RESEARCH ARTICLE

Benchmarking RNA-seq differential expression analysis methods using spike-in and simulation data

Bukyung Baik[1], **Sora Yoon**[1], **Dougu Nam**[1,2]*

1 Department of Biological Sciences, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea, **2** Department of Mathematical Sciences, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea

* dougnam@unist.ac.kr

Abstract

Benchmarking RNA-seq differential expression analysis methods using spike-in and simulated RNA-seq data has often yielded inconsistent results. The spike-in data, which were generated from the same bulk RNA sample, only represent technical variability, making the test results less reliable. We compared the performance of 12 differential expression analysis methods for RNA-seq data, including recent variants in widely used software packages, using both RNA spike-in and simulation data for negative binomial (NB) model. Performance of edgeR, DESeq2, and ROTS was particularly different between the two benchmark tests. Then, each method was tested under most extensive simulation conditions especially demonstrating the large impacts of proportion, dispersion, and balance of differentially expressed (DE) genes. DESeq2, a robust version of edgeR (edgeR.rb), voom with TMM normalization (voom.tmm) and sample weights (voom.sw) showed an overall good performance regardless of presence of outliers and proportion of DE genes. The performance of RNA-seq DE gene analysis methods substantially depended on the benchmark used. Based on the simulation results, suitable methods were suggested under various test conditions.

OPEN ACCESS

Citation: Baik B, Yoon S, Nam D (2020) Benchmarking RNA-seq differential expression analysis methods using spike-in and simulation data. PLoS ONE 15(4): e0232271. https://doi.org/10.1371/journal.pone.0232271

Editor: Li Chen, Indiana University School of Medicine, UNITED STATES

Received: November 1, 2019

Accepted: April 10, 2020

Published: April 30, 2020

Copyright: © 2020 Baik et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: TCGA KIRC count data are obtained from the Broad Institute GDAC (URL: http://gdac.broadinstitute.org), Bottomly count data from ReCount (URL: http://bowtie-bio.sourceforge.net/recount/), and SEQC count data from GEO database with accession number GSE49712 (URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49712). The R package for our benchmark tests is available from github (URL: https://github.com/unistbig/compareDEtools).

Introduction

High-throughput cDNA sequencing (RNA-seq) has been commonly used for transcriptome analysis for the last decade [1]. Whereas the hybridization-based method (microarray) can only be used to measure the expression of preselected genes, RNA-seq is able to cover the whole transcriptome and has additional applications [2–8]. Even for the measurement of gene expression, RNA-seq exhibited better reproducibility and sensitivity, particularly for genes with low expression [9]. One of the main purposes of transcriptome analysis is to identify differentially expressed genes (DE genes) between different cellular conditions. Because the gene expression in RNA-seq data is represented by 'count' of short cDNA reads aligned to corresponding gene (or exon), discrete probabilities, such as Poisson or negative binomial (NB) distributions, have been used to model the read counts [9–13]. Based on these models, a variety of

Funding: National Research Foundation (NRF) of Korea, Genomics Program [2016M3C9A3945893]; The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: DE, differentially expressed; NB, negative binomial; SEQC, Sequencing Quality Control; AUC, area under receiver operating curve; TPR, true positive rate; FDR, false discovery rate; FPC, false positive count.

methods have been developed for differential expression analysis (DE analysis) of RNA-seq data [8, 14–16].

Several comparative studies for DE analysis methods have already been conducted to suggest suitable methods for different test conditions. Some have used real RNA-seq data to test false positive control, compared the reproducibility of DE analysis results using a reduced number of samples, and analyzed similarity of results between different methods [16–18]. Others have used benchmark data such as spike-in RNA-seq [14, 19, 20] or simulated read counts [14, 15, 21] for a more systematic comparison of false positive (or discovery) control, power, and overall discriminatory ability. Especially, simulation-based analysis enabled the comparison of methods under various test conditions related to outlier counts, sample size, proportion of DE genes, and balance of DE genes [14, 15]. Although a few methods with overall good performance have been suggested, their performance was different between studies as different benchmark datasets and test conditions were used. Moreover, new versions or variants of the widely used DE analysis packages, namely, edgeR, DESeq, and limma with improved parameter estimation or outlier treatment have been added in recent years.

In this study, (1) we compared the performance of 12 DE analysis methods using both RNA spike-in [14] and NB-based simulation data. Since the spike-in data were generated from the same bulk RNA sample (technical replicates), they exhibited much smaller dispersion estimates in NB model [22, 23] as compared to regular RNA-seq data. In addition, the proportion of DE genes in spike-in data was only 0.27%. We demonstrate that these two features of spike-in data, which are rarely observed in RNA-seq data based on biological replicates, could have caused some difference in performance between spike-in and simulation-based test results. (2) We then performed simulation analysis under various conditions using parameters estimated from RNA-seq data for biological replicates. Our tests covered the largest range of test conditions to date and demonstrated the large impact of proportion of DE genes on some DE analysis methods (e.g., edgeR and DESeq) and the robust performance of ROTS for unbalanced DE genes, in addition to confirming the dominant effect of dispersion parameter. Based on the test results, methods with a good performance were suggested for each condition. (3) Two real RNA-seq datasets were analyzed for further comparison.

Methods

In this article, only two-sample group comparison and 'gene'-based analysis will be considered to focus on the performance of DE analysis methods. We consider a raw read count matrix R = $[R_{ij}]$, $i = 1,2,\ldots,n$, $j = 1,2,\ldots,(m_1+m_2)$ composed of non-negative integers where n is the number of genes analyzed, and m_1 and m_2 are sample sizes for the test and control groups, respectively.

Modeling read counts using discrete probability

A read count R_{ij} of the ith gene in the jth sample has often been modeled by discrete counting probability, such as Poisson or NB distribution. Poisson distribution is only able to address the experimental variation between replicates caused by random sampling and is appropriate for modeling the counts obtained from technical replicates [9, 10]. However, samples from different individuals (biological replicates) incorporate biological variation as well. Therefore, in biological replicate data, means are surpassed by large variances, whereas means and variances have more similar values in technical-replicate-based data.

To incorporate the increased variability in biological replicates, NB distribution has been widely used for modeling RNA-seq read counts [11–13, 24]. NB distribution has two parameters, i.e., mean $\mu_{ij} \geq 0$ and dispersion $\varphi_i \geq 0$, and its variance is represented as $\sigma_{ij}^2 = \mu_{ij} + \varphi_i \mu_{ij}^2$.

Here, $\mu_{ij} = s_j \mu_i$, where s_j is scaling factor (or depth) of sample j, and μ_i is the mean expression of gene i. If φ_i is zero, NB becomes a Poisson distribution where mean and variance have the same value. Zero-inflated NB distribution was considered to model the sparse single-cell RNA-seq data [25, 26], but may not be appropriate to represent RNA-seq data. In this study, we used NB distribution to simulate RNA-seq read count data.

DE analysis methods and comparison of their performance

We selected such DE analysis methods for RNA-seq data that exhibited relatively good performances in a previous comparative study [14–16, 18, 21]. Additionally, recent variants in widely used R packages were included. The tested methods are as follows: edgeR exact test (edgeR) [11], GLM version of edgeR (edgeR.glm) [27], robust version of edgeR (edgeR.rb) [28], quasi-likelihood edgeR method (edgeR.ql) [29, 30], robust version of edgeR.ql (edgeR.ql.rb) [31], DESeq [12], DESeq with per condition dispersion estimation (DESeq.pc), DESeq2 [32], voom with TMM (voom.tmm) [33] and quantile normalization (voom.qn) [34], voom with sample weights (voom.sw) [35, 36], baySeq [24], PoissonSeq [37], SAMseq [38], and ROTS [19]. baySeq method has been implemented with either TMM or quantile normalization (denoted as baySeq and baySeq.qn, respectively). ROTS was basically applied to voom transformed data with TMM normalization; however, ROTS was also applied to raw count data when DE genes were unbalanced between up- and down-regulated genes. A brief summary of recently developed methods as well as simulation test results of this study are described in Table 1. Detailed simulation conditions are described in 'Simulation conditions' section.

Table 1. Summary and test results of recently developed methods.

Methods	Summary	Test Results
DESeq2 [32]	Empirical shrinkage estimation of dispersions and logarithmic fold-changes. Z-test is used for DE analysis. Both outliers of dispersions and logarithmic fold-changes are treated.	DESeq2 exhibited steady and good performances regardless of outliers, sample size, proportion of DE genes, dispersions, and mean counts.
edgeR.rb [28]	Observations that deviate strongly from the model fit are given lower weights. These observation weights affect both the regression and dispersion estimates. Used when data include outlier counts.	edgeR.rb yielded more DE genes and more false positives compared to other edgeR methods. In presence of outliers and large number of samples ($\geq$10), it exhibited outperforming AUCs.
edgeR.ql [29]	While edgeR exact test assumes the estimated dispersion is true, quasi-likelihood estimation accounts for the uncertainty of the dispersion estimates. This approach improves type I error control.	edgeR.ql showed better AUC, control of true FDR, and FPCs compared with edgeR methods, but exhibited relatively low power.
voom.qn /voom.tmm [34]	Read counts were quantile normalized (voom.qn) or normalized with TMM method (voom.tmm), and then were transformed using voom. A moderated *t*-test is used for DE analysis.	voom.tmm performed better than voom.qn except outlying sample cases. They exhibited overall good performance for most cases, but their powers were relatively low. AUC of voom.qn was noticeably decreased compared to other voom methods as the proportion of DE genes increased.
voom.sw [36]	Observations from highly variable samples are down-weighted for more accurate estimation of regression coefficients. Used when some samples have amplified dispersions	voom.sw performed like voom.tmm rather than voom.qn and showed overall good performance. When samples with amplified dispersions were included, voom.sw outperformed other methods.
ROTS [19]	voom transformation with TMM normalization and bootstrap are used for generalized *t*-statistic that maximally reproduce preselected top k% genes. k = 25 was used for our tests.	ROTS exhibited good AUC and false positive control. ROTS applied to raw count data showed slightly lowered power, but it outperformed other methods when DE genes were unbalanced and a large number of DE genes were included.

https://doi.org/10.1371/journal.pone.0232271.t001

Performance of the aforementioned DE analysis methods was compared using their area under receiver operating curve (AUC), true positive rate (TPR), true false discovery rate (FDR), and false positive counts (FPCs). Here, true FDR is the proportion of non-DE genes among the significant genes and indicates the extent of reliability of predicted DE genes. We calculated true FDR only when five or more significant genes were detected in each method. FPC is the number of significant genes detected from the datasets where no DE genes were included (Type I error). Compared to the previous simulation study [15], we used a liberal threshold q-value < 0.1; this is because we applied smaller effect sizes (1.2 fold or larger) when generating DE genes. Among the compared methods, edgeR and edgeR.glm; and edgeR.ql and edgeR.ql.rb showed virtually the same results in simulation tests. DESeq.pc showed similar or better performance as compared to DESeq across all test conditions. Thus, the results of edgeR.glm, edgeR.ql.rb, and DESeq were removed to simplify the analysis. baySeq.qn results were also removed from the simulation results because the two baySeq results showed nearly the same performance in simulation tests.

SEQC spike-in benchmark data

Sequencing Quality Control (SEQC) data pertained to the Microarray Quality Control (MAQC) study and contained replicated samples of universal human body reference RNA and human brain reference RNA with spike-in controls. We used the same count data as used in a previous DE analysis benchmark study (GEO accession GSE49712) [14] (denoted as SEQC data). SEQC data comprised two sample groups, each having five samples and 21,711 genes including 92 spike-in transcripts from External RNA Controls Consortium (ERCC). They exhibited very small dispersion estimates, which were on an average 22.5 times smaller than those of RNA-seq data of cancer samples (TCGA Kidney Renal Clear Cell dataset). We removed genes that had less than ten mean read counts and analyzed the rest 17,961 genes, which included 63 spike-in transcripts composed of 15, 16, 17, and 15 genes with 0.25, 1.5, 2, and 1 fold changes, respectively. AUCs were calculated using the 63 spike-in genes. In short, SEQC data were characterized by very small dispersion values and low proportion of DE genes (0.27%).

Simulation conditions

The above-described DE analysis methods were tested under the following conditions:

1. (pDE) Different proportion of DE genes: pDE = 0.27%, 5%, 10%, 30% or 60% DE genes were included in the simulation datasets. 0.27% was considered to compare with the results for SEQC data, and 60% DE genes represent, for example, complex disease conditions such as cancer vs. normal conditions.
2. (SS) Sample sizes of three or ten were used in each condition.
3. (OL1) Random outliers: each read count could become an outlier with a probability of 0.05 or less. An outlier count was regenerated with five to ten times larger mean value.
4. (OL2) Low-quality sample: dispersions of 30% of samples were increased fivefold.
5. (OL3) Composite condition of 3 and 4: 30% of samples were generated with fivefold increased dispersions, and the 3% of read counts in the other samples were regenerated as a random outlier.
6. (Bal) Different balances of up and downregulated genes: Bal = 50%, 70% or 90% of DE genes were upregulated, while the remainder were downregulated.

7. (Disp) The same or different dispersion values between conditions were assigned for each gene.
8. (EF) Weak effect sizes (EF = 1.2 or larger fold changes for ten sample cases; 1.5 or larger for three sample cases) were used.

These conditions include diverse situations encountered in RNA-seq data analysis and the test results can provide a guideline for selecting an appropriate method in each situation. Our tests have the following several differences from the previous simulation study by Soneson and Delorenzi [15]:

1. Four additional edgeR methods, two additional limma methods, PoissonSeq, DESeq2, and ROTS were included in our study.
2. Wider range of DE gene proportions from 0.27% to 60% was tested, whereas only 10% and 33% DE genes were considered previously. We demonstrated that both low (0.27% or 5%) and high (60%) proportions of DE genes have substantial effects on performance of some methods (e.g., edgeR, edgeR.ql, and ROTS). The low pDE (e.g. 5%) represents the case, such as gene-knock-out experiment where only a small number of genes are expected to be differentially expressed, whereas the high pDE (e.g., 60%) represents the case of complex disease, such as cancer as compared to normal samples where the majority of genes are expected to be differentially expressed [39].
3. Effects of dispersions and mean counts on DE analysis methods were analyzed. We specifically showed that dispersions have great impacts on both the relative and absolute performances of each method.
4. In our study, weak effect sizes (1.2 or larger fold changes) were used to analyze subtle differences in performance, whereas 1.5 or larger fold changes were used previously.

Simulation of read counts and parameter estimation

To simulate read counts, we first estimated mean and dispersion parameters for each gene from two RNA-seq datasets with different parameter distributions: TCGA Kidney Renal Clear Cell Carcinoma/normal dataset, which is available from GDAC (URL: http://gdac.broadinstitute.org) (denoted as KIRC) and inbred mouse dataset (denoted as Bottomly) [40]. KIRC is composed of independent individual samples and showed large mean counts and dispersions, whereas Bottomly is composed of samples from genetically identical mice and showed smaller means and dispersions (5.6 times and 10 times, respectively). Here, *independent samples* denote those obtained from different individuals with possibly different genetic backgrounds, such as KIRC. Independent samples typically exhibit large dispersion values [22]. We additionally built two kinds of synthetic datasets using hybrid combinations of mean and dispersion parameters obtained from KIRC and Bottomly datasets, in order to compare the effects of mean and dispersion parameters. These parameters were estimated from each of the test and control groups using edgeR package. Especially, edgeR provided the common and tagwise dispersion estimates that were used for the same and different dispersion conditions in our simulation study, respectively. All the read counts were simulated using the 'rnbinom' R function and estimated parameters. The inverse of the dispersion values was used for 'size' argument. To generate fold changes for a DE gene, we added a random number sampled from exponential distribution, exp(1) to the minimum fold changes of 1.5 (three sample case) and 1.2 (ten sample case) [15], respectively. Mean values in the test condition were multiplied or divided by these fold change values to generate up- or down-regulated DE genes, respectively.

To generate outlier counts, we randomly selected approximately 5% (or less) of all read counts and regenerated those counts with five or ten times larger mean values in both test and control conditions. To generate low quality sample data, we randomly selected one sample and three samples from each condition for three and ten sample cases, respectively and regenerated those samples using five times larger dispersion values for all genes. Read counts were generated for 10,000 genes using KIRC parameters and for 5,000 genes using Bottomly or hybrid parameters. In addition, the two real RNA-seq datasets (KIRC and Bottomly) were analyzed using 12 DE analysis methods and the results were compared. All of our analyses can be reproduced using the R package which is available from https://github.com/unistbig/compareDEtools.

Results and discussion

Comparison of test results for SEQC spike-in and simulation data

We tested the 13 DE analysis methods including two different implementations of baySeq for both, SEQC spike-in data and KIRC-based simulation data and compared their AUCs. The AUCs for SEQC data are shown in Fig 1A. ROTS outperformed other methods and was followed by PoissonSeq and baySeq.qn. SAMseq and DESeq2 showed the worst AUCs. baySeq results based on two normalizations showed a considerable difference. Then, read counts of 10,000 genes were simulated using the parameters estimated from KIRC data. Three or five samples were included in each sample group. In particular, a low pDE (0.27%) and 67% up regulated DE genes (Bal = 67%) were assumed to compare with the results for SEQC data. The results for KIRC-based simulation are shown in Fig 1B. edgeR, edgeR.ql, and DESeq.pc showed relatively low AUCs and DESeq2 showed a much better AUCs as compared to the SEQC analysis results. Such a large difference in relative performance of each method between benchmark studies confuses researchers while choosing a suitable method.

Hence, we performed further simulation tests to investigate the cause of such differences, by adjusting dispersion parameters and pDE. First, the dispersions of KIRC simulation data were decreased to the level of SEQC data (22.5 times lower) and AUCs were compared (Fig 1C). Because such small dispersions caused nearly perfect AUCs for all methods, we applied rather reduced fold changes to simulate DE genes (0.625, 1.15, and 1.3). As a result, high AUC of edgeR.ql and relatively low AUC of DESeq2 shown in SEQC results were reproduced. Then, we slightly increased pDE from 0.27% to 1% while keeping the large dispersions of KIRC (S1D Fig). In this case, only the AUC for edgeR.ql increased while DESeq2 remained the same. These results indicate that small dispersions of SEQC caused the relatively low performance of DESeq2, and both, low dispersion and low pDE of SEQC supported the high performance of edgeR.ql in SEQC result. However, the high performances of ROTS, edgeR, and DESeq.pc in SEQC analysis result were not explained by these adjustment experiments. The TPR and true FDR results are also shown in S1 Fig.

Overall, this experiment shows the extremely small dispersion and proportion of DE genes in SEQC data may not provide fully reliable evaluation of DE analysis methods and simulation tests based on parameters for biological replicates should also be considered. A similar concern of benchmarking SEQC data has also been raised when normalizing RNA-seq data [41]. Thus, we herein present the evaluation of 12 DE analysis methods using NB-based simulation under extensive test conditions.

Simulation tests under various conditions

The test results are depicted as boxplots in Fig 2 and S2–S10 Figs. Basically, we compared the performance of methods based on AUCs; however, we have also considered TPRs, true FDR,

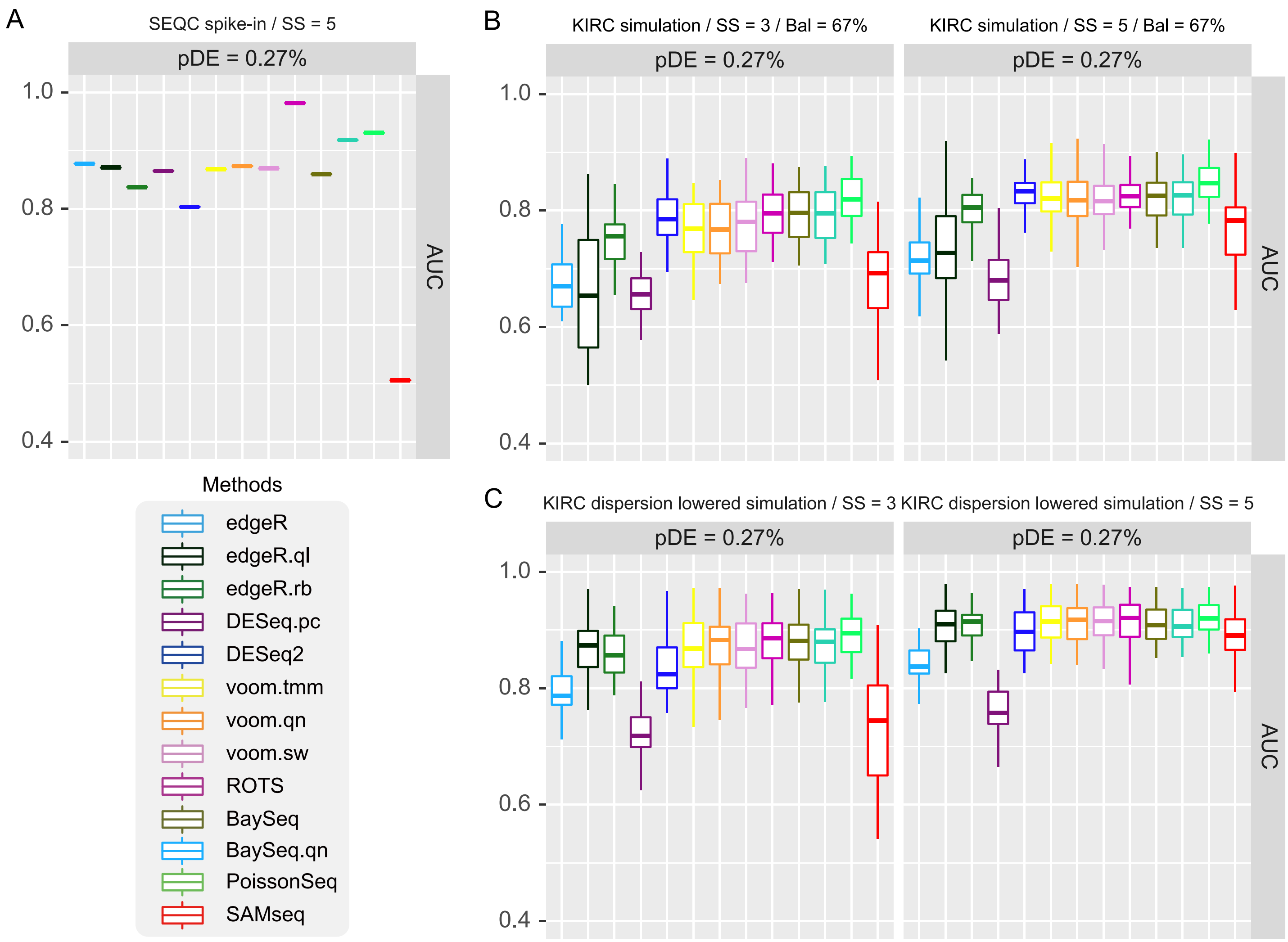

Fig 1. Performance comparison of 13 DE analysis methods for SEQC benchmark data and simulation data based on KIRC parameters. Three and five samples were used in simulation tests. (A) SEQC spike-in data test, 5 sample condition, 0.25, 1.5, 2 fold changes are introduced to DE genes. (B) KIRC based simulation. 1.5 or larger and 1.3 or larger fold changes are introduced to 0.27% of 10,000 genes for 3 and 5 samples, respectively. (C) Same condition as B but fixed fold changes, 0.625, 1.15, and 1.3 were introduced to DE genes. Dispersions were also lowered to SEQC level (22.5 times).

https://doi.org/10.1371/journal.pone.0232271.g001

and FPC to select 'recommended' methods under each test condition (Table 2). We attached '(LP)' if a method exhibited good AUC but rather low power (TPR). If AUCs were similar, the priority was given to 'liberal' methods with high power and high type I error rate, considering that researchers may have additional means or information to reduce false positives. We mainly considered equal proportions of up and down regulated genes among DE genes (Bal = 50%). We further considered the unbalanced cases where 70% and 90% of DE genes were upregulated (Bal = 70% or 90%).

Test results with large dispersions and large mean counts: Independent samples. Read counts of 10,000 genes were generated using the parameters estimated from KIRC data. Three or ten samples were generated in each sample group. To begin with, we tested for the 'base' condition where no outliers and no low quality samples were included, and same dispersions between sample conditions and equal proportions of up and down regulated DE genes (Bal = 50%) were assumed. Then, we further tested for alternative test conditions. Under the

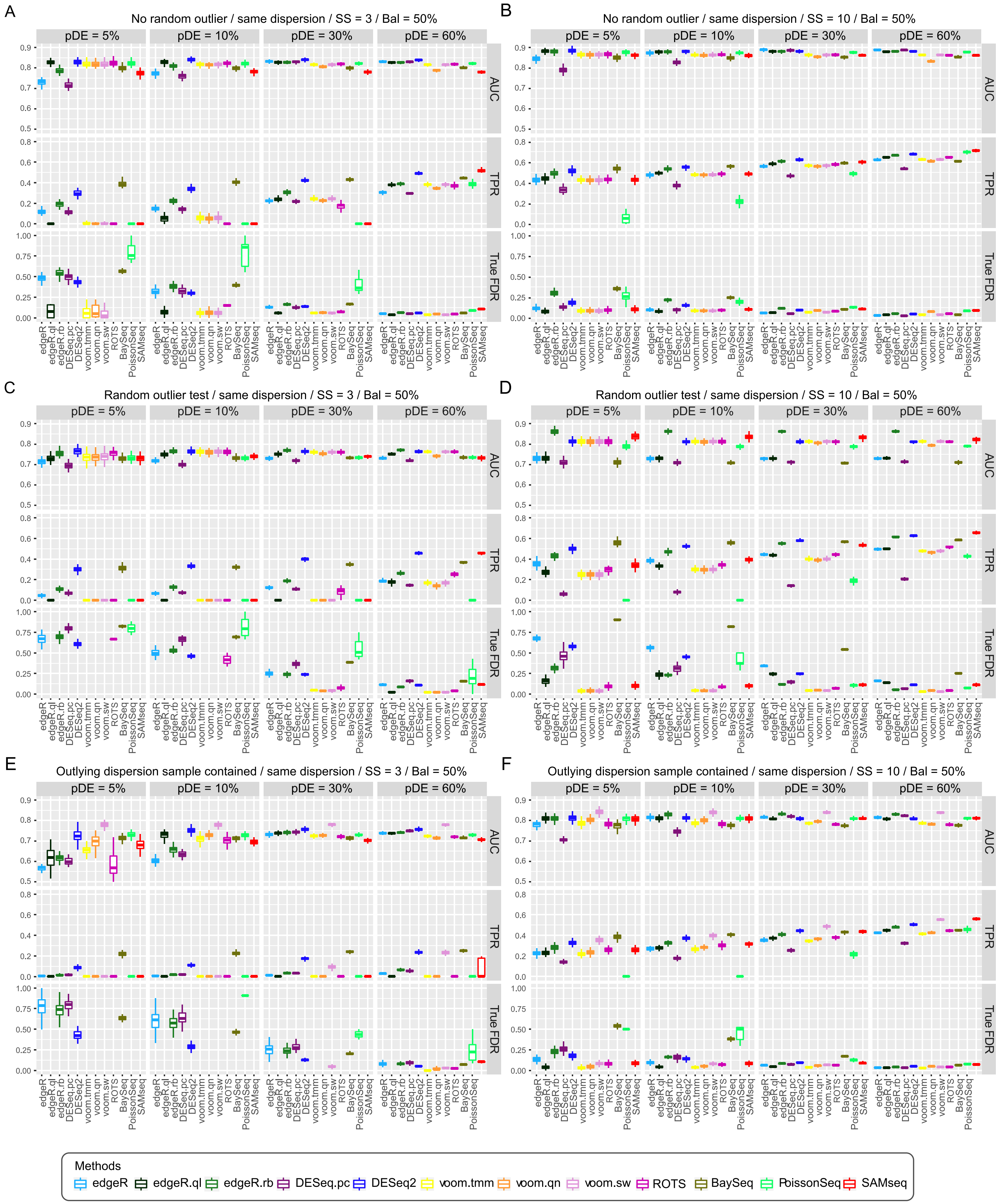
A
No random outlier / same dispersion / SS = 3 / Bal = 50%
B
No random outlier / same dispersion / SS = 10 / Bal = 50%
C
Random outlier test / same dispersion / SS = 3 / Bal = 50%
D
Random outlier test / same dispersion / SS = 10 / Bal = 50%
E
Outlying dispersion sample contained / same dispersion / SS = 3 / Bal = 50%
F
Outlying dispersion sample contained / same dispersion / SS = 10 / Bal = 50%
pDE = 5%
pDE = 10%
pDE = 30%
pDE = 60%
AUC
TPR
True FDR
0.5
0.6
0.7
0.8
0.9
0.0
0.2
0.4
0.6
0.8
0.00
0.25
0.50
0.75
1.00
edgeR
edgeR.ql
edgeR.rb
DESeq.pc
DESeq2
voom.tmm
voom.qn
voom.sw
ROTS
BaySeq
PoissonSeq
SAMseq
Methods
edgeR
edgeR.ql
edgeR.rb
DESeq.pc
DESeq2
voom.tmm
voom.qn
voom.sw
ROTS
BaySeq
PoissonSeq
SAMseq

Fig 2. Performance comparison of 12 DE analysis methods for simulated RNA-seq data with KIRC parameters. Area under ROC curve (AUC), true positive rate (TPR) and true false discovery rate (FDR) are shown. Three and ten samples were used in each sample group for (A, C, E) and (B, D, F), respectively. (A, B) Same dispersion between test and control groups, Bal = 50%, four different proportions of DE genes (5%, 10%, 30%, and 60%) were used. (C, D) Random ourlier counts: the same condition as A-B but each read count can be a random outlier with 5% probability. (E, F) Low quality samples: the same condition as (A, B) but one and three samples of each sample group have fivefold increased dispersion parameters, respectively. True FDR graphs of some methods did not appear in low pDE condition because they detected no DE genes.

https://doi.org/10.1371/journal.pone.0232271.g002

base condition for three samples, DESeq2 performed best (Fig 2A). Difference in the proportion of DE genes (pDE) showed dramatic changes in the edgeR and DESeq.pc test results. These two methods did not show good AUCs under low pDE condition ($\leq$ 5%); however, they performed as well as DESeq2 under high pDE condition ($\geq$ 30%). edgeR.ql showed considerably good AUCs, but showed very low power when three samples were used and pDE was not larger than 10%. Voom methods (voom.qn, voom.tmm, and voom.sw) exhibited similarly good AUCs in low pDE condition (5%); however, the AUC of voom.qn was lowered with the increase of pDE. A larger sample size of ten allowed every method to show better performance (Fig 2B); SAMseq, edgeR, and edgeR.rb showed large improvements in AUC.

In presence of outliers (OL1 = 5%), the best AUCs were observed for DESeq2 and edgeR.rb followed by voom.tmm, and voom.sw when three samples were used. Especially with ten samples, edgeR.rb surpassed every other method, and SAMseq was the second best. DESeq2, voom methods, and ROTS showed similarly good AUCs. We further compared DESeq2 and edgeR.rb by additionally introducing lower proportion of outliers (1% and 3%) to the base condition (S4 Fig). edgeR.rb method tended to exceed DESeq2 when pDE or sample size increased and larger proportion of outliers was introduced, and vice versa. voom.sw and voom.tmm showed similarly good performance under most test conditions. When samples with enhanced dispersions (i.e., low quality samples) were included, voom.sw outperformed other methods (Fig 2E and 2F). Under this condition, edgeR-based methods and DESeq.pc showed low AUCs for low pDE (5%) and small sample size conditions (Fig 2E). Overall, AUCs and powers tended to increase as the proportion of DE genes increased for many of the test conditions.

Next, when we applied different dispersions between conditions, the AUCs of most methods except those for edgeR, edgeR.rb, and DESeq.pc, were slightly decreased (S2 Fig). Lastly,

Table 2. Recommended methods for DE analysis of biological replicate RNA-seq data (simulation results).

Outlier	Sample size	%DE gene	Independent replicates (large dispersions)	Genetically identical replicates (small dispersions)
No outlier counts	3	5%	DESeq2, edgeR.ql (LP)	DESeq2 and most other methods
		≥ 30%	DESeq2, edgeR, DESeq.pc	DESeq2 and most other methods
	10	5%	DESeq2, edgeR.ql, edgeR.rb	DESeq2, edgeR methods, ROTS, PoissonSeq
		≥ 30% [a](Bal≥0.7)	edgeR and DESeq methods, (%ROTS)	edgeR methods, DESeq2, ROTS, (%ROTS)
With outlier counts	3	5%	DESeq2, edgeR.rb	DESeq2, ROTS
		≥ 30%	DESeq2, edgeR.rb	edgeR.rb, DESeq2, ROTS
	10	5%	edgeR.rb, SAMseq, DESeq2, voom methods, ROTS	SAMseq, DESeq2, edgeR.rb
		≥30% [a](Bal≥0.9)	edgeR.rb, SAMseq, DESeq2, voom methods, ROTS, (%ROTS)	edgeR.rb, DESeq2, SAMseq, (%ROTS)
With outlying samples	3	5%	Voom.sw	Voom.sw, edgeR.rb
		≥ 30%	Voom.sw, DESeq2	Voom.sw, edgeR.rb
	10	5%	Voom.sw	Voom.sw, edgeR.rb
		≥ 30%	Voom.sw, edgeR.rb	Voom.sw, edgeR.rb

[a] Under these conditions, ROTS without normalization (denoted as %ROTS) performed best. LP denotes low power.

https://doi.org/10.1371/journal.pone.0232271.t002

when the fraction of upregulated DE genes was increased (Bal = 70% and 90%), the overall performance worsened (S3 Fig). The impact of this imbalance of DE genes increased as the pDE increased. When pDE was 60% and Bal = 90%, the majority of genes were up-regulated (60% * 90% = 54%), and most of the tested methods exhibited a nearly random prediction ($AUC < 0.6$). Interestingly, ROTS applied to raw count data *without any normalization or transformation* was much less affected by the varying proportion and imbalance of DE genes and showed outperforming AUCs when $Bal \geq 0.7$ and $pDE \geq 0.3$ (S3 and S7 Figs). We note that the same depth was applied to each sample in our simulation tests.

True FDRs tended to decrease as pDE increased in most cases. Voom methods, edgeR.ql, and SAMseq exhibited overall quite good control of true FDRs, while PoissonSeq and baySeq showed large true FDRs. ROTS, edgeR, and DESeq exhibited poor true FDRs for three sample case and small DE proportions ($pDE \leq 10\%$), but they were dramatically improved when pDE or sample size was increased. True FDRs were considerably increased as random outlier counts were introduced (Fig 2C and 2D). They were also increased by inclusion of low quality samples for KIRC case when SS = 3 (Fig 2E). True FDRs were not much affected by low quality samples when SS = 10, except for PoissonSeq (Fig 2F).

We then compared the FPCs in each case by randomly sampling two groups from the same sample group (Fig 3); in other words, no DE genes were included. edgeR.ql, voom methods, PoissonSeq, and SAMseq strictly controlled FPCs, while baySeq yielded large FPCs. Although PoissonSeq exhibited very poor true FDR control, its FPC control was observed to be quite good. This shows that FPC control alone is not enough to assess the predicted DE genes, and true FDR control should also be considered. Interestingly, the two methods that account for outlier counts viz. edgeR.rb and DESeq2, did not perform better than other methods in either true FDR or FPC control (Fig 3C and 3D). They exceeded other methods in AUCs.

We also tested the composite conditions of random outliers and increased dispersions (OL3) for a large number of samples (10 and 30 samples in each condition) (S5 Fig). The overall results seemed to reflect both the results of OL1 and OL2, and edgeR.rb showed the best AUC followed by those of limma.sw and SAMseq. Interestingly, DESeq2 which performed well under both OL1 and OL2 showed a low AUC and high true FDR under this complex simulation condition.

Test results with small dispersions and small mean counts (genetically identical replicate case); and hybrid parameters. The same conditions were tested using the parameters estimated from Bottomly data. It is known that large mean counts and low dispersions increase the statistical power of DE analysis [21, 22]. Despite the low depths in Bottomly data, their low dispersions allowed all the methods to show much better AUCs and powers than those obtained using KIRC parameters (S6–S8 Figs). Under the base condition, many methods including DESeq2 and edgeR.rb showed similarly high AUCs; therefore, unique features of each method were not clearly distinguishable. Unlike in the KIRC case, edgeR performed as well as other edgeR-based methods when pDE was low (5%). In presence of outliers and using ten samples, the best performing methods were edgeR.rb, DESeq2, and SAMseq followed by voom based methods. True FDRs for Bottomly were overall much improved as compared to those for KIRC, and their patterns were overall similar to each other.

To compare the impacts of mean and dispersion parameters, we additionally created simulation data with hybrid parameters by combining KIRC means and Bottomly dispersions (denoted as mKdB) and Bottomly means and KIRC dispersions (denoted as mBdK), respectively. As shown in S10 and S11 Figs, all 12 methods showed high powers for low dispersion datasets (mKdB and Bottomly), while they showed relatively low powers for high dispersion datasets (mBdK and KIRC). In contrast, mean had only limited effects; performance for mKdB dataset was only slightly better than that for Bottomly dataset, and the performances for mBdK and KIRC datasets were observed to be nearly the same.

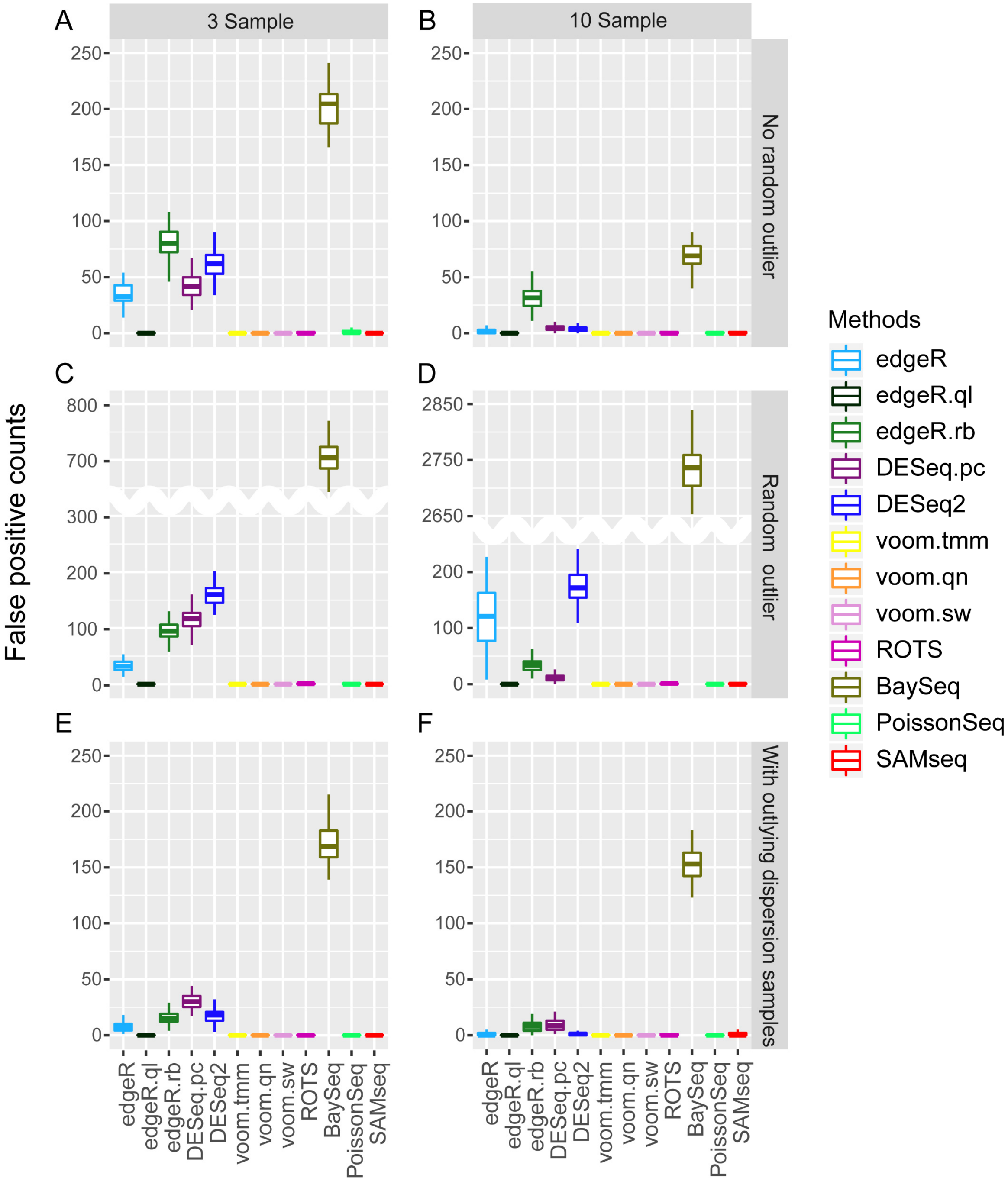

Fig 3. False positive count comparison of 12 DE analysis methods for simulated RNA-seq data with KIRC parameters. Three and ten samples in each sample group and three different conditions (No random outlier, random outlier, and outlying dispersion sample) were used for comparison.

https://doi.org/10.1371/journal.pone.0232271.g003

A guide for choosing methods. We have shown through simulation tests that dispersion values, proportion of DE genes, and outlier counts largely impact the performance of each method. However, it is not always clear which condition in Table 2 corresponds to the data at hand except for the sample size. We have previously demonstrated that RNA-seq data generated from independent samples tended to exhibit large dispersion values (approximately, ≥

0.1), whereas those obtained from genetically identical replicates yielded relatively small dispersion values (approximately, 0.01–0.1) [22]. Thus, the replicate type used is able to roughly suggest the distribution of dispersions. To be more precise, the researcher may also use existing packages, such as edgeR, directly to estimate dispersion values from their own read count data.

For outlier counts, DESeq2 automatically truncates them or simply remove genes that contained a potential outlier count depending on the sample size. However, for edgeR package, the user has to choose a method between edgeR and its robust version. If the given data are highly noisy, edgeR.rb may be most effective. For moderately noisy data, both DESeq2 and edgeR.rb perform better than their original methods. Recently developed methods can help detect outlier counts in RNA-seq data [42, 43].

Lastly, the proportion of DE genes are not known to the user. However, we can roughly estimate the proportion from the test condition. In complex disease, such as cancer, the majority of genes are differentially expressed compared to the normal condition, where the condition $\geq$30% can be considered. If a single gene with potentially limited effects was knocked out, much less proportion of DE genes are expected compared to complex disease conditions. In ambiguous cases, several DE analysis methods can be tested together to roughly estimate DE genes from significantly detected genes.

Real RNA-seq data analysis

We also compared the 12 DE analysis methods by analyzing the two RNA-seq datasets (KIRC and Bottomly) that exhibited different distribution of parameters. As the true DE genes are not known for real data, we only compared the number of significantly detected genes and FPCs. KIRC contained 72 normal and 72 cancer samples. Bottomly contained ten C57BL/6J strain and eleven DBA/2J strain samples. Three, five, ten, and twenty (KIRC only) sample sizes were considered. After removing the genes with average read count of less than ten, 16,621 genes and 8,550 genes remained for KIRC and Bottomly, respectively. We used q-value < 0.1 to select DE genes. To compare FPCs, we used only the normal sample group for KIRC, where two contrasting groups were randomly sampled for DE analysis. The same analysis was done for the DBA/2J sample group of Bottomly. Each experiment was repeated fifty times and the results are represented as boxplots in Fig 3A and 3B.

Most results showed similar patterns between the two datasets except that PoissonSeq detected relatively small number of DE genes for Bottomly. SAMseq and voom.sw detected the largest numbers of DE genes, while DESeq.pc and ROTS detected the smallest numbers of DE genes for both datasets. We assumed that KIRC included many DE genes between cancer and normal conditions (e.g., pDE = 60%), and found high similarity between the number of detected DE genes (Fig 4A; 3 and 10 samples) and the TPR of simulation results (Fig 2A and 2B; pDE = 60%); we may expect such a similarity though DE genes and TPR represent different concepts. We also found a similarity in FPCs in that edgeR.ql, voom methods, ROTS, PoissonSeq, and SAMseq exhibited very small or no FPCs while baySeq exhibited relatively large FPCs in both datasets.

Additionally, we compared the similarity of DE analysis results between the 12 methods. Five samples were randomly selected from each sample group, and DE analysis was repeated fifty times using each method. Based on our observation that most methods detected approximately 5,000 and 500 DE genes with five samples for KIRC and Bottomly, respectively (Fig 4A and 4B), the top 5,000 and 500 DE genes in each method were selected based on DE q-values for KIRC and Bottomly, respectively. The selected DE genes obtained using each method were pooled and sorted by their average DE q-values. Then, Spearman rank correlations between each pair of methods were calculated. The similarity matrix as well as clustering results for

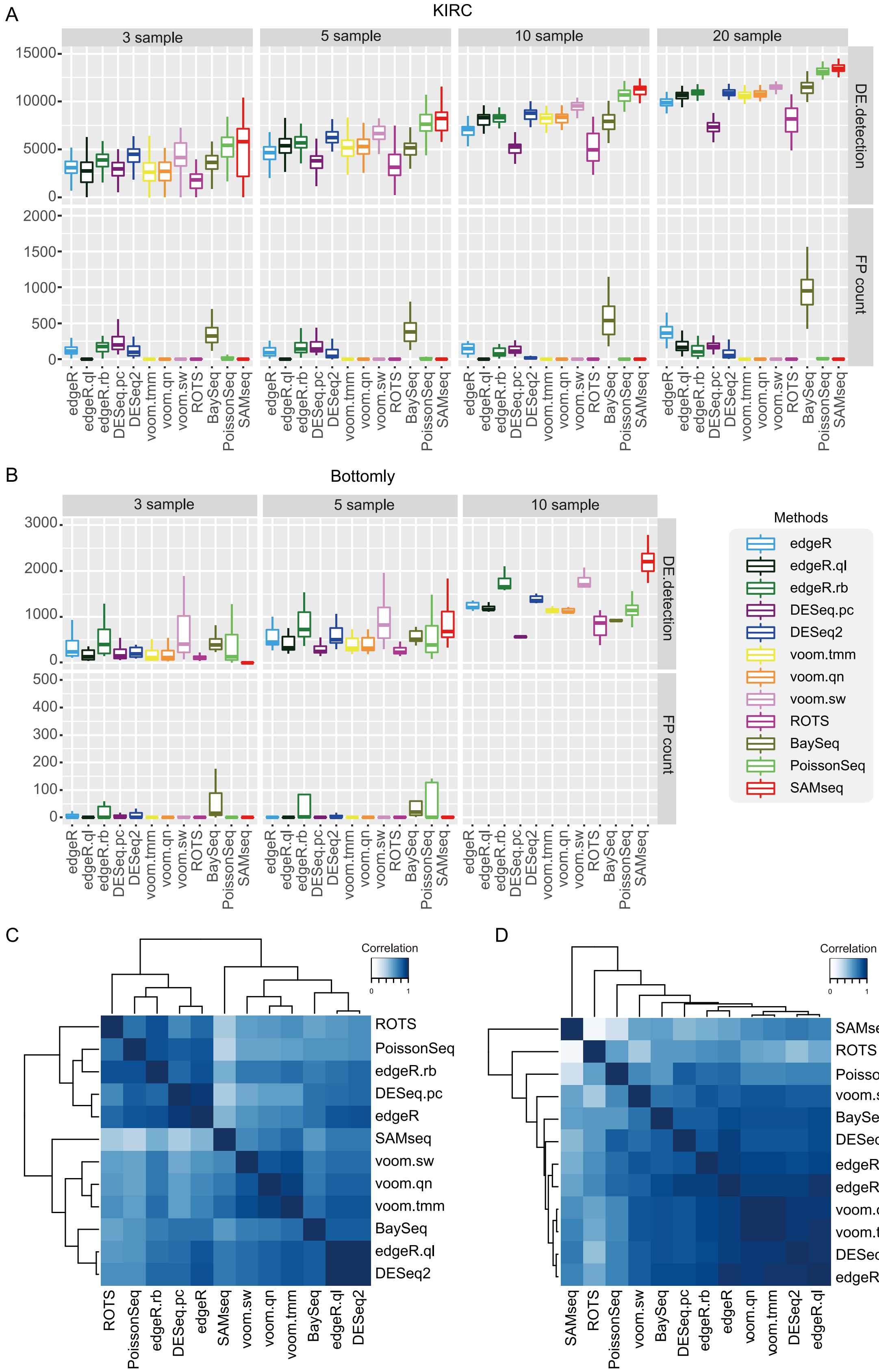
A
KIRC
3 sample
5 sample
10 sample
20 sample
DE.detection
FP count
15000
10000
5000
0
2000
1500
1000
500
0
edgeR
edgeR.ql
edgeR.rb
DESeq.pc
DESeq2
voom.tmm
voom.qn
voom.sw
ROTS
BaySeq
PoissonSeq
SAMseq
B
Bottomly
3 sample
5 sample
10 sample
DE.detection
FP count
3000
2000
1000
0
500
400
300
200
100
0
Methods
edgeR
edgeR.ql
edgeR.rb
DESeq.pc
DESeq2
voom.tmm
voom.qn
voom.sw
ROTS
BaySeq
PoissonSeq
SAMseq
C
Correlation
0
1
ROTS
PoissonSeq
edgeR.rb
DESeq.pc
edgeR
SAMseq
voom.sw
voom.qn
voom.tmm
BaySeq
edgeR.ql
DESeq2
D
Correlation
0
1
SAMseq
ROTS
PoissonSeq
voom.sw
BaySeq
DESeq.pc
edgeR.rb
edgeR
voom.qn
voom.tmm
DESeq2
edgeR.ql

Fig 4. Analysis on KIRC and Bottomly. (A, B) The number of DE genes between two sample groups and FP counts under varying sample sizes. (A) DE genes between normal and cancer sample groups and 3 to 20 sample size conditions were tested for KIRC. (B) DE genes between C57BL/6J strain and DBA/2J strain sample groups and 3 to 10 sample size conditions were tested for Bottomly. (C, D) Hierarchical clustering of DE genes obtained from 12 DE analysis methods based on spearman rank correlation. Top 5000 and 500 DE genes are used to calculate spearman rank correlation for KIRC and Bottomly datasets, respectively.

https://doi.org/10.1371/journal.pone.0232271.g004

KIRC and Bottomly are shown in Fig 4C and 4D, respectively. In both cases, edgeR.ql and DESeq2 were closely combined. Voom methods were also closely combined to each other. However, voom.sw was separated from other voom methods in Bottomly, thus implying the presence of outlying samples in Bottomly. Indeed, principal component analysis for the two datasets (S12 Fig) showed samples in KIRC exhibited a relatively homogeneous and compact distribution in each condition, while those in Bottomly showed a rather heterogeneous distribution and less clear difference between conditions. ROTS was most distantly clustered from other methods in both cases. Similar clustering patterns were reproduced when we used top 3,000 and 300 genes in KIRC and Bottomly, respectively (S13 Fig).

Comparison of the computing time

Lastly, the running times for each method for analyzing KIRC data were compared (Fig 5). Each method was executed ten times using ‘proc.time’ R function. A subset of KIRC containing 16,621 genes with 3, 5, 10, and 20 samples were used for this measurement. All the methods were run on a Linux machine with Inter X5670 hexa core processor and 20Gb of 1333MHz DDR3 memory. The two fastest methods were voom.qn and voom.tmm followed by

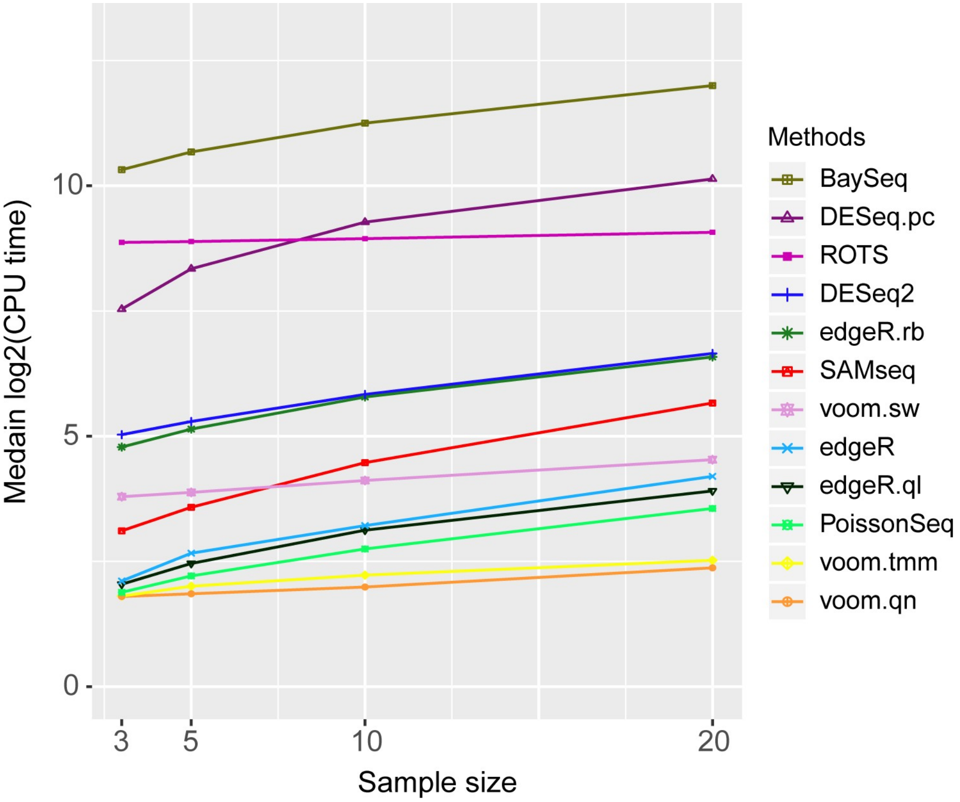

Fig 5. CPU-times to run DE analysis methods for KIRC data. Each method was run 10 times for different sample sizes 3, 5, 10, and 20.

https://doi.org/10.1371/journal.pone.0232271.g005

PoissonSeq and edgeR.ql. Most methods finished computing within 100 cpu-time. The slowest was observed to be baySeq (1,000 to 4,000 cpu-time) followed by DESeq.pc and ROTS. As sample size increased, many methods including baySeq, DESeq.pc, DESeq2, edgeR, and SAMseq showed a linear increase in the computing time.

Conclusion

We compared the performances of DE analysis methods including recent developments available from popularly used packages, for RNA spike-in, simulated read counts, and real RNA-seq data. We demonstrated the extremely small dispersions and proportion of DE genes in the spike-in data could cause quite different benchmarking results. We performed most extensive simulation tests based on NB model and obtained many new or more detailed results in addition to confirming previously tested results. Our test results suggest appropriate methods in each experimental condition.

Supporting information

S1 Fig. Performance comparison of 13 DE analysis methods for SEQC benchmark and simulation data. Area under ROC curve (AUC), true positive rate (TPR) and true false discovery rate (true FDR) are shown. Three and five samples were used in simulation tests. (A) Test results for SEQC spike-in data (5 samples). 0.25, 1.5, and 2 fold changes were introduced to DE genes. (B) KIRC based simulation. Not less than 1.5 or 1.3 fold changes are introduced to 0.27% of 10,000 genes for 3 and 5 samples, respectively. (C) Two conditions are different from B: fixed fold changes of 0.625, 1.15, and 1.3 were introduced to DE genes, and dispersions were also lowered to SEQC level (22.5 times) (D) The same condition as A but DE proportion was increased to 1%.
(EPS)

S2 Fig. Simulation study with KIRC parameters. Dispersions between test and control groups are different (A) Three samples per each group and four different proportions of DE genes (5%, 10%, 30%, and 60%) were used. (B) Sample size is increased to ten. (C, D) Each read count can be a random outlier with 5% probability.
(EPS)

S3 Fig. Simulation study with KIRC parameters for imbalanced DE proportions where 70% or 90% DE genes are upregulated (Bal = 70% and 90%). No random outlier was included, and same dispersions were used between conditions. (A) Three samples per each group and Bal = 70% were used for four different proportions DE genes (5%, 10%, 30%, and 60%). (B) Bal = 90%. (C, D) The same condition as A-B, but sample size is increased to ten.
(EPS)

S4 Fig. Simulation study for low probabilities (1% and 3%) of each read count being an outlier (KIRC parameters). (A) 1% probability for each read count to be an outlier was applied. Three samples per each group and four different proportions DE genes (5%, 10%, 30%, and 60%) were used. (B) The same condition as A, but the sample size is increased to ten. (C, D) The same as A and B, but the outlier probability of 3% is used.
(EPS)

S5 Fig. Simulation test with KIRC parameters for complex conditions. 30% of each sample group have fivefold increased dispersion parameters and 3% of counts in the other samples were regenerated as outliers. Ten and 30 samples were used in each sample group for (A) and

(B), respectively.
(EPS)

S6 Fig. Performance comparison of 12 DE analysis methods for simulated RNA-seq data with Bottomly parameters. Area under ROC curve (AUC), true positive rate (TPR) and true false discovery rate (true FDR) are shown. Three and ten samples were used in each sample group for (A, C, E) and (B, D, F), respectively. (A, B) Same dispersion between test and control groups, Bal = 50%, four different proportions DE genes (5%, 10%, 30%, and 60%) were used. (C, D) Random outlier counts: the same condition as A-B but each read count can be a random outlier with 5% probability. (E, F) Low quality samples: the same condition as (A, B) but one and three samples of each sample group have fivefold increased dispersion parameters.
(EPS)

S7 Fig. Simulation study with Bottomly parameters. Dispersions between test and control groups are different (A) Three samples per each group and four different proportions of DE genes (5%, 10%, 30%, and 60%) were used. (B) The sample size is increased to ten. (C, D) The same condition as A-B but each read count can be a random outlier with 5% probability.
(EPS)

S8 Fig. Simulation study with Bottomly parameters for imbalanced DE proportions where 70% or 90% DE genes are upregulated (Bal = 70% and 90%). Same dispersions were used between conditions. (A) Three samples per each group and Bal = 70% were used for four different proportions of DE genes (5%, 10%, 30%, and 60%). (B) Bal = 90%. (C, D) Sample size is increased to ten.
(EPS)

S9 Fig. False positive count comparison of 12 DE analysis methods for simulated RNA-seq data with Bottomly parameters. Three and ten samples were used in each sample group for (A, C, E) and (B, D, F), respectively. (A, B) Same dispersions between test and control groups are assumed. (C, D) Random outlier counts: the same condition as A-B but each read count can be a random outlier with 5% probability. (E, F) Low quality samples: the same condition as (A, B) but one and three samples of each sample group have fivefold increased dispersion parameters.
(EPS)

S10 Fig. Simulation study with hybrid parameters (mBdK: Means of Bottomly and dispersions of KIRC). Same dispersions between test and control groups are used. (A) Three samples per each group and four different proportions of DE genes (5%, 10%, 30%, and 60%) were used. (B) Sample size is increased to ten. (C, D) Random outlier counts: the same condition as A-B, but each read count can be a random outlier with 5% probability.
(EPS)

S11 Fig. Simulation study with hybrid parameters (mKdB: Means of KIRC and dispersions of Bottomly). Same dispersions between test and control groups are used. (A) Three samples per each group and four different proportions of DE genes (5%, 10%, 30%, and 60%) were used. (B) Sample size is increased to ten. (C, D) Random outlier counts: the same condition as A-B, but each read count can be a random outlier with 5% probability.
(EPS)

S12 Fig. Principal Component Analysis for (A) KIRC and (B) Bottomly datasets.
(EPS)

S13 Fig. Similarity of DE analysis methods. Hierarchical clustering of DE genes based on spearman rank correlation. The union of top 3000 and 300 DE genes obtained from 12 DE analysis methods were clustered between different methods based on spearman rank correlation for (A) KIRC and (B) Bottomly datasets, respectively.
(EPS)

Author Contributions

Conceptualization: Bukyung Baik, Sora Yoon, Dougu Nam.

Formal analysis: Bukyung Baik, Sora Yoon.

Funding acquisition: Dougu Nam.

Investigation: Bukyung Baik, Sora Yoon.

Methodology: Bukyung Baik, Sora Yoon, Dougu Nam.

Project administration: Dougu Nam.

Supervision: Dougu Nam.

Visualization: Bukyung Baik.

Writing – original draft: Bukyung Baik.

Writing – review & editing: Dougu Nam.

References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009; 10(1):57–63. https://doi.org/10.1038/nrg2484 WOS:000261866500012. PMID: 19015660
2. Peng ZY, Cheng YB, Tan BCM, Kang L, Tian ZJ, Zhu YK, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nature Biotechnology. 2012;30(3):253-+. https://doi.org/10.1038/nbt.2081 WOS:000301303800021.
3. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biology. 2011; 12(1). ARTN R6 https://doi.org/10.1186/gb-2011-12-1-r6 WOS:000289437000010. PMID: 21247443
4. Vidal RO, do Nascimento LC, Mondego JMC, Pereira GAG, Carazzolle MF. Identification of SNPs in RNA-seq data of two cultivars of Glycine max (soybean) differing in drought resistance. Genetics and Molecular Biology. 2012; 35(1):331–U258. WOS:000304829600014.
5. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011; 27(17):2325–9. https://doi.org/10.1093/bioinformatics/btr355 WOS:000294067300001. PMID: 21697122
6. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25(9):1105–11. Epub 2009/03/18. https://doi.org/10.1093/bioinformatics/btp120 PMID: 19289445; PubMed Central PMCID: PMC2672628.
7. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5(7):621–8. Epub 2008/06/03. https://doi.org/10.1038/nmeth.1226 PMID: 18516045.
8. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome biology. 2016; 17:13. Epub 2016/01/28. https://doi.org/10.1186/s13059-016-0881-8 PMID: 26813401; PubMed Central PMCID: PMC4728800.
9. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008; 18(9):1509–17. Epub 2008/06/14. https://doi.org/10.1101/gr.079558.108 PMID: 18550803; PubMed Central PMCID: PMC2527709.
10. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010; 26(1):136–8. Epub 2009/10/27. https://doi.org/10.1093/bioinformatics/btp612 PMID: 19855105.

11. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616 WOS:000273116100025. PMID: 19910308
12. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010; 11(10). ARTN R106 https://doi.org/10.1186/gb-2010-11-10-r106 WOS:000287378900008. PMID: 20979621
13. Lu J, Tomfohr JK, Kepler TB. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. BMC Bioinformatics. 2005; 6:165. Epub 2005/07/01. https://doi.org/10.1186/1471-2105-6-165 PMID: 15987513; PubMed Central PMCID: PMC1189357.
14. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013; 14(9):R95. Epub 2013/09/12. https://doi.org/10.1186/gb-2013-14-9-r95 PMID: 24020486; PubMed Central PMCID: PMC4054597.
15. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013; 14:91. Epub 2013/03/19. https://doi.org/10.1186/1471-2105-14-91 PMID: 23497356; PubMed Central PMCID: PMC3608160.
16. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform. 2015; 16(1):59–70. Epub 2013/12/05. https://doi.org/10.1093/bib/bbt086 PMID: 24300110; PubMed Central PMCID: PMC4293378.
17. Lin Y, Golovnina K, Chen ZX, Lee HN, Negron YL, Sultana H, et al. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. BMC Genomics. 2016; 17(1):28. Epub 2016/01/07. https://doi.org/10.1186/s12864-015-2353-z PMID: 26732976; PubMed Central PMCID: PMC4702322.
18. Schurch NJ, Schofield P, Gierlinski M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? Rna. 2016; 22(6):839–51. Epub 2016/03/30. https://doi.org/10.1261/rna.053959.115 PMID: 27022035; PubMed Central PMCID: PMC4878611.
19. Seyednasrollah F, Rantanen K, Jaakkola P, Elo LL. ROTS: reproducible RNA-seq biomarker detector-prognostic markers for clear cell renal cell cancer. Nucleic Acids Res. 2016; 44(1):e1. Epub 2015/08/13. https://doi.org/10.1093/nar/gkv806 PMID: 26264667; PubMed Central PMCID: PMC4705679.
20. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. PLoS One. 2017; 12(12):e0190152. https://doi.org/10.1371/journal.pone.0190152 PMID: 29267363.
21. Ching T, Huang S, Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. Rna. 2014; 20(11):1684–96. Epub 2014/09/24. https://doi.org/10.1261/rna.046011.114 PMID: 25246651; PubMed Central PMCID: PMC4201821.
22. Yoon S, Nam D. Gene dispersion is the key determinant of the read count bias in differential expression analysis of RNA-seq data. BMC Genomics. 2017; 18(1):408. Epub 2017/05/27. https://doi.org/10.1186/s12864-017-3809-0 PMID: 28545404; PubMed Central PMCID: PMC5445461.
23. Holik AZ, Law CW, Liu R, Wang Z, Wang W, Ahn J, et al. RNA-seq mixology: designing realistic control experiments to compare protocols and analysis methods. Nucleic Acids Res. 2017; 45(5):e30. Epub 2016/12/03. https://doi.org/10.1093/nar/gkw1063 PMID: 27899618; PubMed Central PMCID: PMC5389713.
24. Hardcastle TJ, Kelly KA. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. Bmc Bioinformatics. 2010; 11. Artn 422 https://doi.org/10.1186/1471-2105-11-422 WOS:000281443300002. PMID: 20698981
25. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014; 11(7):740–2. https://doi.org/10.1038/nmeth.2967 PMID: 24836921.
26. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018; 9(1):284. Epub 2014/05/20. https://doi.org/10.1038/s41467-017-02554-5 PMID: 29348443; PubMed Central PMCID: PMC4112276.
27. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012; 40(10):4288–97. Epub 2012/01/31. https://doi.org/10.1093/nar/gks042 PMID: 22287627; PubMed Central PMCID: PMC3378882.
28. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. 2014; 42(11):e91. Epub 2014/04/23. https://doi.org/10.1093/nar/gku310 PMID: 24753412; PubMed Central PMCID: PMC4066750.
29. Lund SP, Nettleton D, McCarthy DJ, Smyth GK. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. Stat Appl Genet Mol Biol. 2012;11(5). Epub 2012/10/30. https://doi.org/10.1515/1544-6115.1826 PMID: 23104842.

30. Lun AT, Chen Y, Smyth GK. It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. Methods Mol Biol. 2016; 1418:391–416. Epub 2016/03/24. https://doi.org/10.1007/978-1-4939-3578-9_19 PMID: 27008025.
31. Phipson B, Lee S, Majewski IJ, Alexander S, Smyth GK. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. Ann Appl Stat. 2016; 10(2).
32. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014; 15(12). ARTN 550 https://doi.org/10.1186/s13059-014-0550-8 WOS:000346609500022. PMID: 25516281
33. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010; 11(3):R25. Epub 2010/03/04. https://doi.org/10.1186/gb-2010-11-3-r25 PMID: 20196867; PubMed Central PMCID: PMC2864565.
34. Law CW, Chen YS, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014; 15(2). ARTN R29 https://doi.org/10.1186/gb-2014-15-2-r29 WOS:000336256600012. PMID: 24485249
35. Ritchie ME, Diyagama D, Neilson J, van Laar R, Dobrovic A, Holloway A, et al. Empirical array quality weights in the analysis of microarray data. BMC Bioinformatics. 2006; 7:261. Epub 2006/05/23. https://doi.org/10.1186/1471-2105-7-261 PMID: 16712727; PubMed Central PMCID: PMC1564422.
36. Liu R, Holik AZ, Su S, Jansz N, Chen K, Leong HS, et al. Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. Nucleic Acids Res. 2015; 43(15):e97. Epub 2015/05/01. https://doi.org/10.1093/nar/gkv412 PMID: 25925576; PubMed Central PMCID: PMC4551905.
37. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics. 2012; 13(3):523–38. Epub 2011/10/18. https://doi.org/10.1093/biostatistics/kxr031 PMID: 22003245; PubMed Central PMCID: PMC3372940.
38. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013; 22(5):519–36. Epub 2011/12/01. https://doi.org/10.1177/0962280211428386 PMID: 22127579; PubMed Central PMCID: PMC4605138.
39. Peng L, Bian XW, Li DK, Xu C, Wang GM, Xia QY, et al. Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types. Sci Rep. 2015; 5:13413. https://doi.org/10.1038/srep13413 PMID: 26292924.
40. Bottomly D, Walter NA, Hunter JE, Darakjian P, Kawane S, Buck KJ, et al. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. PLoS One. 2011; 6(3): e17820. Epub 2011/04/02. https://doi.org/10.1371/journal.pone.0017820 PMID: 21455293; PubMed Central PMCID: PMC3063777.
41. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014; 32(9):896–902. Epub 2014/08/26. https://doi.org/10.1038/nbt.2931 PMID: 25150836; PubMed Central PMCID: PMC4404308.
42. Brechtmann F, Mertes C, Matuseviciute A, Yepez VA, Avsec Z, Herzog M, et al. OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. Am J Hum Genet. 2018; 103(6):907–17. Epub 2018/12/07. https://doi.org/10.1016/j.ajhg.2018.10.025 PMID: 30503520; PubMed Central PMCID: PMC6288422.
43. George NI, Bowyer JF, Crabtree NM, Chang CW. An Iterative Leave-One-Out Approach to Outlier Detection in RNA-Seq Data. PLoS One. 2015; 10(6):e0125224. Epub 2015/06/04. https://doi.org/10.1371/journal.pone.0125224 PMID: 26039068; PubMed Central PMCID: PMC4454687.