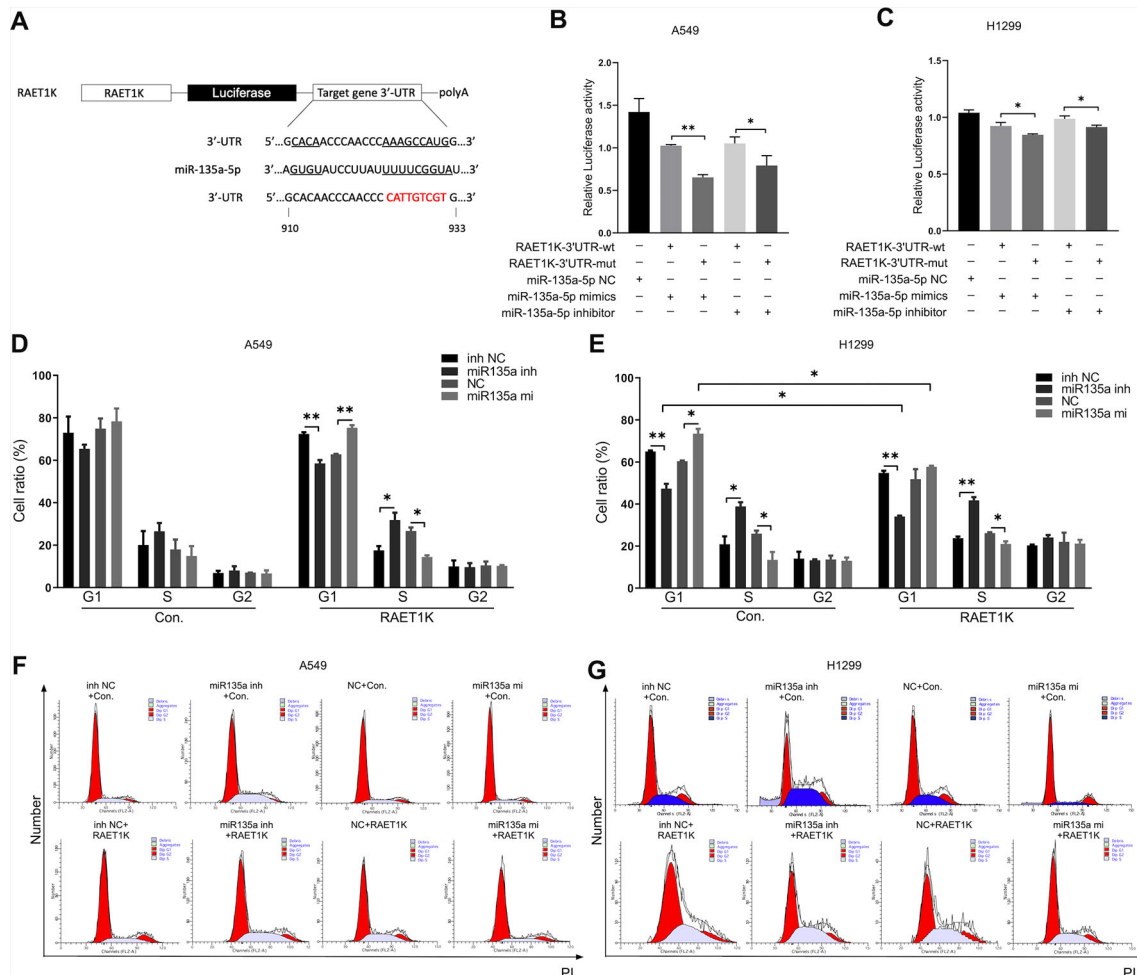
number of A549<sup>RAET1K</sup> cells in the G1 phase, whereas a larger proportion were observed in the S phase (**Figure 4D**).

Similar, yet significant, tendencies were observed in H1299 cells. As compared with the NC inhibitor group, the use of the miR-135a-5p inhibitor resulted in fewer H1299<sup>Con</sup> and H1299<sup>RAET1K</sup> cells arrested in the G1 phase than in the S phase (**Figure 4E**). In addition, lncRNA RAET1K overexpression enhanced the inhibition of cells arrested in the G1 phase. As compared with the NC group, transfection of H1299<sup>Con</sup> cells with the miR-135a-5p mimics increased the number of cells accumulated in the G1 phase; however, RAET1K overexpression rescued this accumulation (**Figure 4E**). Moreover, histograms of the cell cycle were created (**Figures 4F, G**). The results showed that RAET1K overexpression with decreased miR-135a-5p could synergistically arrest the A549 and H1299 cells in the G1 phase and hinder cell cycle transformation from the G1 to S phase.

## DISCUSSION

To identify significant lncRNAs in LUAD, comprehensive computational analysis of transgenic cells was performed. The results showed that lncRNA RAET1K regulated the expression of CCNE1 in LUAD and served as ceRNA to sponge miR-135a, whereas CCNE1 was targeted in cells arrested at the G1-S phase boundary. It is important to understand the pathological cell cycle process that is associated with the dysregulation of cell proliferation leading to cancer (Bertoli et al., 2013). The dynamic progression of the cell cycle consists of four sequential phases: S (chromosome replication), M (chromosome segregation), and G1 and G2 (gap), which are regulated by cyclin/cyclin-dependent kinases (Dai et al., 2018). In particular, cyclin E/Cdk2 interacts and forms complexes that promote G1 progression and G1/S transition (Sonntag et al., 2018). The amplification of cyclin E, which functions in cell cycle progression, inhibition of apoptosis, transcription, and replication, and DNA repair, has been observed in various types of cancer (Kanska et al., 2016; Vijayaraghavan et al., 2017). Furthermore, cyclin E1 can be modulated by multiple regulators, such as the transcription factors c-Myc, retinoblastoma, and E2F (Thurlings and de Bruin, 2016), as well as by miRNA-mediated inhibitors miR-15/16 (Yuan et al., 2019) and miR-424-5p (Jiang et al., 2019) at the transcriptional, posttranscriptional, and translational levels.

The rapid evolution of genomic technologies and analytical tools has improved the understanding of traditional simple gene mutations in cancer genomics. Furthermore elucidation of the complex networks of genomic alterations in LC has provided a basic understanding of the biological consequences and alterations of signal transduction pathways (Chin et al., 2011). A range of evidence suggests that diversity and complex molecular functions of lncRNAs may regulate epigenetic processes, particularly by acting as ceRNAs to sponge miRNAs. To identify novel LUAD-specific lncRNAs, differential analysis was performed during the early and advanced stages using



**FIGURE 4 |** Upregulated RAET1K arrested G1 phase by targeted miR-135a-5p in lung adenocarcinoma (LUAD) cells. **(A)** Schematic representation of the putative binding target and modified sequence site of RAET1K for miR-135a-5p. Luciferase activity between RAET1K-3'UTR-wt/mut and miR-135a-5p detected by dual luciferase reporter assays in A549 **(B)** and H1299 **(C)**. The percentage of cell at different cell cycle phases were in the lower histograms in A549 **(D)** and H1299 **(E)**, while flow cytometry assay results showed cell cycle distribution by PI staining were presented in A549 **(F)** and H1299 **(G)**. Bands were quantitatively compared with relative negative control groups. Data are represented as means  $\pm$  S.D. from three independent experiments. \* $P < 0.05$ , \*\* $P < 0.01$ . wt, wild type; mut, mutant type; inh NC, miRNA-135a-5p inhibitor negative control; inh, inhibitor; NC, negative control; mi, mimics; PI, propidium iodide.

normal tissues in the TCGA LUAD cohort. Different genes in both subsets were selected to facilitate the next step. The co-expression gene network was detected by WGCNA, which is a systematic biological method to identify synergistically altered gene clusters, candidate biomarkers, and therapeutic targets. According to the WGCNA results, DIFF-genes in the blue module were related to the LUAD clinical stage and were enriched in cell cycle-related functions. Cell cycle dysfunction in LUAD was consistent with our results. A recent study demonstrated that cell cycle-related genes, such as E2F1 (Chen et al., 2019), were enriched during the regulation of the cell cycle progression (Li et al., 2018; Qi et al., 2019). In the present study, we found that lncRNA RAET1K could promote cell cycle dysfunction, providing insight into the crosstalk regulatory mechanism between lncRNAs and coding genes. Interestingly,

GSEA results also showed that some cell cyclin proteins and CDK family members were classified by the median of RAET1K expression level including PBK, KIF14, NEK2, CCNE1, CDC45, and DENPF, among others. In addition to the survival prediction of RAET1K, a Kaplan-Meier curve and a nomogram of integrating clinical traits were constructed. Indeed, RAET1K attracted our attention. Liang et al. (Sui et al., 2019) reported that RAET1K was predictive of the prognosis of LUAD patients in a TCGA cohort, which is consistent with our results; however, this was not further verified at the molecular level. To the best of our knowledge, no study has investigated the underlying molecular mechanism of RAET1K in patients with LUAD.

lncRNA RAET1K is a conversely processed transcript at 6q25.1 composed of four exons and is 1,883 bp in length. The key mechanism of lncRNA RAET1K as a ceRNA is to

competitively combine the same miRNA with cross-regulated genes by sharing the miRNA response elements in the 3'-UTR of the target genes. We hypothesized that RAET1K functions as a ceRNA that influences CCNE1 expression and the cell cycle process *via* miR-135a-5p. The role of RAET1K in A549, H1299, and PC-9 cells was investigated to determine why PC-9 cells did not survive puromycin-selection of cells transfected with a lentivirus overexpressing RAET1K. As a possible explanation, the epidermal growth factor receptor gene might be mutated in PC-9 cells, whereas A549 and H1299 cell lines carried the WT phenotype. Therefore, the effects of miR-135a-5p and co-transfection of RAET1K/miR-135a-5p in A549 and H1299 cells were investigated. The results of the PC-9 cells transfected with miR-135a-5p are provided in the **Supplementary Figure S1**. In the A549 and H1299 cell lines, CCNE1 expression was silenced by increased miR-135a-5p, which also affected the cell cycle process. In contrast, the miR-135a-5p inhibitor had opposite effects. The results revealed that overexpression of RAET1K partially absorbed miR-135a-5p and enhanced the miR-135a-5p-mediated biological effects. The tumor suppressive function of miR-135a in LUAD has been consistently demonstrated in previous studies. For instance, miR-135a-5p promoted the progression of head and neck squamous cell carcinoma by targeting HOXA10 (Guo et al., 2018), the progression of thyroid carcinoma by VCAN (Zhao et al., 2017), and the progression of gastric cancer by KIFC1 (Zhang et al., 2016). Conversely, miR-135a was found to target SIAH1 to promote cell transformation in cervical cancer *via* the  $\beta$ -catenin pathway (Leung et al., 2014). Furthermore, Zhang et al. (2019) reported that miR-135a-5p promoted LC progression *via* modulating LOXL4 and blockage of LC cells arrested at the G1 phase. The reasons for these findings could be the differences in the samples used for *in vivo* (LC tissue) vs. *in vitro* (LC cell lines) studies. However, the results above were in agreement regarding the influence of the G1 phase of the cell cycle.

Furthermore, the results of this study indicated that co-transfection of A549<sup>RAET1K</sup> and H1299<sup>RAET1K</sup> cells with the miR-135a-5p inhibitor could act synergistically to reduce the expression level of CCNE1 and accumulate the proportion of cells arrested at the G1-S phase boundary, thereby suggesting the possible existence of an oncogenic RAET1K/miR-135a-5p axis. As predicted and verified by the bioinformatics algorithms and luciferase reporter assay, RAET1K and CCNE1 are potential targets of miR-135a-5p at the 7-mer-m8 site. The lncRNA RAET1K/miR-135a-5p axis might have a stronger synergistic effect on the regulation of cell cycle phase-dependent CCNE1 and transformation from the G1 to S phase. Here, the role of

RAET1K as a putative oncogene in LUAD was revealed, suggesting that targeting the cyclin E1-CDK signaling provides a novel targeted therapeutic option for the treatment of LUAD. However, further investigations are required to verify the crucial molecules and signaling pathways involved in lncRNA RAET1K-mediated LUAD tumorigenesis.

## CONCLUSION

The major finding of this study was that RAET1K acted as a ceRNA and increased the expression of CCNE1 by directly competing with miRNA-135a-5p, which influenced the function of the cyclin E1 protein. Furthermore, the RAET1K/miR-135a-5p axis, which drives cell cycle progression, was arrested at the G1 phase in LUAD onset and progression. These findings are expected to be useful for the development of a novel biomarkers and pathways regulating the the cell cycle in LUAD.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the Cancer Genome Atlas at (<https://portal.gdc.cancer.gov>).

## AUTHOR CONTRIBUTIONS

Conceptualization, analysis and validation: CZ and XL. Software: YR and ZY. Writing: CZ. Funding acquisition: BZ and XL.

## FUNDING

This project was supported by the National Natural Science Foundation of China (No.81773524 and No.81502878).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2019.01348/full#supplementary-material>

## REFERENCES

- Bertoli, C., Skotheim, J. M., and de Bruin, R. A. (2013). Control of cell cycle transcription during G1 and S phases. *Nat. Rev. Mol. Cell Biol.* 14 (8), 518–528. doi: 10.1038/nrm3629
- Besse, B., and Le Chevalier, T. (2012). Developments in the treatment of early NSCLC: when to use chemotherapy. *Ann. Oncol.* 23 Suppl 10, x52–x59. doi: 10.1093/annonc/mds347

- Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25 (8), 1091–1093. doi: 10.1093/bioinformatics/btp101
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries *CA Cancer J. Clin.* 68 (6), 394–424. doi: 10.3322/caac.21492



- Chen, R., Xia, W., Wang, S., Xu, Y., Ma, Z., Xu, W., et al. (2019). long noncoding RNA SBF2-AS1 is critical for tumorigenesis of early-stage lung adenocarcinoma. *Mol. Ther. Nucleic Acids* 16, 543–553. doi: 10.1016/j.omtn.2019.04.004
- Chin, L., Andersen, J. N., and Futreal, P. A. (2011). Cancer genomics: from discovery science to personalized medicine. *Nat. Med.* 17, 297. doi: 10.1038/nm.2323
- Dai, L., Zhao, T., Bisteau, X., Sun, W., Prabhu, N., Lim, Y. T., et al. (2018). Modulation of protein-interaction states through the cell cycle. *Cell* 173 (6), 1481–1494.e1413. doi: 10.1016/j.cell.2018.03.065
- Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of transcription in human cells. *Nature* 489 (7414), 101–108. doi: 10.1038/nature11233
- Dong, P., Xiong, Y., Yue, J., Hanley, S. J. B., Kobayashi, N., Todo, Y., et al. (2018). Long non-coding RNA NEAT1: a novel target for diagnosis and therapy in human tumors. *Front. Genet.* 9, 471. doi: 10.3389/fgene.2018.00471
- Dong, P., Xiong, Y., Yue, J., Xu, D., Ihira, K., Konno, Y., et al. (2019). Long noncoding RNA NEAT1 drives aggressive endometrial cancer progression via miR-361-regulated networks involving STAT3 and tumor microenvironment-related genes. *J. Exp. Clin. Cancer Res.* 38 (1), 295. doi: 10.1186/s13046-019-1306-9
- Ettinger, D. S., Wood, D. E., Akerley, W., Bazhenova, L. A., Borghaei, H., Camidge, D. R., et al. (2015). Non-small cell lung cancer, version 6.2015. *J. Natl. Compr. Canc. Netw.* 13 (5), 515–524. doi: 10.6004/jnccn.2015.0071
- Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., and Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer* 127 (12), 2893–2917. doi: 10.1002/ijc.25516
- Guo, L. M., Ding, G. F., Xu, W., Ge, H., Jiang, Y., Chen, X. J., et al. (2018). miR-135a-5p represses proliferation of HNSCC by targeting HOXA10. *Cancer Biol. Ther.* 19 (11), 1–28. doi: 10.1080/15384047.2018.1450112
- Ingham, M., and Schwartz, G. K. (2017). Cell-cycle therapeutics come of age. *J. Clin. Oncol.* 35 (25), 2949–2959. doi: 10.1200/jco.2016.69.0032
- Jeggari, A., Marks, D. S., and Larsson, E. (2012). miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics* 28 (15), 2062–2063. doi: 10.1093/bioinformatics/bts344
- Jiang, B., Wu, D., Huang, L., and Fang, H. (2019). miR-424-5p inhibited malignant behavior of colorectal cancer cells by targeting CCNE1 *Panminerva Med.* doi: 10.23736/s0031-0808.19.03708-x
- Kanska, J., Zakhour, M., Taylor-Harding, B., Karlan, B. Y., and Wiedemeyer, W. R. (2016). Cyclin E as a potential therapeutic target in high grade serous ovarian cancer. *Gynecol. Oncol.* 143 (1), 152–158. doi: 10.1016/j.ygyno.2016.07.111
- Langfelder, P., and Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.* 1, 54. doi: 10.1186/1752-0509-1-54
- Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf.* 9, 559. doi: 10.1186/1471-2105-9-559
- Leung, C. O., Deng, W., Ye, T. M., Ngan, H. Y., Tsao, S. W., Cheung, A. N., et al. (2014). miR-135a leads to cervical cancer cell transformation through regulation of beta-catenin via a SIAH1-dependent ubiquitin proteasomal pathway. *Carcinogenesis* 35 (9), 1931–1940. doi: 10.1093/carcin/bgu032
- Li, X., Li, B., Ran, P., and Wang, L. (2018). Identification of ceRNA network based on a RNA-seq shows prognostic lncRNA biomarkers in human lung adenocarcinoma. *Oncol. Lett.* 16 (5), 5697–5708. doi: 10.3892/ol.2018.9336
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi: 10.1186/s13059-014-0550-8
- Otto, T., and Sicinski, P. (2017). Cell cycle proteins as promising targets in cancer therapy. *Nat. Rev. Cancer* 17 (2), 93–115. doi: 10.1038/nrc.2016.138
- Qi, G., Kong, W., Mou, X., and Wang, S. (2019). A new method for excavating feature lncRNA in lung adenocarcinoma based on pathway crosstalk analysis. *J. Cell Biochem.* 120 (6), 9034–9046. doi: 10.1002/jcb.28177
- Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146 (3), 353–358. doi: 10.1016/j.cell.2011.07.014
- Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer Statistics, 2017. *CA Cancer J. Clin.* 67 (1), 7–30. doi: 10.3322/caac.21387
- Sonntag, R., Giebler, N., Nevzorova, Y. A., Bangen, J. M., Fahrenkamp, D., Lambertz, D., et al. (2018). Cyclin E1 and cyclin-dependent kinase 2 are critical for initiation, but not for progression of hepatocellular carcinoma. *Proc. Natl. Acad. Sci. U.S.A.* 115 (37), 9282–9287. doi: 10.1073/pnas.1807155115
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102 (43), 15545–15550. doi: 10.1073/pnas.0506580102
- Sui, J., Yang, S., Liu, T., Wu, W., Xu, S., Yin, L., et al. (2019). Molecular characterization of lung adenocarcinoma: a potential four-long noncoding RNA prognostic signature. *J. Cell Biochem.* 120 (1), 705–714. doi: 10.1002/jcb.27428
- Thurlings, I., and de Bruin, A. (2016). E2F transcription factors control the roller coaster ride of cell cycle gene expression. *Methods Mol. Biol.* 1342, 71–88. doi: 10.1007/978-1-4939-2957-3\_4
- Vijayaraghavan, S., Karakas, C., Doostan, I., Chen, X., Bui, T., Yi, M., et al. (2017). CDK4/6 and autophagy inhibitors synergistically induce senescence in Rb positive cytoplasmic cyclin E negative cancers. *Nat. Commun.* 8, 15916. doi: 10.1038/ncomms15916
- Yuan, Z., Zhong, L., Liu, D., Yao, J., Liu, J., Zhong, P., et al. (2019). MiR-15b regulates cell differentiation and survival by targeting CCNE1 in APL cell lines. *Cell Signal* 60, 57–64. doi: 10.1016/j.cellsig.2019.04.005
- Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article17. doi: 10.2202/1544-6115.1128
- Zhang, C., Chen, X., Chen, X., Wang, X., Ji, A., Jiang, L., et al. (2016). miR-135a acts as a tumor suppressor in gastric cancer in part by targeting KIFC1. *Oncotargets Ther.* 9, 3555–3563. doi: 10.2147/ott.s105736
- Zhang, Y., Jiang, W. L., Yang, J. Y., Huang, J., Kang, G., Hu, H. B., et al. (2019). Downregulation of lysyl oxidase-like 4 LOXL4 by miR-135a-5p promotes lung cancer progression *in vitro* and *in vivo*. *J. Cell Physiol.* 234 (10), 18679–18687. doi: 10.1002/jcp.28508
- Zhao, X., Sun, Z., Li, H., Jiang, F., Zhou, J., and Zhang, L. (2017). MiR-135a-5p modulates biological functions of thyroid carcinoma cells via targeting VCAN 3'-UTR. *Cancer Biomark* 20 (2), 207–216. doi: 10.3233/cbm-170566
- Zheng, C., Li, X., Qian, B., Feng, N., Gao, S., Zhao, Y., et al. (2018). The lncRNA myocardial infarction associated transcript-centric competing endogenous RNA network in non-small-cell lung cancer. *Cancer Manag. Res.* 10, 1155–1162. doi: 10.2147/cmar.s163395

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Zheng, Li, Ren, Yin and Zhou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Analysis of Key Genes Involved in Potato Anthocyanin Biosynthesis Based on Genomics and Transcriptomics Data

Nie Tengkun<sup>1\*</sup>, Wang Dongdong<sup>1</sup>, Ma Xiaohui<sup>1</sup>, Chen Yue<sup>1\*</sup> and Chen Qin<sup>2\*</sup>

<sup>1</sup> State Key Laboratory of Crop Stress Biology for Arid Areas, College of Agronomy, Northwest A&F University, Yangling, China, <sup>2</sup> State Key Laboratory of Crop Stress Biology for Arid Areas, College of Food Science and Engineering, Northwest A&F University, Yangling, China

## OPEN ACCESS

### Edited by:

Monica Bianchini,  
University of Siena, Italy

### Reviewed by:

Dinesh Kumar,  
Indian Council of Agricultural  
Research (ICAR), India  
Izabela Makalowska,  
Adam Mickiewicz University  
in Poznań, Poland

### \*Correspondence:

Nie Tengkun  
chinantk@126.com  
Chen Yue  
xnchenyue@nwfafu.edu.cn  
Chen Qin  
chenpeter2289@nwsuaf.edu.cn

### Specialty section:

This article was submitted to  
Bioinformatics and Computational  
Biology,  
a section of the journal  
Frontiers in Plant Science

**Received:** 28 November 2018

**Accepted:** 24 April 2019

**Published:** 14 May 2019

### Citation:

Tengkun N, Dongdong W,  
Xiaohui M, Yue C and Qin C (2019)  
Analysis of Key Genes Involved  
in Potato Anthocyanin Biosynthesis  
Based on Genomics  
and Transcriptomics Data.  
Front. Plant Sci. 10:603.  
doi: 10.3389/fpls.2019.00603

The accumulation of secondary metabolites, such as anthocyanins, in cells plays an important role in colored plants. The synthesis and accumulation of anthocyanins are regulated by multiple genes, of which the R2R3-MYB transcription factor gene family plays an important role. Based on the genomic data in the Potato Genome Sequencing Consortium database (PGSC) and the transcriptome data in the SRA, this study used potato as a model plant to comprehensively analyze the plant anthocyanin accumulation process. The results indicated that the most critical step in the synthesis of potato anthocyanins was the formation of *p*-coumaroyl-CoA to enter the flavonoid biosynthetic pathway. The up-regulated expression of the *CHS* gene and the down-regulated expression of *HCT* significantly promoted this process. At the same time, the anthocyanins in the potato were gradually synthesized during the process from leaf transport to tubers. New transcripts of *stAN1* and *PAL* were cloned and named *stAN1-like* and *PAL-like*, respectively, but the functions of these two new transcripts still need further study. In addition, the sequence characteristics of amino acids in the R2-MYB and R3-MYB domains of potato were preliminarily identified. The aims of this study are to identify the crucial major genes that affect anthocyanin biosynthesis through multi-omics joint analysis and to transform quantitative traits into quality traits, which provides a basis and reference for the regulation of plant anthocyanin biosynthesis. Simultaneously, this study provides the basis for improving the anthocyanin content in potato tubers and the cultivation of new potato varieties with high anthocyanin content.

**Keywords:** anthocyanin, potato, multi-omics analysis, *stAN1*, *PAL*, R2R3-MYB

## INTRODUCTION

It is well known that some plants are colorful, and there are many reasons why plants display multiple colors. For example, the pH of plant cytoplasmic substrates, the accumulation of secondary metabolites, such as anthocyanins, and environmental factors, such as light, all have an effect on plant color formation (Asen et al., 1972; Dai and Mumper, 2010; Xu X. et al., 2015). The accumulation of anthocyanins and other flavonoids in cells results in plants displaying colors

other than green (Tanaka et al., 2008). Biosynthesis and metabolic pathways of anthocyanins in plants have been studied in depth, and many key genes have been cloned.

Among the many phenylalanine metabolic pathways, the pathway based on the biosynthesis process of phenylpropanoids is an important source of flavonoids in plants (Salvatierra et al., 2010). Phenylalanine is deaminated by phenylalanine ammonia lyase (PAL) to form *trans*-cinnamic acid; *trans*-cinnamic acid produces cinnamoyl-CoA under 4-coumarate-CoA ligase (4CL); then cinnamoyl-CoA is catalyzed by *trans*-cinnamate 4-monooxygenase (C4H) to form *p*-coumaroyl-CoA; finally *p*-coumaroyl-CoA is involved in the biosynthesis of flavonoids (Vogt, 2010). *p*-coumaroyl-CoA, through chalcone synthase (CHS), shikimate *O*-hydroxycinnamoyltransferase (HCT), chalcone isomerase (CHI), flavonoid 3', 5'-hydroxylase (F3'5'H), flavonoid 3'-monooxygenase (F3'H), naringenin 3-dioxygenase (F3H), dihydroflavonol 4-reductase (DFR), anthocyanidin synthase (ANS) and other enzymes, catalyzes the final formation of pelargonidin, cyanidin and delphinidin, involved in anthocyanin biosynthesis (Martens et al., 2010; Tanaka et al., 2010). Anthocyanin mainly accumulates in plant cell vacuoles in the form of glycosides (Pietrini et al., 2002).

The MYB-bHLH-WD40 transcription factor complex (MBW) is a regulator that has been thoroughly studied and has an important regulatory effect on the synthesis of flavonoids such as anthocyanins (Jaakola, 2013). The main transcription factor involved in the regulation of anthocyanin synthesis in the MYB gene family is the R2R3-MYB transcription factor (Stracke et al., 2007). A study of the *Arabidopsis* MBW complex TT2-TT8-TTG1 showed that the target gene of the complex might be mainly determined by a R2R3-MYB transcription factor-encoded protein (Xu W. et al., 2015). The bHLH proteins involved in the MBW complex have some common features and most belong to the IIIF subfamily (Zimmermann et al., 2004). The *Arabidopsis thaliana* TT8 gene belongs to the bHLH gene family, which can regulate the synthesis of flavonoids by feedback regulation (Baudry et al., 2006). Studies have indicated that the WD40 protein does not participate in the recognition of gene promoters or regulate the expression of target genes; its effect is to link the two other protein subunits in the MBW complex (Hichri et al., 2011). In the synthesis of flavonoids, for some specific genes, MYB transcription factors can activate the corresponding gene transcription directly without binding to bHLH transcription factors (Jaakola, 2013). Thus, it is important that the R2R3-MYB transcription factor plays a role in the synthesis of flavonoids.

Anthocyanin is an important component of polyphenolic antioxidant active substances, and such compounds are easily absorbed and utilized by the human digestive system (Fernandes et al., 2014). Anthocyanins have a special chemical structure, which allows them to exert a variety of physiological and biochemical functions in mammals such as humans (Stintzing and Carle, 2004). On the one hand, anthocyanins have the effect of scavenging free radicals in living organisms and improving the antioxidant capacity of organisms themselves (Miguel, 2011); on the other hand, anthocyanins have many

important pharmacological effects, for example, anthocyanins have significant effects in preventing many major human-related diseases, such as cardiovascular and cerebrovascular diseases, diabetes and its complications, cancer, and so on (Scalbert et al., 2005). Because of the above characteristics, anthocyanins are gradually being valued by chemists and pharmacologists. Potato is an important plant food for humans to obtain antioxidant active substances such as ascorbic acid and polyphenols (Lobo et al., 2010). Nutrients such as anthocyanins accumulate in colored potato tubers. In addition, it is considered that the anthocyanin content of potato with red or purple tubers is significantly higher than that of common potato with white or yellow tubers (Brown et al., 2005; Lachman and Hamouz, 2005). Since anthocyanins have favorable biological functions for humans, the key genes controlling the synthesis and accumulation of potato anthocyanins can be studied, and then the accumulation of anthocyanins in potato tubers can be regulated. This study attempted to control the content of anthocyanins in potato tubers, making it easier for humans to take antioxidant active substances such as anthocyanins, thereby preventing a variety of diseases and making humans healthier.

Potato is a good model plant for studying the formation of plant color by studying the process of anthocyanin biosynthesis. Firstly, potato plants reproduce mainly through asexual reproduction, and the genetic composition is stable. Secondly, different potato varieties have different colors, and for a single potato, the whole plant is consistent in color. In addition, mature potato plants have a large biomass, which is convenient for the determination of various secondary metabolites. Numerous key genes regulating anthocyanin synthesis have been cloned, but it is unclear which of these key genes is the most important. At the same time, whether there are other gene regulatory pathways controlling anthocyanin accumulation in plants is also worthy of further study.

In this experiment, we analyzed the R2R3-MYB transcription factor gene family, which plays a major role in the anthocyanin synthesis process, based on the genomic data of existing diploid potato (*Solanum phureja* DM1-3). Then, potato transcriptomics data from the NCBI Sequence Read Archive (SRA) database were used to determine which key genes were enriched in anthocyanin synthesis. Finally, based on the above analysis results, we aimed to identify the most critical genes involved in the regulation of anthocyanin biosynthesis and to explore new genes that may be involved in the regulation of anthocyanin synthesis.

## MATERIALS AND METHODS

### Identification of the R2R3-MYB Subfamily Genes in Potato Proteome Data

We downloaded proteomic data PGSC\_DM\_v3.4\_pep.fasta (Amino acid sequences corresponding to all gene coding

sequences) from the potato group database PGSC<sup>1</sup>. The identification of R2R3-MYB subfamily genes used *stAN2* as a reference sequence (Jung et al., 2009); local Blast analysis was performed using blast-2.6.0+ software, and the e-value was set to e-5. After removal of short sequences of amino acids with a length less than 100 and repeated sequences, the SMART<sup>2</sup> database was submitted for retrieval. MEME 4.11.4<sup>3</sup> was used to determine the conserved domain boundaries of the MYB-R2 and MYB-R3 domains in potato. Only the amino acid sequences having both the MYB-R2 and MYB-R3 domains were retained for subsequent analysis.

## Construction of the Phylogenetic Tree of the Potato R2R3-MYB Gene and Collinear Analysis

Using MEGA7<sup>4</sup> software, an unrooted tree was constructed using the minimal evolution method, and the phylogenetic tree was tested using Bootstrap = 1000. The potato genome collinearity analysis was performed based on the PGSC\_DM\_v3.4\_cds.fasta application MCSanX<sup>5</sup>, and circos-0.69<sup>6</sup> was used to visualize the results of the potato genome collinearity analysis.

## Transcriptional Data of Potato Color Changes Were Analyzed

The potato transcriptome data were downloaded from the SRA database<sup>7</sup> the downloaded data format was transformed by the SRA-Toolkit<sup>8</sup>, and then the downloaded data were regrouped. According to the color of the potato stem and tuber used in sequencing, they were reclassified into a colored group and colorless group. The regrouped colored group contained 21 biological replicates; the regrouped colorless group contained 36 biological replicates. The colorless group was the control group, and the data and grouping information are shown in **Supplementary Table S5** (Hannapel et al., 2013; Liu et al., 2015; Gálvez et al., 2016; Pham et al., 2017). In this experiment, the NGSQC Toolkit (Patel and Jain, 2012) was used to filter the reads; Trimmomatic<sup>9</sup> was used to remove the linkers used for sequencing; and the PCR repeats generated during the sequencing process were eliminated by FastUniq<sup>10</sup>. Using the doubled monoploid *S. tuberosum* Group Phureja clone DM1-3 (DM) as the reference genome (Xu et al., 2011), TopHat and Cufflinks were used to splice the transcriptome data and obtain differentially expressed genes (Trapnell et al., 2012). Finally, InterProScan-5.29-68.0<sup>11</sup> and

KOBAS 3.0<sup>12</sup> were used for preliminary annotations of the differentially expressed genes.

## GO Annotation and KEGG Enrichment Analysis Based on Genomic and Transcriptome Analysis Results

Comprehensive genomic and transcriptome analysis results were analyzed by GO annotation and KEGG enrichment using AnnotationDbi<sup>13</sup>, AnnotationHub<sup>14</sup> and clusterProfiler<sup>15</sup>. Only GO annotations and KEGG enrichment analysis results with *p*-value < 0.05 were retained. The GOplot<sup>16</sup> was applied to visualize the results of GO annotation. The KEGG analysis results were confirmed by the KEGG online database<sup>17</sup>.

## Semi-Quantitative RT-PCR to Detect Gene Expression

Semi-quantitative RT-PCR was used to verify the expression of the key genes obtained from the above studies. We applied the potato variety Shepody and the colored potato material, Yellow Meigui 1, Red Meigui 3, Purple Meigui 2, which were bred in our laboratory. The color performance of each potato material is shown in **Figure 5C**. In this experiment, total RNA of roots, stems, leaves, and tubers of potato seedlings was extracted by TRNzol. After reverse transcription, semi-quantitative RT-PCR was carried out with *EF-1α* as the reference gene. The semi-quantitative RT-PCR experiment of each plant tissue was performed with 5 biological replicates. The primers used in the above experiments are shown in **Supplementary Table S6**. Finally, ImageJ<sup>18</sup> was used to measure the agarose gel gray value and perform statistical analysis.

## Application of Tobacco Leaves for Subcellular Localization

The *stAN1-like*-GFP vector and the *PAL-like*-GFP vector were constructed and transformed into *Agrobacterium* strain LBA4404 by the freeze-thaw method. The transformed *Agrobacterium* was cultured at 28°C with shaking until the OD<sub>600</sub> = 0.6–0.8, and the cells were centrifuged. We used a suspension (MES = 10 mmol/L; MgCl<sub>2</sub> = 10 mmol/L; acetosyringone = 0.3 mmol/L; pH = 5.8) to resuspend the cells. The resuspended cells were allowed to stand at room temperature for 2 h, and the resuspended bacteria were injected into the tobacco leaves using a disposable syringe. Under the condition of maintaining the humidity, green fluorescence was observed by laser scanning confocal microscopy (LSCM) after 48 h of tobacco leaf injection. The injected tobacco leaves were treated with a 0.25 g/ml sucrose solution, and the plasmolysis was observed by LSCM (**Supplementary Figure S1**). The GFP excitation wavelength

<sup>1</sup>[http://solanaceae.plantbiology.msu.edu/pgsc\\_download.shtml](http://solanaceae.plantbiology.msu.edu/pgsc_download.shtml)

<sup>2</sup><http://smart.embl-heidelberg.de/>

<sup>3</sup><http://meme-suite.org/index.html>

<sup>4</sup><http://megasoftware.net/>

<sup>5</sup><http://chibba.pgml.uga.edu/mcscan2/>

<sup>6</sup><http://circos.ca/software/download/circos/>

<sup>7</sup><https://www.ncbi.nlm.nih.gov/sra>

<sup>8</sup><https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>

<sup>9</sup><http://www.usadellab.org/cms/index.php?page=trimmomatic>

<sup>10</sup><https://sourceforge.net/projects/fastuniq/files/>

<sup>11</sup><http://www.ebi.ac.uk/interpro/download.html>

<sup>12</sup><http://kobas.cbi.pku.edu.cn/>

<sup>13</sup><http://www.bioconductor.org/packages/devel/bioc/html/AnnotationDbi.html>

<sup>14</sup><http://www.bioconductor.org/packages/release/bioc/html/AnnotationHub.html>

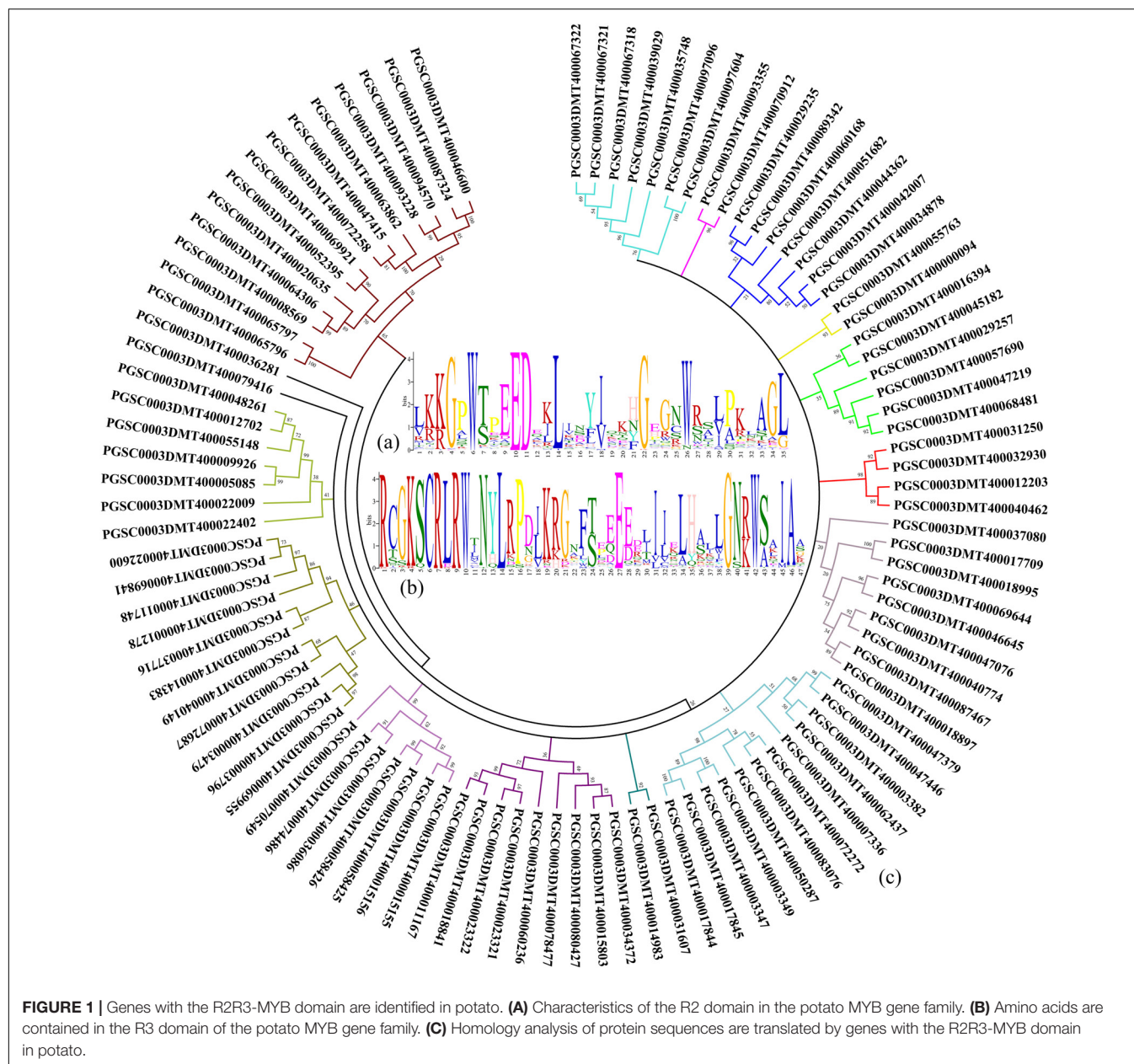
<sup>15</sup><http://www.bioconductor.org/packages/release/bioc/html/clusterProfiler.html>

<sup>16</sup><http://wencke.github.io/>

<sup>17</sup><https://www.genome.jp/kegg/pathway.html>

<sup>18</sup><https://imagej.nih.gov/ij/>





was 488 nm, and the chloroplast autofluorescence excitation wavelength was 633 nm.

## RESULTS

### Identification of Genes Containing Only the R2 and R3 Domains in the Potato MYB Family

In the potato genome data, a total of 101 genes with the R2R3-MYB domain were found by a literature search and sequence alignment (Jung et al., 2009; Zhao et al., 2013; Liu et al., 2016). By comparing the protein sequences found using the

above genes, the common features of the functional structure of the potato R2R3-MYB gene family were obtained. The results of the alignment of the R2 domain, which contains a total of 35 amino acids, are shown in **Figure 1A**. Analysis of the R3 domain revealed a total of 47 amino acids in its domain (**Figure 1B**). In the R2 and R3 domains, the conserved amino acids in order from the N-Terminal to the C-Terminal are glycine (G), tryptophan (W), glutamic acid (E), glycine (G), and tryptophan (W). Therefore, the G-W-E-G-W structure may have an important function in the process of binding the MYB transcription factor to the target promoter.

A phylogenetic tree was constructed using the amino acid sequence corresponding to the gene with the R2R3-MYB domain found in potato. As shown in **Figure 1C**, the population

of genes could be initially divided into 16 subpopulations based on the amino acid homology alignment. Amino acid homology analysis provided a reference for finding genes with the R2R3-MYB domain in the potato genome associated with anthocyanin accumulation.

## Collinearity Analysis of Potato R2R3-MYB Genes

The whole genome of potato was analyzed by collinearity analysis. The results showed that the potato genes were divided into five types: no repeat genes (singleton); modes other than segmental, tandem and proximal (dispersed duplication); nearby chromosomal region but not adjacent (proximal); consecutive repeat (tandem); and collinear genes in collinear blocks (WGD/segmental). Among them, the proximal type had a minimum of 1441 genes; the WGD/segmental type had a maximum of 21372 genes. The remaining types were 4797 genes for the singleton type, 7408 genes for the dispersed duplication type, and 4011 genes for the tandem type gene (Figures 2A–E). There were 25,383 collinear genes and tandem replication genes in the potato genome, accounting for 65.04% of the total number of potato genes. It could be seen that most genes had multiple copies in the potato genome, and there was a high number of genes with similar sequence characteristics or functions.

R2R3-MYB genes were present on each chromosome of potato. The R2R3-MYB genes were most abundantly distributed on the ch05 chromosome, with a total of 14 R2R3-MYB genes on this chromosome. Furthermore, the R2R3-MYB genes were also extensively distributed on the ch01, ch02, ch03, ch06, ch07, and ch10 chromosomes (Figure 2G). The distributions of the collinear genes and the tandem genes in the potato genome were relatively uniform on each chromosome, but there were fewer in the 41–46 Mb region of ch00 and the 1–10 Mb region of ch02. The lines in Figure 2 indicated the collinear relationship between R2R3-MYB genes in the potato genome and between the R2R3-MYB genes and other genes in potato. Based on the above results, a total of 31 other genes were found in the potato genome, which were collinear with the members of the R2R3-MYB gene family identified above (Supplementary Table S1). Genes that were collinear with the R2R3-MYB gene family members could also be used as key candidate genes for the regulation of potato anthocyanin synthesis.

## Transcriptome Analysis Results

Based on the re-grouping transcriptome sequencing data, a total of 12,913 genes with different expression levels were found, of which 420 ( $p \leq 0.05$ ) were significantly different in terms of expression (Supplementary Table S2). There were 11030 genes with different expression levels  $|\log_2\text{FC}| \geq 1$ ; the colored group up-regulated genes accounted for 58.52%, and the colored group down-regulated genes accounted for 41.48% (Figure 2H). Compared with the colorless group, the number of up-regulated genes in the colored group was significantly higher. This indicated that the change in plant color and the

accumulation of anthocyanins were achieved by the simultaneous up-regulation of multiple genes.

## GO Enrichment and KEGG Path Analysis

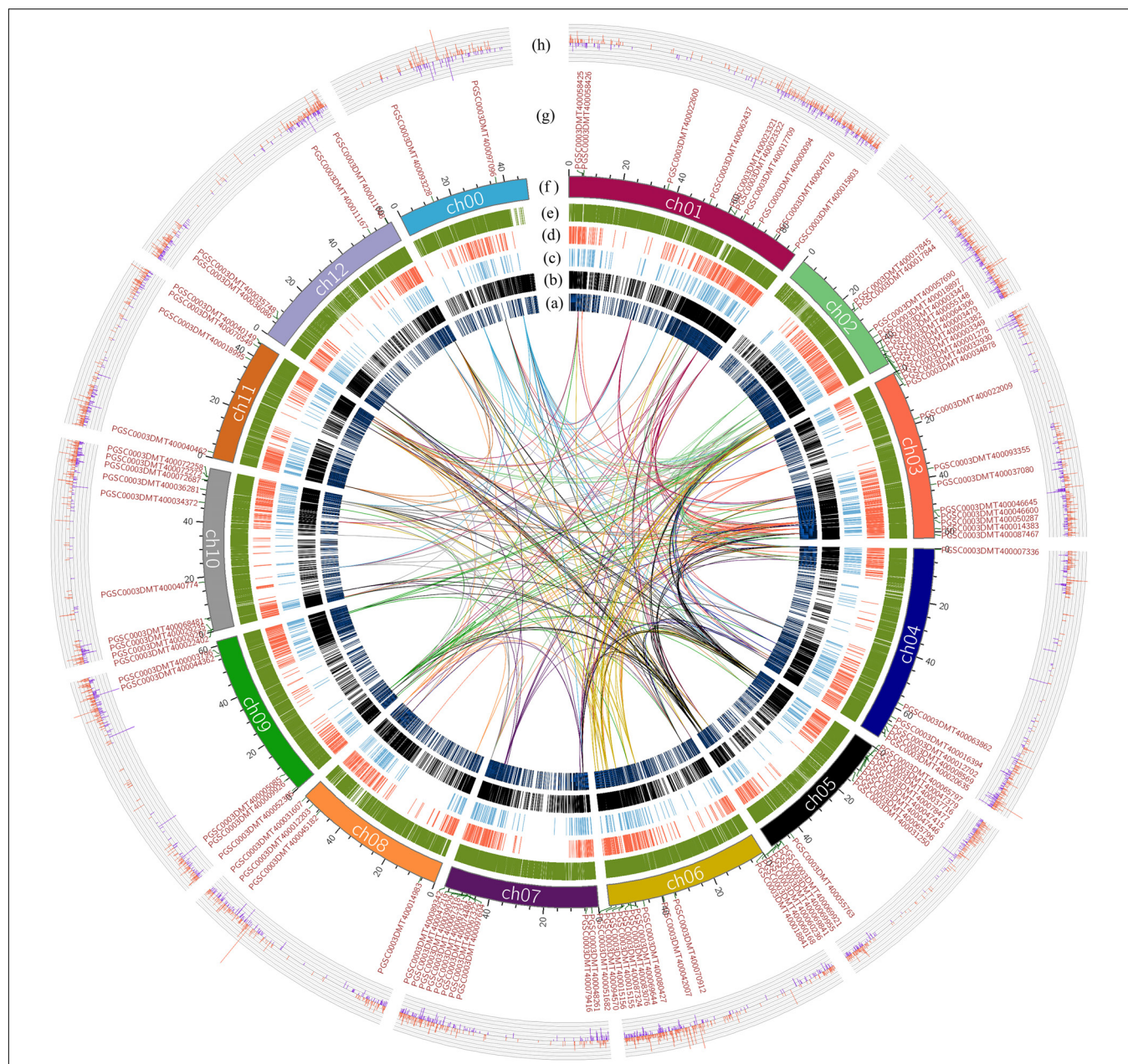
GO enrichment analysis was performed on transcriptome data using interproscan and clusterProfiler software (Yu et al., 2012; Jones et al., 2014). A total of 23 valid GO annotation terms ( $p$ -value  $< 0.05$ ) were enriched, of which there were 7 annotation results with  $p$ -value  $< 0.01$  (Figure 3A). The content of anthocyanins or polyphenols in plants has a close positive correlation with the antioxidant activity of plants (Velioglu et al., 1998). Among the 23 GO analysis results, 10 were significantly associated with plant color changes or plant antioxidant activity. Among them, the GO:0015035, GO:0004601, GO:0016684, GO:0046906, GO:0016747, GO:0010333, and GO:0016829 pathways were enhanced in the colored potato group, whereas the GO:0004866, GO:0030414, and GO:0004857 pathway were weakened in the colored potato group (Figures 3B,D).

It could be seen that the antioxidant activity of potato in the colored group was stronger than that in the colorless group, and the acyltransferase activity of potato in the colored group was also higher than that in the colorless group. This indicated that the high expression of some antioxidant genes and acyltransferase genes contributes to the accumulation of substances such as anthocyanins in plants. At the same time, it also showed that colored potato had higher antioxidant activity, and the antioxidant activity was improved by the simultaneous up-regulation of multiple key genes. A total of 104 differentially expressed genes were enriched in 10 significantly GO pathways, and these gene expressions may play an important role in the accumulation of potato anthocyanins (Figure 3C). Therefore, the above genes can be used as key candidate genes for further study of the synthesis of plant flavonoids and changes in plant antioxidant activity.

The transcriptome data were enriched by KEGG analysis to obtain 23 metabolic pathways ( $p$ -value  $< 0.05$ ), including two pathways closely related to anthocyanin synthesis and accumulation (Figure 4A). These two pathways were sot00940 (phenylpropanoid biosynthesis) and sot00941 (flavonoid biosynthesis). The biological processes related to the accumulation of anthocyanins were sorted, and the up- and down-regulated expression changes of the potato genes in the colored group are shown in Figure 4B. The role of PAL (4.3.1.24) in phenylpropanoid biosynthesis is very important, but this study found that its up-regulated expression in colored potatoes was not obvious. However, the enhancement of the enzyme activity of caffeoyl-CoA *O*-methyltransferase (2.1.1.104), cinnamyl-alcohol dehydrogenase (1.1.1.195), and peroxidase (1.11.1.7) in the colored group promoted the formation of various phenolic substances, represented by lignin, and also promoted the transformation of cinnamoyl-CoA into *p*-coumaroyl-CoA.

*P*-cinnamoyl-CoA is a key precursor of synthetic anthocyanins, and its increased content contributes to the accumulation of potato anthocyanins (Besseau et al., 2007). The up-regulated expression of PGSC0003DMT400022254 and PGSC0003DMT400022255 genes increased the content of the CHS (2.3.1.74) enzyme and promoted the accumulation

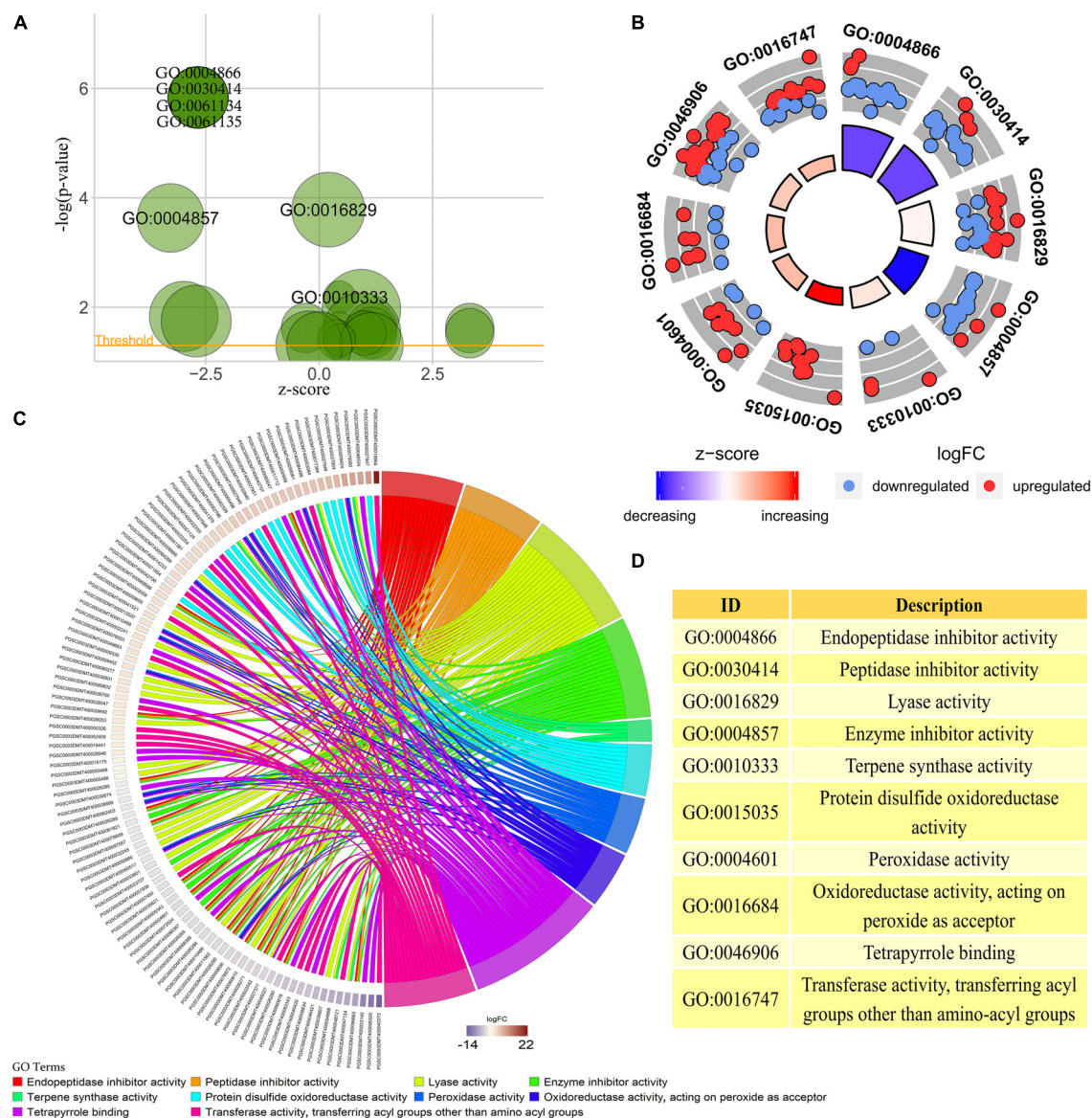




**FIGURE 2 |** Collinearity analysis of the R2R3-MYB domain genes in the potato genome, and gene expression differences based on transcriptome analysis. The lines indicate the collinearity of R2R3-MYB genes in the potato genome, and the line color is the same as the chromosome color corresponding to R2R3-MYB genes. **(A)** The distribution of the singleton-type genes in the potato genome. **(B)** Distribution of dispersed duplication type genes in the potato genome. **(C)** Distribution of the proximal type genes. **(D)** Distribution of the tandem type genes. **(E)** Distribution of WGD/segmental type genes in potato. **(F)** The length of each chromosome of the potato is expressed in units of Mb, where in the ch00 chromosome, are unanchored sequences based on the sequencing result of the potato DM genome. **(G)** The distribution of genes with the R2R3-MYB domain on each chromosome of potato. **(H)** The difference in the expression of each gene obtained by transcriptome analysis showed that the data were differential genes of  $|\log_2FC| \geq 1$ . The red color indicates that the genes were up-regulated in the colored group; the blue color indicates that the genes were down-regulated in the colored group.

of downstream products, which is of great significance in the whole process of anthocyanin accumulation. During the whole process of anthocyanin synthesis, the expression level of the PGSC0003DMT400018861 gene was significantly decreased, resulting in a decrease in the HCT (2.3.1.133) content. This could effectively reduce the loss of *p*-coumaroyl-CoA to caffeic acid

metabolism and promote *p*-coumaroyl-CoA in the flavonoid synthesis pathway, which also had a positive significance for the accumulation of anthocyanins. In addition, the up-regulation of F3'5'H (1.14.14.81) could effectively counteract the effect of HCT (2.3.1.133) down-regulated expression on the anthocyanin composition type. This resulted in the contents of



**FIGURE 3 |** GO enrichment analysis of transcriptome sequencing results. **(A)** All GO enrichment results with  $p$ -value  $< 0.05$ , and the results noted in the figure are GO enrichment results with  $p$ -value  $< 0.01$ . **(B)** GO annotation pathways associated with anthocyanin biosynthesis based on GO enrichment results with  $p$ -value  $< 0.05$ . **(C)** Differentially expressed potato genes based on the results of the GO pathway associated with anthocyanin biosynthesis in the above analysis. The higher the logFC, the higher the expression of genes in the potato colored group, and vice versa. **(D)** Detailed description of key GO enrichment pathways.

the delphinidin, pelargonidin, and cyanidin classes remaining relatively balanced.

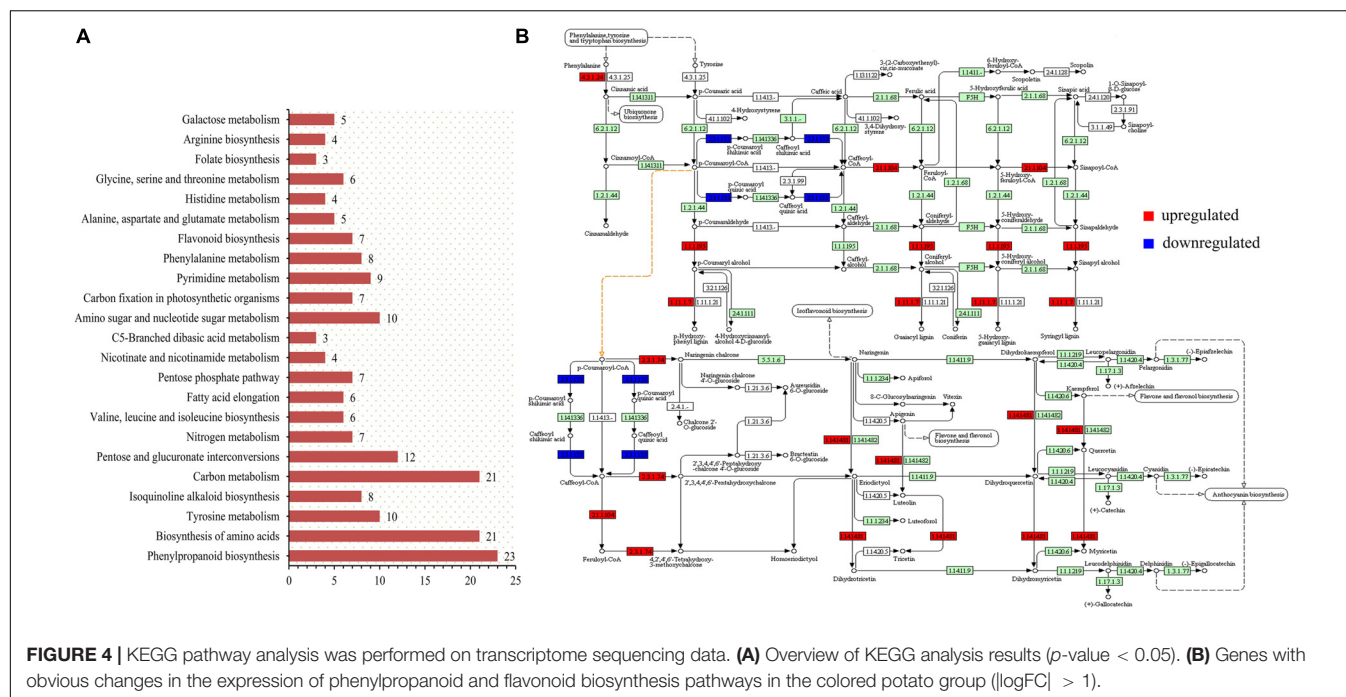
## Semi-Quantitative RT-PCR to Verify the Expression of Related Genes

The members of the potato R2R3-MYB gene family were preliminarily identified by sequence alignment and construction of a phylogenetic tree, and the characteristics of R2 and R3 domains in potato were determined. Based on the collinearity analysis of the R2R3-MYB gene family, the R2R3-MYB gene family members were further enriched. A total of 104 potato

R2R3-MYB gene family members were identified by combining phylogenetic analysis and collinearity analysis. Combined with the results of transcriptome analysis, the differentially expressed genes were searched for among the 104 R2R3-MYB members, and the most differentially expressed genes may be related to the synthesis of anthocyanins and changes in potato color.

Based on a comprehensive comparison of genomic and transcriptome analysis results (**Supplementary Tables S3, S4**), a total of 9 genes were further confirmed. The results of transcriptome analysis were verified by semi-quantitative RT-PCR using colored potatoes as material (**Figure 5C**). The expression of 7 genes was the same as that of transcriptome





**FIGURE 4 |** KEGG pathway analysis was performed on transcriptome sequencing data. **(A)** Overview of KEGG analysis results ( $p$ -value < 0.05). **(B)** Genes with obvious changes in the expression of phenylpropanoid and flavonoid biosynthesis pathways in the colored potato group ( $|\log FC| > 1$ ).

analysis, and the expression of PGSC0003DMT400062326 and PGSC0003DMT400062403 was opposite to that of transcriptome analysis (Figures 5A,B). PGSC0003DMT400040774, PGSC0003DMT400055148, and PGSC0003DMT400009404 were mainly expressed in the potato stem. The expression level of PGSC0003DMT400064555 in various tissues of colored potatoes was generally lower than that of the control Shepody, but higher in the root of Red Meigui 3. PGSC0003DMT400055488 (*PAL-like*) was expressed in leaves and tubers of colored potatoes, but the expression did not increase with the deepening of potato color. The expression levels of PGSC0003DMT400036281 (*stAN1-like*) and PGSC0003DMT400055489 (*PAL*) increased as the color of the potato deepened. *Solanum tuberosum* anthocyanin 1 like (*stAN1-like*) was mainly expressed in the roots, stems and tubers of potato; its expression in Red Meigui 3 and Purple Meigui 2 potato tubers was significantly increased. The expression of phenylalanine ammonia-lyase (*PAL*) was mainly concentrated in the leaves of colored potatoes, but the expression level in the leaves of the control variety Shepody was significantly reduced.

### Subcellular Localization of *stAN1-Like* and *PAL-Like*

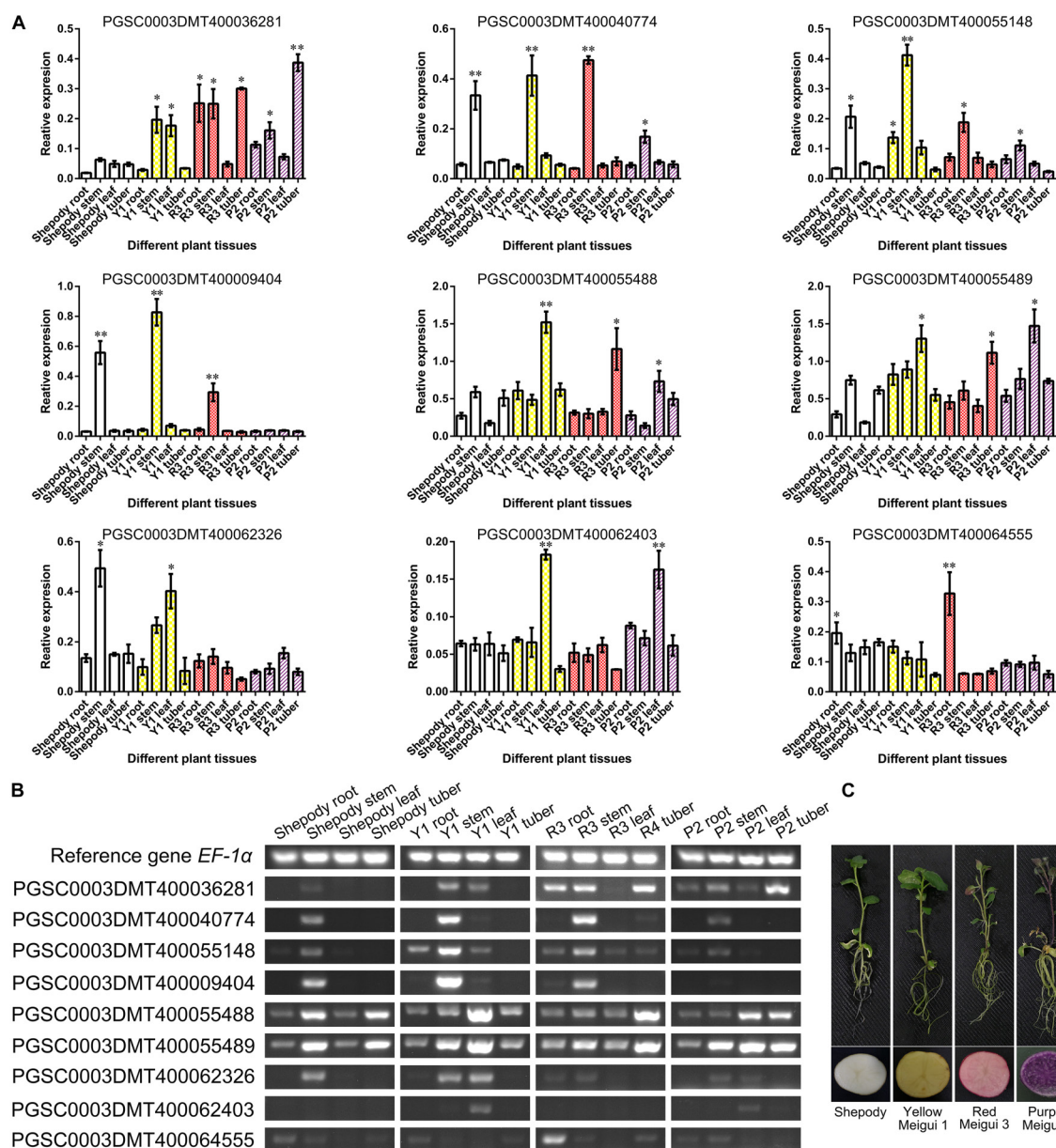
The total RNA of leaves was extracted from the Red Meigui 3 potato, and the new transcripts *stAN1-like* and *PAL-like* of *stAN1* and *PAL* genes were cloned by RT-PCR. The length of the CDS sequence of *stAN1-like* is 798 bp, which indicates that the resulting protein peptide chain contains 265 amino acids. The length of the CDS sequence of *PAL-like* is 2169 bp, and 722 amino acids are included in the protein peptide chain. The subcellular localization results of *stAN1-like* (PGSC0003DMT400036281) and *PAL-like* (PGSC0003DMT400055488) genes are shown in

Figure 6. It could be seen that the proteins produced by the *stAN1-like* gene were mainly concentrated in the nucleus. This suggested that *stAN1-like* might have the function of initiating downstream gene expression. The protein translated by *PAL-like* was concentrated on the cell membrane (Supplementary Figure S1), which is consistent with its function as a functional protein to promote the conversion of phenylalanine to anthocyanin-producing precursor phenylpropanoids. Some of the phenylpropanoids are further metabolized to form lignins involved in cell wall synthesis (Zhou et al., 2009).

## DISCUSSION

### Distribution of Potato R2R3-MYB Transcription Factor on Chromosomes

The R2R3-MYB transcription factor genes have important functions in the process of anthocyanin biosynthesis (Feller et al., 2011). Their primary function in the MBW transcriptional complex is binding to a gene (Xu W. et al., 2015). In this study, 101 R2R3-MYB family genes were found in the potato genome, which were distributed on all of the chromosomes of potato. This indicates that R2R3-MYB transcription factor genes have important biological functions in potato. R2R3-MYB family genes not only participate in the synthesis and regulation of flavonoids, such as anthocyanins, but also participate in many physiological and biochemical processes, such as floral induction, photoperiod response, and plant drought resistance, and so on (Albert et al., 2011; Yang et al., 2012; Zhang et al., 2012; Liu et al., 2013). In addition, in other crops, such as *Arabidopsis* and *Oryza sativa*, the R2R3-MYB transcription factors were also found to be distributed on all of the chromosomes (Katiyar et al., 2012).



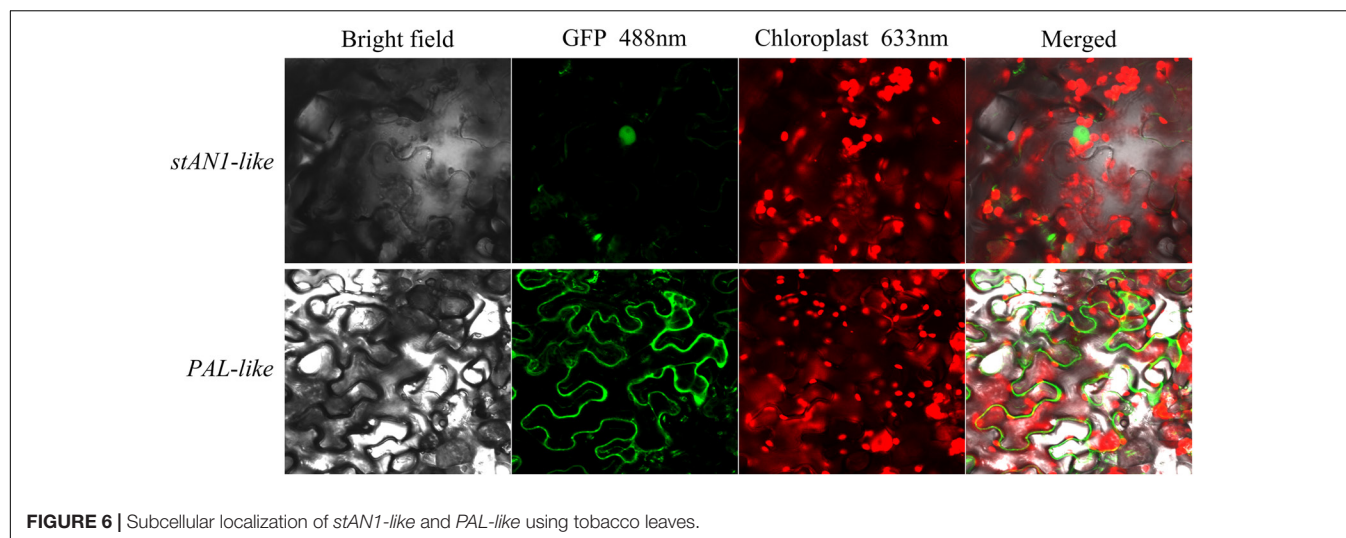
**FIGURE 5 |** Verification of key gene expression. **(A)** Relative expression levels of key genes obtained by semi-quantitative RT-PCR. The semi-quantitative RT-PCR experiment of each plant tissue was performed on 5 biological replicates. **(B)** The results of agarose gel electrophoresis corresponding to semi quantitative RT-PCR tests. **(C)** Potato materials used in this experiment. Y1, Yellow Meigui 1; R3, Red Meigui 3; P2, Purple Meigui 2. \*Significant difference ( $p$ -value  $< 0.05$ ); \*\*highly significant difference ( $p$ -value  $< 0.01$ ).

This further demonstrates that the functions of the R2R3-MYB transcription factors are important for plants.

## The Function of New Transcripts of *stAN1* and *PAL*

The new transcripts of *stAN1* and *PAL* in this experimental clone were from our own laboratory material Red Meigui 3. The new transcripts were named *stAN1-like* and *PAL-like*, respectively. The cloned *stAN1-like* amino acid sequence differs from *stAN1*

(Supplementary Figure S2), which has been reported to regulate potato color (Zhang et al., 2009; Liu et al., 2016). The *stAN1-like* transcript has 21 bases more than the 5' end of the *stAN1* reference transcript. By comparing the *stAN1-like* transcript with the *stAN1* reference gene sequence, it was found that the 21 bases were completely identical to the *stAN1* reference gene sequence. It can be clarified that the production of *stAN1-like* transcripts is caused by the changes of transcription initiation sites or splicing sites of the pre-mRNA. Therefore, it is necessary to further study the role of *stAN1-like* in potato anthocyanins synthesis



**FIGURE 6 |** Subcellular localization of *stAN1-like* and *PAL-like* using tobacco leaves.

and plant color change. The *PAL* gene also plays an important role in the accumulation of potato anthocyanins (Zhang and Liu, 2015), but *PAL-like* is different from the typical *PAL* gene (Supplementary Figure S3). Therefore, it is impossible to rule out the possibility that proteins produced by *PAL-like* guidance have other functions. The function of *PAL-like* needs further research through molecular biological methods.

## Biosynthesis and Accumulation Process of Anthocyanins

The R2R3-MYB transcription factor mainly regulates the transcription of downstream genes controlling anthocyanin synthesis, such as DFR (Nesi et al., 2001). The results of comprehensive transcriptome analysis showed that the upstream genes controlling the synthesis of anthocyanin precursors represented by *PAL* (PGSC0003DMT400055489) were mainly expressed in leaves. However, the R2R3-MYB transcription factor genes represented by *stAN1-like* (PGSC0003DMT400036281) were mainly concentrated in stems and tubers. This indicates that there is a transport process during the synthesis and accumulation of anthocyanins throughout the potato. Anthocyanin precursors such as phenylalanine and tyrosine accumulate in leaves; then the intermediate products are gradually catalyzed to form the final product (anthocyanins) in the process of transport to the tubers; the final end product accumulates in the tuber in the form of anthocyanins. The whole process is synthesized while transporting, rather than directly accumulating the final product of anthocyanin biosynthesis in the leaves and then transferring to the tubers.

Analysis of transcriptomic data revealed that the role of *PAL* gene in the overall anthocyanin biosynthesis process is not critical. In anthocyanin biosynthesis, the metabolic step that really plays a pivotal role should be the following process: The anthocyanin synthesis precursor *p*-cinnamoyl-CoA is transformed into naringenin chalcone as much as possible, thereby entering the subsequent synthesis process of

anthocyanins, so that *p*-cinnamoyl-CoA enters the synthesis pathway of lignin as little as possible. In colored potatoes, the expression of *CHS* was up-regulated, and the down-regulated expression of *HCT* effectively realized this process. Therefore, the up-regulation of *CHS* and the down-regulation of *HCT* should be the most critical link to promote plant anthocyanin synthesis and increase the plant anthocyanin content. In addition, the high expression of the *F3'5'H* gene effectively offsets the effect of the down-regulated expression of *HCT* on the anthocyanin composition type, so that the composition of each type of anthocyanin can remain relatively balanced.

## Application of Multi-Omics Joint Analysis in Experiments

With the development of bioinformatics and the accumulation of experimental data in the field of plant life sciences, it has become possible for multi-omics to jointly analyze a certain life phenomenon (Zhang et al., 2010; Lakshmanan et al., 2015). The transcriptomics data used in this paper were different from the traditional RNA-seq data. This experiment combined multiple RNA-seq results for comprehensive analysis. Potato could be used as a good model plant to study the process of anthocyanin synthesis and accumulation. However, due to the lack of research on potato gene function, it is difficult to perform transcriptome analysis and annotation, especially for transcription factor-related genes. At the same time, potato proteomics and metabolomics experimental data are still insufficient, and the analytical methods are limited, which make the relevant life phenomena unable to be fully analyzed. Future scientific research needs to further complement data on potato-related proteomics, metabolomics, and phenomics. With the advancement of life sciences, the above problems will surely be gradually solved.

The anthocyanin metabolism and synthesis process is a typical quantitative trait, and the synthesis process is controlled by multiple genes. In this experiment, the genomic and



transcriptome analysis indicated that the most important step in the anthocyanin synthesis process was to transfer *p*-cinnamoyl-CoA into the flavonoid biosynthesis process instead of further metabolism-producing lignin species. Up-regulation of *CHS* and down-regulation of *HCT* played a central role in anthocyanin biosynthesis. Through this analysis, we strived to find the major genes that regulate quantitative traits and convert quantitative traits into quality traits. At the same time, it was preliminarily found that anthocyanins synthesized precursor substances in leaves that were then gradually transformed during transport, and finally, end products (anthocyanins) accumulated in potato tubers. After a comprehensive analysis, two new transcripts with research potential were found, namely, *stAN1-like* and *PAL-like*, and their functions were preliminarily studied. However, the specific functions of these two transcripts still require the construction of transgenic plants for further research and validation. This study provides a reference for the comprehensive analysis and application of multiple transcriptomics data in the context of big data. At the same time, it also provides a reference for the application of R programming language in GO and KEGG analysis of non-model plants. Finally, the results of this study provide a solid theoretical basis for increasing the anthocyanin content in potato tubers, cultivating new potato varieties with high anthocyanin content and regulating plant color.

## AUTHOR CONTRIBUTIONS

NT completed the main content of this manuscript. WD and MX made language retouching of this manuscript. CQ and CY provided guidance for the experiments.

## REFERENCES

- Albert, N. W., Lewis, D. H., Zhang, H., Schwinn, K. E., Jameson, P. E., and Davies, K. M. (2011). Members of an R2R3-MYB transcription factor family in *Petunia* are developmentally and environmentally regulated to control complex floral and vegetative pigmentation patterning. *Plant J.* 65, 771–784. doi: 10.1111/j.1365-3113X.2010.04465.x
- Asen, S., Stewart, R. N., and Norris, K. H. (1972). Co-pigmentation of anthocyanins in plant tissues and its effect on color. *Phytochemistry* 11, 1139–1144. doi: 10.1016/S0031-9422(00)88467-8
- Baudry, A., Caboche, M., and Lepiniec, L. (2006). TT8 controls its own expression in a feedback regulation involving TTG1 and homologous MYB and bHLH factors, allowing a strong and cell-specific accumulation of flavonoids in *Arabidopsis thaliana*. *Plant J.* 46, 768–779. doi: 10.1111/j.1365-3113X.2006.02733.x
- Besseau, S., Hoffmann, L., Geoffroy, P., Lapiere, C., Pollet, B., and Legrand, M. (2007). Flavonoid accumulation in *Arabidopsis* repressed in lignin synthesis affects auxin transport and plant growth. *Plant Cell* 19, 148–162. doi: 10.1105/tpc.106.044495
- Brown, C. R., Culley, D., Yang, C., Durst, R., and Wrolstad, R. (2005). Variation of anthocyanin and carotenoid contents and associated antioxidant values in potato breeding lines. *J. Am. Soc. Hortic. Sci.* 130, 174–180. doi: 10.21273/JASHS.130.2.174
- Dai, J., and Mumper, R. J. (2010). Plant phenolics: extraction, analysis and their antioxidant and anticancer properties. *Molecules* 15, 7313–7352. doi: 10.3390/molecules15107313

## FUNDING

This work was supported by the National Natural Science Foundation of China (No. 31601358), the Project of Science and Technology from Shaanxi Province (No. 2017ZDXM-NY-004) and major collaborative innovation projects for production, education and research in Yangling demonstration zone (No. 2016CXY-05), the National Key Research and Development Program of China (2018YFD0200805).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2019.00603/full#supplementary-material>

**FIGURE S1** | *PAL-like* transient expression of tobacco leaves with plasmolysis.

**FIGURE S2** | The amino acid sequence of *stAN1-like* was aligned with the reference sequence.

**FIGURE S3** | The amino acid sequence of *PAL-like* was aligned with the reference sequence.

**TABLE S1** | Other candidate genes found by collinear analysis.

**TABLE S2** | Genes with significantly different expression in the transcriptome.

**TABLE S3** | Genes with differential expression in the potato MYB-R2R3 gene family.

**TABLE S4** | The results of GO and KEGG analysis related to anthocyanin biosynthesis and accumulation based on transcriptome data.

**TABLE S5** | RNA sequencing data regrouping.

**TABLE S6** | Primers used in this experiment.

- Feller, A., Machemer, K., Braun, E. L., and Grotewold, E. (2011). Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. *Plant J.* 66, 94–116. doi: 10.1111/j.1365-3113X.2010.04459.x
- Fernandes, I., Faria, A., Calhau, C., de Freitas, V., and Mateus, N. (2014). Bioavailability of anthocyanins and derivatives. *J. Funct. Foods* 7, 54–66. doi: 10.1016/j.jff.2013.05.010
- Gálvez, J. H., Tai, H. H., Lagüe, M., Zebarth, B. J., and Strömvik, M. V. (2016). The nitrogen responsive transcriptome in potato (*Solanum tuberosum* L.) reveals significant gene regulatory motifs. *Sci. Rep.* 6:26090. doi: 10.1038/srep26090
- Hannapel, D. J., Sharma, P., and Lin, T. (2013). Phloem-mobile messenger RNAs and root development. *Front. Plant Sci.* 4:257. doi: 10.3389/fpls.2013.00257
- Hichri, I., Barrieu, F., Bogs, J., Kappel, C., Delrot, S., and Lauvergeat, V. (2011). Recent advances in the transcriptional regulation of the flavonoid biosynthetic pathway. *J. Exp. Bot.* 62, 2465–2483. doi: 10.1093/jxb/erq442
- Jaakola, L. (2013). New insights into the regulation of anthocyanin biosynthesis in fruits. *Trends Plant Sci.* 18, 477–483. doi: 10.1016/j.tplants.2013.06.003
- Jones, P., Binns, D., Chang, H., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031
- Jung, C. S., Griffiths, H. M., De Jong, D. M., Cheng, S., Bodis, M., Kim, T. S., et al. (2009). The potato developer (D) locus encodes an R2R3 MYB transcription factor that regulates expression of multiple anthocyanin structural genes in tuber skin. *Theor. Appl. Genet.* 120, 45–57. doi: 10.1007/s00122-009-1158-3
- Katihar, A., Smita, S., Lenka, S. K., Rajwanshi, R., Chinnusamy, V., and Bansal, K. C. (2012). Genome-wide classification and expression analysis of MYB transcription factor families in rice and *Arabidopsis*. *BMC Genomics* 13:544. doi: 10.1186/1471-2164-13-544



- Lachman, J., and Hamouz, K. (2005). Red and purple coloured potatoes as a significant antioxidant source in human nutrition—a review. *Plant Soil Environ.* 51, 477–482. doi: 10.1007/s11104-004-0337-x
- Lakshmanan, M., Lim, S., Mohanty, B., Kim, J. K., Ha, S., and Lee, D. (2015). Unraveling the light-specific metabolic and regulatory signatures of rice through combined in silico modeling and multi-omics analysis. *Plant Physiol.* 169, 3002–3020. doi: 10.1104/pp.15.01379
- Liu, S., Wang, X., Li, E., Douglas, C. J., Chen, J., and Wang, S. (2013). R2R3 MYB transcription factor PtrMYB192 regulates flowering time in Arabidopsis by activating flowering locus C. *J. Plant Biol.* 56, 243–250. doi: 10.1007/s12374-013-0135-1
- Liu, Y., Lin-Wang, K., Deng, C., Warran, B., Wang, L., Yu, B., et al. (2015). Comparative transcriptome analysis of white and purple potato to identify genes involved in anthocyanin biosynthesis. *PLoS One* 10:e129148. doi: 10.1371/journal.pone.0129148
- Liu, Y., Lin-Wang, K., Espley, R. V., Wang, L., Yang, H., Yu, B., et al. (2016). Functional diversification of the potato R2R3 MYB anthocyanin activators AN1, MYBA1, and MYB113 and their interaction with basic helix-loop-helix cofactors. *J. Exp. Bot.* 67, 2159–2176. doi: 10.1093/jxb/erw014
- Lobo, V., Patil, A., Phatak, A., and Chandra, N. (2010). Free radicals, antioxidants and functional foods: impact on human health. *Pharm. Rev.* 4, 118–126. doi: 10.4103/0973-7847.70902
- Martens, S., Preuß, A., and Matern, U. (2010). Multifunctional flavonoid dioxygenases: flavonol and anthocyanin biosynthesis in *Arabidopsis thaliana* L. *Phytochemistry* 71, 1040–1049. doi: 10.1016/j.phytochem.2010.04.016
- Miguel, M. G. (2011). Anthocyanins: antioxidant and/or anti-inflammatory activities. *J. Appl. Pharm. Sci.* 1, 7–15.
- Nesi, N., Jond, C., Debeaujon, I., Caboche, M., and Lepiniec, L. (2001). The Arabidopsis TT2 gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *Plant Cell* 13, 2099–2114. doi: 10.1105/tpc.13.9.2099
- Patel, R. K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 7:e30619. doi: 10.1371/journal.pone.0030619
- Pham, G. M., Newton, L., Wiegert Rinninger, K., Vaillancourt, B., Douches, D. S., and Buell, C. R. (2017). Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. *Plant J.* 92, 624–637. doi: 10.1111/tpj.13706
- Pietrini, F., Iannelli, M. A., and Massacci, A. (2002). Anthocyanin accumulation in the illuminated surface of maize leaves enhances protection from photo-inhibitory risks at low temperature, without further limitation to photosynthesis. *Plant Cell Environ.* 25, 1251–1259. doi: 10.1046/j.1365-3040.2002.00917.x
- Salvatierra, A., Pimentel, P., Moya-Leon, M. A., Caligari, P. D., and Herrera, R. (2010). Comparison of transcriptional profiles of flavonoid genes and anthocyanin contents during fruit development of two botanical forms of *Fragaria chiloensis* ssp. *chiloensis*. *Phytochemistry* 71, 1839–1847. doi: 10.1016/j.phytochem.2010.08.005
- Scalbert, A., Manach, C., Morand, C., Rémésy, C., and Jiménez, L. (2005). Dietary polyphenols and the prevention of diseases. *Crit. Rev. Food Sci. Nutr.* 45, 287–306. doi: 10.1080/10408690509096
- Stintzing, F. C., and Carle, R. (2004). Functional properties of anthocyanins and betalains in plants, food, and in human nutrition. *Trends Food Sci. Technol.* 15, 19–38. doi: 10.1016/j.tifs.2003.07.004
- Stracke, R., Ishihara, H., Huep, G., Barsch, A., Mehrrens, F., Niehaus, K., et al. (2007). Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the *Arabidopsis thaliana* seedling. *Plant J.* 50, 660–677. doi: 10.1111/j.1365-313X.2007.03078.x
- Tanaka, Y., Brugliera, F., Kalc, G., Senior, M., Dyson, B., Nakamura, N., et al. (2010). Flower color modification by engineering of the flavonoid biosynthetic pathway: practical perspectives. *Biosci. Biotechnol. Biochem.* 74, 1760–1769. doi: 10.1271/bbb.100358
- Tanaka, Y., Sasaki, N., and Ohmiya, A. (2008). Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. *Plant J.* 54, 733–749. doi: 10.1111/j.1365-313X.2008.03447.x
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016
- Velioglu, Y. S., Mazza, G., Gao, L., and Oomah, B. D. (1998). Antioxidant activity and total phenolics in selected fruits, vegetables, and grain products. *J. Agric. Food Chem.* 46, 4113–4117. doi: 10.1021/jf9801973
- Vogt, T. (2010). Phenylpropanoid biosynthesis. *Mol. Plant* 3, 2–20. doi: 10.1093/mp/ssp106
- Xu, W., Dubos, C., and Lepiniec, L. (2015). Transcriptional control of flavonoid biosynthesis by MYB-bHLH-WDR complexes. *Trends Plant Sci.* 20, 176–185. doi: 10.1016/j.tplants.2014.12.001
- Xu, X., Paik, I., Zhu, L., and Huq, E. (2015). Illuminating progress in phytochrome-mediated light signaling pathways. *Trends Plant Sci.* 20, 641–650. doi: 10.1016/j.tplants.2015.06.010
- Xu, X., Pan, P., Cheng, S., Zhang, B., Mu, D., Ni, P., et al. (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475, 189–195. doi: 10.1038/nature10158
- Yang, A., Dai, X., and Zhang, W. (2012). A R2R3-type MYB gene, OsMYB2, is involved in salt, cold, and dehydration tolerance in rice. *J. Exp. Bot.* 63, 2541–2556. doi: 10.1093/jxb/err431
- Yu, G., Wang, L., Han, Y., and He, Q. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118
- Zhang, L., Zhao, G., Xia, C., Jia, J., Liu, X., and Kong, X. (2012). A wheat R2R3-MYB gene, TaMYB30-B, improves drought stress tolerance in transgenic Arabidopsis. *J. Exp. Bot.* 63, 5873–5885. doi: 10.1093/jxb/ers237
- Zhang, W., Li, F., and Nie, L. (2010). Integrating multiple ‘omics’ analysis for microbial biology: application and methodologies. *Microbiology* 156, 287–301. doi: 10.1099/mic.0.034793-0
- Zhang, X., and Liu, C. (2015). Multifaceted regulations of gateway enzyme phenylalanine ammonia-lyase in the biosynthesis of phenylpropanoids. *Mol. Plant* 8, 17–27. doi: 10.1016/j.molp.2014.11.001
- Zhang, Y., Jung, C. S., and De Jong, W. S. (2009). Genetic analysis of pigmented tuber flesh in potato. *Theor. Appl. Genet.* 119, 143–150. doi: 10.1007/s00122-009-1024-3
- Zhao, L., Gao, L., Wang, H., Chen, X., Wang, Y., Yang, H., et al. (2013). The R2R3-MYB, bHLH, WD40, and related transcription factors in flavonoid biosynthesis. *Funct. Integr. Genomics* 13, 75–98. doi: 10.1007/s10142-012-0301-4
- Zhou, J., Lee, C., Zhong, R., and Ye, Z. (2009). MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. *Plant Cell* 21, 248–266. doi: 10.1105/tpc.108.063321
- Zimmermann, I. M., Heim, M. A., Weisshaar, B., and Uhrig, J. F. (2004). Comprehensive identification of *Arabidopsis thaliana* MYB transcription factors interacting with R/B-like BHLH proteins. *Plant J.* 40, 22–34. doi: 10.1111/j.1365-313X.2004.02183.x

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Tengkun, Dongdong, Xiaohui, Yue and Qin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Advantages of publishing in Frontiers



## OPEN ACCESS

Articles are free to read  
for greatest visibility  
and readership



## FAST PUBLICATION

Around 90 days  
from submission  
to decision



## HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,  
and constructive  
peer-review



## TRANSPARENT PEER-REVIEW

Editors and reviewers  
acknowledged by name  
on published articles

## Frontiers

Avenue du Tribunal-Fédéral 34  
1005 Lausanne | Switzerland

Visit us: [www.frontiersin.org](http://www.frontiersin.org)

Contact us: [info@frontiersin.org](mailto:info@frontiersin.org) | +41 21 510 17 00



## REPRODUCIBILITY OF RESEARCH

Support open data  
and methods to enhance  
research reproducibility



## DIGITAL PUBLISHING

Articles designed  
for optimal readership  
across devices



## FOLLOW US

@frontiersin



## IMPACT METRICS

Advanced article metrics  
track visibility across  
digital media



## EXTENSIVE PROMOTION

Marketing  
and promotion  
of impactful research



## LOOP RESEARCH NETWORK

Our network  
increases your  
article's readership