# Bias, robustness and scalability in single-cell differential expression analysis

Soneson, Charlotte ; Robinson, Mark D

Abstract: Many methods have been used to determine differential gene expression from single-cell RNA (scRNA)-seq data. We evaluated 36 approaches using experimental and synthetic data and found considerable differences in the number and characteristics of the genes that are called differentially expressed. Prefiltering of lowly expressed genes has important effects, particularly for some of the methods developed for bulk RNA-seq data analysis. However, we found that bulk RNA-seq analysis methods do not generally perform worse than those developed specifically for scRNA-seq. We also present conquer, a repository of consistently processed, analysis-ready public scRNA-seq data sets that is aimed at simplifying method evaluation and reanalysis of published results. Each data set provides abundance estimates for both genes and transcripts, as well as quality control and exploratory analysis reports.

DOI: https://doi.org/10.1038/nmeth.4612

# Bias, robustness and scalability in differential expression analysis of single-cell RNA-seq data

Charlotte Soneson[1,2] and Mark D. Robinson[1,2]

[1] Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
[2] SIB Swiss Institute of Bioinformatics, Zurich, Switzerland

Correspondence to C.S. (charlotte.soneson@uzh.ch) or M.D.R. (mark.robinson@imls.uzh.ch)

## Abstract

We perform an extensive evaluation of the performance and characteristics of 36 approaches for differential gene expression analysis in single-cell RNA-seq, using both experimental and synthetic data. Considerable differences are found between the methods in terms of the number and characteristics of the genes that are called differentially expressed. Prefiltering of lowly expressed genes is shown to have important effects on the results, particularly for some of the methods originally developed for analysis of bulk RNA-seq data. Generally, however, methods developed for bulk RNA-seq analysis do not perform notably worse than those developed specifically for scRNA-seq. We also present *conquer*, a repository of consistently processed, analysis-ready public single-cell RNA-seq datasets, aimed at simplifying method evaluation and reanalysis of published results. Each dataset provides abundance estimates for both genes and transcripts, as well as quality control and exploratory analysis reports.

Keywords: single-cell RNA-seq, comparison, differential expression

## Introduction

RNA-seq is used routinely to characterize transcriptomes, but until recently sequencing libraries had to be prepared from pools of thousands or more cells, and any measurement would represent an *average* across these cells. However, recent advances enable library preparation from minute amounts of RNA and thus profiling of the transcriptomes of individual cells[1–5]. An increasing number of such single-cell RNA-seq (scRNA-seq) datasets are being generated and deposited in public repositories, which typically contain both raw read files and processed data tables with, e.g., estimated gene abundances. Since the aims of different studies vary widely, public datasets are often processed using very different pipelines. Furthermore, the abundances may be represented in different units and sometimes a fraction of the cells and/or genes are filtered out. This can make reuse of the preprocessed public datasets, and especially comparisons across datasets, challenging. To simplify this aspect, we have developed *conquer*, a collection of consistently processed, analysis-ready public scRNA-seq datasets. Each dataset has abundance estimates for all annotated genes and transcripts, as well as quality assessment and exploratory analysis reports to help users determine whether a particular dataset is suitable for their purposes.

One of the most commonly performed computational tasks for RNA-seq data is differential gene expression (DE) analysis. While well-established tools exist for such analysis in bulk RNA-seq data[6–8], methods for scRNA-seq data are just emerging. Due to the special characteristics of scRNA-seq data, including generally low library sizes, high noise levels and a large fraction of so-called "dropout" events, it is unclear whether DE methods developed for bulk RNA-seq are suitable also for scRNA-seq. A few recent studies have started to investigate this question, suggesting that the optimal method choice may depend on the number of cells and the strength of the signal[9], and illustrating that also methods that were not initially developed for RNA-seq analysis can perform well[10]. In this study we use processed datasets, from *conquer* and other sources, to evaluate DE methods in scRNA-seq data. Our study extends the previous comparisons to a larger set of methods and a broader range of experimental datasets, and additionally includes evaluations based on simulated data. We also investigate the effect of filtering out lowly expressed genes and extend the set of employed evaluation criteria. We focus on contrasting two predefined groups of cells since this setup can be accommodated by all considered methods. However, it should be noted that some scRNA-seq datasets contain cells from multiple subjects, or from multiple plates, introducing a hierarchical variance structure that is not accounted for by such a simple model[11]. Moreover, single-cell measurements allow additional questions that can not be addressed with bulk RNA-seq data, such as testing whether different groups of cells show different levels of variability or multimodality[12,13].

## Results

Currently, *conquer* contains 36 datasets: 31 generated with full-length protocols and 5 with 3'-end sequencing (UMI) protocols. With consistent processing and

representation of the datasets, we envision that *conquer* can be useful for a range of applications. It can lower the barriers for evaluations and comparisons of computational methods, for developers as well as end-users, and having easy access to processed data is useful for teaching and tutorial construction. In addition, *conquer* can be used for exploring the generality of biological hypotheses across datasets from different species and cell types.

Seven datasets from *conquer* (six full-length and one UMI dataset) and two additional UMI count datasets were used for the evaluation (Supplementary Table 1, Supplementary Figures 1-2). We keep two predefined groups of cells from each dataset, and generate multiple dataset instances with varying number of cells. For eight datasets, we generate null datasets by subsampling from a single group. Three datasets are used to simulate datasets with signal (10% of the genes differentially expressed) as well as null datasets. For each instance, we apply 36 DE approaches (Supplementary Table 2). Some methods failed to run for certain datasets (Supplementary Figure 3), and these combinations are excluded from the evaluations.

## Number of differentially expressed and non-tested genes

Using all instances of the nine "signal" scRNA-seq datasets, we compare the number of differentially expressed genes called by the different methods at an adjusted p-value cutoff of 0.05 (Supplementary Figures 4-7). For full-length datasets, SeuratBimod[14] (without the default internal filtering) detects the largest number of significant genes. edgeR/QLF[7,15] detects large numbers of genes if the dataset is not prefiltered to remove lowly expressed genes, but shows the largest decrease in the number of significant genes after filtering (Supplementary Figure 8). Conversely, SeuratBimod with non-zero expression threshold, metagenomeSeq[16] and scDD[13] consistently detect few differentially expressed genes. For UMI datasets, the performance of the methods based on the voom transformation[8] is highly variable without gene prefiltering.

Many DE methods implement internal filtering, which means that not all quantified genes are actually being tested for DE. Such filtering is typically performed to exclude lowly expressed genes and increase the power to detect differences in the retained genes[17,18]. For some methods, the model fitting procedure can also fail to converge for some genes. While most evaluated methods report valid results for all genes, some indeed exclude many genes if run with default settings (Supplementary Figures 9-10). This is, however, not specific to scRNA-seq data, and similar patterns can be seen if a subset of the methods are applied to a large bulk RNA-seq dataset[19] (Supplementary Figure 11). If the datasets are filtered before the DE analysis, the fraction of non-reported results decreases, indicating that they mostly correspond to lowly expressed genes.

## Type I error control

Using the eight real null datasets, where no truly differential genes are expected, we evaluate the type I error control by recording the fraction of tested genes that

are assigned a nominal p-value below 0.05 (Figure 1A). For unfiltered datasets, many methods struggle to correctly control the type I error, and the best performance is obtained by ROTS[20,21] and SeuratTobit. Several of the other methods are too liberal, with SeuratBimod and edgeR/QLF standing out with a large number of false positive findings. Setting a non-zero expression threshold in Seurat (SeuratBimodIsExpr2) improves the error control, but at the price of detecting much fewer significant genes (Supplementary Figures 4-7). Conversely, metagenomeSeq, scDD, SCDE[22] and DESeq2[6] on Census counts[23] instead control the false positive rate well below the imposed level. Methods based on voom mostly perform well, but sometimes the number of false positives is very high (Supplementary Figure 12). For UMI datasets, monocle[24] performs best when applied to transcript counts (monoclecount), whereas converting these values to TPMs and applying a tobit model (monocle) deteriorates performance. For full-length datasets, however, the TPM values lead to a slightly better performance than the read counts. After filtering out lowly expressed genes (Figure 1B) the performance of voom-limma, ROTSvoom and edgeR/QLF stabilizes and improves, along with most other methods, while SeuratBimod still assigns low p-values to a large fraction of the tested genes. P-value histograms further illustrate that without filtering, few methods return uniformly distributed p-values while after the applied filtering, results are considerably improved (Supplementary Figures 13-14). The results are largely similar for the three simulated datasets (Supplementary Figure 15).

## Characteristics of false positive genes

To investigate the presence of biases in the DE calling, we use the eight unfiltered real null datasets to characterize the set of genes that are (falsely) called significant by the different methods. For each gene in each dataset instance, we estimate the average, variance and coefficient of variation of the CPM values across all cells as well as the fraction of cells in which the gene is undetected. For each instance, and for each method calling at least five genes DE, we calculate a signal-to-noise statistic comparing the values of each of the four gene characteristics between the significant and non-significant (including non-tested) genes (Figure 2, Supplementary Figure 16). The results show striking differences between the types of genes detected by the different methods. False positives of NODES[25], ROTS, SAMseq[26] and SeuratBimod have few zeros, high expression and mostly a relatively low coefficient of variation. Conversely, false positives of edgeR/QLF, SeuratTobit, MAST[27] and metagenomeSeq have relatively many zeros. The same evaluation performed on the simulated datasets shows largely similar results (Supplementary Figure 17).

## Between-method similarity

Using the nine real scRNA-seq "signal" datasets we quantify the concordance between gene rankings returned by different methods (within-method consistency is investigated in Supplementary Figure 18). For each dataset we calculate the area under the concordance curve (AUCC) for the top-ranked 100 genes for each pair of methods (Online Methods). Averaging the AUCCs across all datasets and clustering based on the resulting similarities (Figure 3) shows, for

example, that while the four MAST modes give overall similar rankings, the inclusion of the detection rate as a covariate has a larger effect on the rankings than changing the type of expression values from CPMs to TPMs. Moreover, the count-based bulk RNA-seq methods cluster together, as do some of the general non-parametric methods (the Wilcoxon test and D3E[28]), which are also similar to the robust count-based methods and several approaches based on log-like transformations of the data. The methods using Census transcript counts as input give similar rankings. The degree of similarity between any given pair of methods can vary widely across the dataset instances (Supplementary Figure 19), but for most method pairs, it is somewhat positively associated with the number of cells per group (Supplementary Figure 20).

## FDR control and power

Using the simulated datasets, we evaluate the false discovery rate control and statistical power of the methods. Several methods, such as voom/limma, ROTStpm, MAST, the methods applied to Census counts, SeuratTobit, SeuratBimod with non-zero expression cutoff and SAMseq, robustly control the FDR close to the imposed level (Figure 4A). SCDE, scDD, the t-test, D3E, limma-trend[8,29], the Wilcoxon test, and the other variants of ROTS control the FDR at a lower level than imposed. The worst FDR control for the unfiltered data is obtained by monocle, SeuratBimod and edgeR/QLF. After filtering, edgeR/QLF improves dramatically (Figure 4B), whereas MAST and SCDE yield even lower false discovery proportions (FDPs). Most methods perform closer to the optimal level for large sample sizes (Supplementary Figure 21). Adjusting the nominal p-values for multiple testing using independent hypothesis weighting[18] with the average expression as covariate rather than using the values returned by the respective methods has only minor impact (Supplementary Figure 22).

Practically all methods show increased power with increased sample size (Figure 4C-D, Supplementary Figure 23). Among the methods with good, robust FDR control after filtering, edgeR/QLF, SAMseq, DEsingle[30] and voom-limma achieve high power, whereas for methods like metagenomeSeq, SeuratTobit, SeuratBimodIsExpr2 and the methods applied to Census counts, the FDR control comes at the price of reduced power. The power to detect true differences is weakly related to the fraction of genes that are excluded by internal filtering procedures (Supplementary Figure 24). However, DESeq2 and NODES achieve high power despite strong filtering. The area under the ROC curve (AUROC), indicating whether the methods are able to rank truly differentially expressed genes ahead of truly non-differential ones, shows favourable performance of edgeR, followed by MAST, limma (voom and trend), SCDE, DEsingle, DESeq2 and SeuratBimod without filtering and the non-parametric methods (Figure 4E). After prefiltering the rankings of most methods are improved (Figure 4F), and the AUROC is typically higher for datasets with more cells (Supplementary Figure 25).

### Other aspects

As the number of cells that are studied in a dataset increases, computational efficiency becomes important for method selection. For comparative purposes, we ran all methods on a single core in this study. However, DESeq2, BPSC[31], MAST, SCDE, scDD and monocle all feature explicit arguments to take advantage of parallelization, and methods that perform gene-wise tests without information sharing between genes, such as the Wilcoxon test, the t-test and D3E, can be run in parallel after splitting the data into chunks. Four dedicated single-cell methods, namely BPSC, DEsingle, D3E and SCDE, are the slowest for most datasets, while the bulk methods (edgeR, DESeq2 and especially the limma variants) are generally faster (Supplementary Figure 26A). Most single-cell methods (with the exception of SCDE) scale well with increasing number of cells, while the computational time required for the bulk RNA-seq methods is more sample size dependent (Supplementary Figures 26B, 27-31).

While the evaluations in this study are centered on the simplest experimental situation, comparing two groups of cells, many real studies require a more complex experimental design, which not all evaluated methods can accommodate. Specifically, the Wilcoxon test, the t-test, scDD, NODES, SCDE, Seurat, ROTS, DEsingle and D3E are limited to two-group comparisons, while SAMseq can perform a limited number of analysis types. The remaining methods implement statistical frameworks that can accommodate more complex (fixed effect) designs, including comparisons across multiple groups and adjustments for batch effects and other covariates.

Other important aspects are the availability and documentation of the software packages. Most methods are available either via Bioconductor[32] or CRAN, or via a public GitHub repository (Supplementary Table 2). NODES was obtained via a Dropbox link provided by the authors. The Bioconductor packages have extensive documentation, including help pages for individual functions and a vignette to guide the user through a typical workflow, all tested to work with the current version of the package. Some packages, such as Seurat, D3E, monocle and SCDE, have dedicated webpages with instructions for users, examples and tutorials.

## Discussion

We have presented an extensive evaluation and comparison of methods for DE analysis of scRNA-seq data, using mainly real datasets from *conquer*, a repository of consistently processed public single-cell RNA-seq datasets. The fact that *conquer* provides gene expression estimates in multiple units allowed us to compare methods requiring different types of input values, and also to investigate the effect of using different input values for the same method. We have shown that prefiltering of genes is essential to obtain good, robust performance for several of the evaluated methods, most notably edgeR/QLF, which tends to call lowly expressed genes with many zeros significant if these are present in the data but otherwise performs well, and voom-limma, which also performs more robustly after filtering out lowly expressed genes.

We noted a large variability among the number of genes called differential with the different methods, as well as in the ability to control the type I error rate and the false discovery rate. After appropriate filtering, a subset of the methods managed to control the FDR and FPR close to the imposed level while achieving a high power while for many other methods, appropriate error control was associated with a lack of power.

We also showed that the DE methods are biased in different ways in terms of the types of genes they preferentially detect as differential, which can have important implications in practical applications. In agreement with previous evaluations, methods originally developed for bulk RNA-seq analysis did not perform worse than methods specifically developed for scRNA-seq data, but sometimes showed a stronger dependence on the data being appropriately prefiltered.

Figure 5 summarizes the performance of the different methods across the main evaluation criteria in our study. For each evaluation aspect, each method was classified as “good”, “intermediate” or “poor” (Online Methods). While it is difficult to capture the full complexity of the evaluation in a crude categorization, the table provides a convenient summary of our results and can be used to select an appropriate method based on the criteria that are most important for a specific application.

The number of cells per group ranged between 6 and 400 in our datasets. While these are relatively small numbers compared to the thousands of cells that can be sequenced in an actual experiment, DE analysis is typically used to compare sets of homogeneous cells (e.g., from given, well-defined cell types), and these collections are likely to be much smaller. Thus, we believe that the range of sample sizes considered in our comparisons are relevant for real applications and that it is important to know how the methods perform under these circumstances.

## Acknowledgements

The authors would like to acknowledge M Love and V Svensson for helpful online instructions regarding automated download of raw data from ENA. This study was supported by the Forschungskredit of the University of Zurich, grant no. FK-16-107 to C.S.

## Author Contributions

C.S. and M.D.R. designed analyses and wrote the manuscript. C.S. performed analyses. Both authors have read and approved the final manuscript.

## Competing interests

The authors declare that they have no competing interests.

## References


1. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6,** 377–382 (2009).
2. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10,** 1096–1098 (2013).
3. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161,** 1187–1201 (2015).
4. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
5. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8,** 14049 (2017).
6. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15,** 550 (2014).
7. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
8. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15,** R29 (2014).
9. Miao, Z. & Zhang, X. Differential expression analyses for single-cell RNA-Seq: old questions on new data. *Quant Biol* **4,** 243–260 (2016).
10. Jaakkola, M. K., Seyednasrollah, F., Mehmood, A. & Elo, L. L. Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief Bioinform* bbw057 (2016).

11. Lun, A. T. L. & Marioni, J. C. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics* (2017).
12. Vallejos, C. A., Richardson, S. & Marioni, J. C. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biol* **17,** 70 (2016).
13. Korthauer, K. D. *et al.* A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol* **17,** 222 (2016).
14. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33,** 495–502 (2015).
15. Lun, A. T. L., Chen, Y. & Smyth, G. K. It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. in *Statistical Genomics* (eds. Mathé, E. & Davis, S.) 391–416 (Springer New York, 2016).
16. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* **10,** 1200–1202 (2013).
17. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci U A* **107,** 9546–9551 (2010).
18. Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat Methods* **13,** 577–580 (2016).

19. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).

20. Elo, L. L., Filén, S., Lahesmaa, R. & Aittokallio, T. Reproducibility-optimized test statistic for ranking genes in microarray studies. *IEEEACM Trans Comput Biol Bioinform* **5,** 423–431 (2008).

21. Seyednasrollah, F., Rantanen, K., Jaakkola, P. & Elo, L. L. ROTS: reproducible RNA-seq biomarker detector-prognostic markers for clear cell renal cell cancer. *Nucleic Acids Res* **44,** e1 (2016).

22. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat Methods* **11,** 740–742 (2014).

23. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* **14,** 309–315 (2017).

24. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32,** 381–386 (2014).

25. Sengupta, D., Rayan, N. A., Lim, M. & Lim, B. Fast, scalable and accurate differential expression analysis for single cells. *bioRxiv* (2016). doi:10.1101/049734

26. Li, J. & Tibshirani, R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res* **22,** 519–536 (2013).

27. Finak, G. *et al.* MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA-seq data. *Genome Biol* **16,** 278 (2015).

28. Delmans, M. & Hemberg, M. Discrete distributional differential expression (D3E)–a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics* **17,** 110 (2016).

29. Smyth, G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat Appl Genet Mol Biol* **3,** Article 3 (2004).

30. Miao, Z. & Zhang, X. DEsingle: A new method for single-cell differentially expressed genes detection and classification. *bioRxiv* (2017). doi:10.1101/173997

31. Vu, T. N. *et al.* Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32,** 2128–2135 (2016).

32. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12,** 115–121 (2015).

## Figure legends

### Figure 1

Type I error control across several instances from eight single-cell null datasets, with a range of sample sizes. Values are split between full-length and UMI datasets, and the methods are ordered by the median FPR across all datasets (separately for unfiltered and prefiltered datasets). A. Without any prefiltering of genes (only excluding genes with zero counts across all cells). B. After filtering, retaining only genes with an estimated expression above 1 TPM in more than 25% of the cells. Only methods returning nominal p-values are included. The black line indicates the target FPR=0.05, and the y-axis is square-root transformed for increased visibility. Center line, median; hinges, first and third quartiles; whiskers, most extreme values within 1.5 IQR from the box; *n*, number of data set instances.

## Figure 2

Characteristics of genes falsely called significant by the evaluated methods. For each instance of the eight real scRNA-seq null datasets, we record characteristics of each gene (average CPM, variance and coefficient of variation of CPM, fraction zeros across all cells) and use a signal-to-noise statistic to compare each of these characteristics between genes called significant and the rest of the genes. A positive statistic indicates that the corresponding characteristic is more pronounced in the set of genes called significant than in the remaining genes. Note that ROTSvoom, D3E, limma-trend, the t-test and the Wilcoxon test did not return enough false positive findings to be included in the evaluation. Center line, median; hinges, first and third quartiles; whiskers, most extreme values within 1.5 IQR from the box; *n*, number of data set instances.

## Figure 3

Dendrogram illustrating the average similarities between the gene rankings obtained by the evaluated methods. The dendrogram is obtained by complete-linkage hierarchical clustering based on the matrix of average AUCC values across all datasets. The labels of the internal nodes represent their stability across datasets, in terms of the fraction of instances where they are observed. Only nodes with stability scores of at least 0.1 are labeled. The colored boxes below the methods represent characteristics of the methods.

## Figure 4

Differential expression detection performance, summarized across all instances of the three simulated datasets. The methods are stratified by their ability to control the FDR at the 0.05 level across the datasets. A method where more than 75% of the observed FDPs are above 0.05 or where the median FDP is above 0.15 is considered to have "high FDP", whereas a method where more than 75% of the observed FDPs are below 0.05 or where the median FDP is below 0.0167 is considered to have "low FDP". A-B. Observed FDP at an adjusted p-value cutoff at 0.05. The horizontal line represents the target FDR of 0.05, and the y-axis is square-root transformed for increased visibility. C-D. Observed TPR at an adjusted p-value cutoff at 0.05. E-F. Observed area under the ROC curve. Center line, median; hinges, first and third quartiles; whiskers, most extreme values within 1.5 IQR from the box; *n*, number of data set instances.

## Figure 5

Summary of the performance of the evaluated methods across all major evaluation criteria in the current study. A description of the criteria and the cutoff values for assigning a method to a performance category is available in the Online Methods. The methods are ranked by their average performance across the criteria, with the numerical encoding good=2, intermediate=1, poor=0. NODES and SAMseq do not return nominal p-values and are therefore not evaluated in terms of the FPR.

# Online Methods

## *conquer*

The *conquer* pipeline processes (sc)RNA-seq datasets using the steps outlined in Supplementary Table 3, including quality control, abundance estimation, exploratory analysis and summarization.

Many of the processed datasets contain not only scRNA-seq samples (single cells), but also bulk RNA-seq samples for comparison, or technical control samples. Whenever these could be identified, they are excluded from the processed data. A list of the excluded samples is provided in the online repository. Cells belonging to the same SRA/GEO dataset but sequenced on different platforms are separated into different repository entries. No filtering based on poor quality or low abundance is performed, since that may introduce unwanted biases for certain downstream analyses and since no universally adopted filtering approach or threshold currently exists. However, the provided quality control and exploratory analysis reports can be used to determine whether some cells need to be excluded for specific applications. The Ensembl catalog (v38)[33] was used as reference when processing the currently available datasets. Information about the underlying reference is also included as metadata in the processed datasets and displayed in the exploratory report. Since TPMs and read counts are estimated using the same reference annotation, with the same software and using the same data, the *conquer* datasets can be used to compare computational methods that require different types of input, with minimal bias. The processed datasets and the resulting reports can be browsed and downloaded from http://imlspenticton.uzh.ch:3838/conquer/, and the underlying code used to process all datasets is available from https://github.com/markrobinsonuzh/conquer.

## Evaluation of differential expression methods

### Experimental and simulated data

Seven of the real datasets from *conquer*, with a large number of cells, are selected as the basis for the evaluation of DE analysis methods. For each of the datasets, we retain only cells from two of the annotated cell groups (Supplementary Table 1), attempting to select large and relatively homogeneous populations among the ones annotated by the data generators. The selected datasets span a wide spectrum of signal strengths and population homogeneities (Supplementary Figures 1 and 2). For each dataset, we then generate one instance of "maximal" size (with the number of cells per group equal to the size of the smallest of the two selected cell populations) and several subsets with fewer cells per group by random subsampling from the maximal size subset (see Supplementary Table 1 for exact group sizes). For each non-maximal sample size, we generate five replicate dataset instances, and thus each original dataset contribute 11-21 separate instances, depending on the number of different sample sizes (Supplementary Table 1). Moreover, for each dataset with enough cells we generate *null* datasets with different sample sizes (again, five instances per

sample size except for the maximal size) by sampling randomly from one of the two selected cell populations. Finally, three of the datasets (GSE45719, GSE74596 and GSE60749-GPL13112) are used as the basis for simulation of data using a slightly modified version of the *powsim* R package[34]. Individual reports generated by *countsimQC*[35] and verifying the similarity between the simulated and real datasets across a range of aspects are provided as Supplementary Data. As for the original, experimental datasets, we subsample dataset instances with varying number of cells per group, and further generate null datasets by random sampling from one of the simulated groups. In each simulated dataset, 10% of the genes are selected to be differentially expressed between the two groups, with fold changes sampled from a Gamma distribution with shape 4 and rate 2. The direction of the DE is randomly determined for each gene, with equal probability of up- and downregulation. Mean and dispersion parameters used as basis for the simulations are estimated from the respective real datasets using edgeR[7]. For each of the three datasets, the rounded length-scaled TPMs for all genes with at least two non-zero counts are used as input to the simulator, and a dataset with the same number of genes is generated. The counts for each simulated gene are based on one of the original genes (however, the same original gene can be the basis for more than one simulated gene), and by retaining this information we can link average transcript lengths (calculated by *tximport*[36] for the original data) to each simulated gene, and thus estimate approximate TPMs also for the simulated data.

In addition to the seven datasets from *conquer*, we downloaded and processed two additional UMI datasets. First, the UMI counts corresponding to the GEO entry GSE59739[37] were downloaded from http://linnarssonlab.org/drg/ (accessed December 18, 2016). The provided UMI RPMs were used in the place of TPMs, and were combined with the provided information about the total number of reads per cell to generate gene counts. Empty wells were filtered out. Second, we downloaded UMI count matrices for C14+ monocytes and cytotoxic T-cells processed with the 10X Genomics GemCode protocol[5] (https://support.10xgenomics.com/single-cell-gene-expression/datasets, accessed September 17, 2017). For this dataset, as well as for the UMI dataset obtained from *conquer* (GSE62270-GPL17021), the UMI counts were used as "raw counts" in the DE analysis, and since these counts are supposed to be proportional to the concentration of transcript molecules, we estimated the TPM by scaling the UMI counts to sum to 1 million. Although this may be suboptimal due to the low capture efficiency of single-cell protocols, it allows us to apply methods consistently across full-length and UMI datasets.

For comparison, we also downloaded a bulk RNA-seq dataset from the Geuvadis project[19] from http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/ and estimated gene expression levels using the same pipeline as for the *conquer* datasets. For this dataset, we perform DE analysis using a subset of the methods applied to the single-cell RNAseq datasets, comparing samples from the CEU and YRI populations generated at the University of Geneva.

For each real and simulated dataset, we perform the DE analysis evaluation both on the full, "unfiltered", dataset (excluding only genes with 0 counts in all

considered cells) and on a filtered dataset, where we retain only genes with an estimated TPM above 1 in more than 25% of the considered cells. Depending on the dataset and the number of considered cells, between 4 and 50% of the genes are retained after this filtering (Supplementary Figure 32).

## Differential expression analysis methods

For each of the real and simulated scRNA-seq datasets, we apply 36 statistical approaches for DE analysis to compare the expression levels in the two groups of cells (Supplementary Table 2). As representatives for methods developed for differential analysis of bulk RNA-seq data, we include edgeR[7], DESeq2[6], voom-limma[8] and limma-trend[8]. For edgeR, we apply both the likelihood ratio test (LRT)[38] and the more recent quasi-likelihood approach (QLF)[15]. For the LRT, in addition, we use both the default dispersion estimates[39] and the robust dispersion estimates developed to address outlier counts[40], and we apply edgeR both with the default TMM normalization[41] and with the recently developed deconvolution normalization approach for scRNA-seq[42]. In addition, we run edgeR/QLF including the cellular detection rate (the fraction of detected genes per cell) as a covariate. DESeq2 is run in three modes, after rounding the length-scaled TPM values to integers: with default settings, without the log-fold change shrinkage (beta prior), and after disabling the internal independent filtering and outlier detection and replacement. Additionally, both edgeR/LRT and DESeq2 are applied to both the read counts (length-scaled TPMs as described above) and Census transcript counts[23], aimed at converting relative abundances such as TPMs into transcript counts, based on the assumption that the most common signal among the genes detectable with current single-cell library preparation protocols corresponds to a single molecule. The Census counts are calculated from the estimated TPMs using monocle[24] with default settings. We note that it is possible that modifications of these settings, optimized for the library preparation parameters for each individual dataset, would lead to different absolute count values, and thus potentially altered performance, in some of the datasets.

Three non-parametric methods are included in the comparison: SAMseq[26], the Wilcoxon test[43] and NODES[25]. SAMseq is applied to the length-scaled TPMs, while the Wilcoxon test is applied to TPM estimates after applying TMM normalization to address the compositionality of the TPMs. NODES was initially run in two modes: with default settings, and after disabling the internal filtering steps. However, disabling the internal filtering caused the method to fail in subsequent steps, and thus we retain only the runs with default settings.

We include a broad range of methods developed specifically for scRNA-seq DE analysis. BPSC[31] is applied to CPMs (calculated using edgeR) as suggested by the package authors. D3E[28] is run with the method-of-moments approach to parameter estimation, the non-parametric Cramer-von Mises test to compare distributions and without removing zeros before the analysis. MAST[27] is applied to both $\log_2$(CPM+1) and $\log_2$(TPM+1) values, both with and without including the cellular detection rate (the fraction of genes that are detected with non-zero

counts) as a covariate in the model. For monocle[24], the input is either TPM estimates (with a tobit model), raw counts (read counts or UMI counts, depending on the dataset, with a Negative Binomial model) or Census counts (with a Negative Binomial model), calculated from the TPMs as for edgeR and DESeq2 above. SCDE[22] is applied to rounded length-scaled TPMs, following the instructions provided in the package documentation, and p-values are calculated from the provided z-scores. Seurat[14] is applied using either the default "bimod" likelihood ratio test[44] (applied to the length-scaled TPMs, which are log-normalized internally), both with default settings and disabling the internal filtering steps, as well as after setting the internal expression threshold to 2 instead of the default of 0, or the "tobit" test[24] (applied to the TPMs). scDD[13] was applied to counts normalized with the median normalization, and using the default "fast" procedure based on the Kolmogorov-Smirnov test, without permutations. We applied DEsingle[30] to rounded counts.

Given the similarities between single-cell RNA-seq data and operational taxonomic unit (OTU) count data from 16S marker studies in metagenomics applications, we also apply metagenomeSeq[16] to the count values, fitting the zero-inflated log-normal model using the *fitFeatureModel* function from the metagenomeSeq package and testing for differences in abundance.

Finally, we include ROTS (reproducibility-optimized test statistic)[20,21], which is a general test, originally developed for microarray data, in which a t-like test statistic is optimized for reproducibility across bootstrap resamplings. We apply ROTS to CPM and TPM values, as well as to the log-transformed CPM values calculated by the voom function in the limma package[8]. For comparison, we also apply a Welch t-test[45] to TMM-normalized TPM values, after adding 1 and applying a log-transformation.

All code used for the DE analysis and evaluation is accessible via https://github.com/csoneson/conquer_comparison.

## Evaluation strategies

Most of the evaluations in this study are performed using real, experimental data, where no independently validated truth is available. The advantage of this approach is that no assumptions or restrictions are made regarding data distributions or specific structures of the data. However, the set of evaluation measures is more limited than in situations where the ground truth is accessible. Our first battery of evaluation approaches aim to catalog the number of genes found to be significantly differentially expressed, as well as the number and characteristics of the false positive detections from each method. For the latter evaluations we use the null datasets, where no truly differential genes are expected and thus all significant genes are false positives. First, we investigate the fraction of genes for which no interpretable test results are returned by the applied methods (e.g., due to internal filtering or convergence failure of fitting procedures). Then, for all methods returning nominal p-values, we calculate the fraction of performed tests that give a nominal p-value below 0.05. For a well-calibrated test, this fraction should be around 5%. Next, we calculate

characteristics such as the expression level (CPM), the fraction of zero counts and the expression variability (variance and coefficient of variation for CPM estimates) for all genes, and compare these characteristics between genes called differentially expressed (with an adjusted p-value/FDR threshold of 0.05) and genes not considered DE, for each of the methods. More precisely, for each characteristic and for each method detecting at least five differentially expressed genes at this threshold, we calculate a signal-to-noise statistic:

$$\frac{\mu_S - \mu_{NS}}{\sigma_S + \sigma_{NS}}$$

where $\mu_S$ ($\mu_{NS}$) and $\sigma_S$ ($\sigma_{NS}$) represent the mean and standard deviation of the gene characteristic among the significant (nonsignificant) genes. Genes with non-interpretable test results (e.g., NA adjusted p-values) are considered non-significant in this evaluation. This approach gives insights into the inherent biases of the different methods, in the sense of the type of genes that are preferentially called significantly differential. Note that since the evaluation is done on the null datasets, the results are not confounded by the characteristics of *truly* differentially expressed genes.

The second type of evaluations focus on *robustness* of methods when applied to different subsets of the same dataset. In a dataset where there is a true underlying signal (i.e., truly differential genes between cell populations), ideally, this signal will be detected regardless of the set of cells that are sampled for the analysis. Thus, a high concordance between results obtained from different subsets of the cells is positive, and indicative of robust performance. For a dataset without truly differential genes, however, any detections should be random, and a high similarity between results obtained from different subsets can rather indicate a bias in the DE calling. Thus, we first calculate a measure of concordance between the gene rankings from each pair of instances of a dataset with the same number of cells per group (five such instances were generated for each group size, giving 10 pairwise comparisons). Then, we match "signal" and null instances from the same original dataset and with the same number of cells per group, and compare the robustness values between signal and null instances. A large difference indicates a significant difference between the cross-instance concordance in a dataset with a true underlying signal and a dataset without a true signal, suggesting that the method is able to robustly detect underlying effects, and that this robustness is not due to a strong bias in the significance testing. As a measure of concordance, we use the area under the *concordance curve* for the top-$K$ genes ranked by significance, with $K$=100 (cf. Irizarry *et al*.[46]). More precisely, for each dataset instance and each DE method, we rank the genes by statistical significance (nominal p-value or adjusted p-value). Then, for each pair of dataset instances with the same sample size, for $k$=1,…,$K$, we count the number of genes that are ranked among the top $k$ in both the corresponding rankings. Plotting the number of shared genes against $k$ gives a curve, and the area under this curve is used as a measure of the concordance. To obtain more interpretable values, we divide the calculated area with the maximal possible value ($K^2/2$). Thus, a normalized value of 1 indicates that the two compared rankings are identical, whereas a value of 0 indicates that the sets of top-$K$ genes

from the two rankings don't share any genes. The rationale for using this type of concordance index to evaluate robustness is that it is independent of the number of genes that are actually called significant (which can vary widely across methods), and it is applicable to situations where not all compared rankings have interpretable results for the same sets of genes (e.g., due to different internal filtering criteria), which would cause a problem for e.g. overall correlation estimation. Furthermore, as opposed to a simple intersection of the top-$K$ genes in the two rankings, the concordance score incorporates the actual ranking of these top-$K$ genes.

A similar approach is used to evaluate similarities between methods. Briefly, for each dataset instance, we rank the genes by significance using each of the DE methods. Then, for each pair of methods, we construct a concordance curve and calculate the area under this curve as a measure of similarity between the results from the two methods. This evaluation is only performed on the "signal" datasets.

Finally, we use the simulated data to evaluate false discovery rate (FDR) control and true positive rate (TPR, power), as well as the area under the receiver operating characteristic (ROC) curve, indicating the ability of a method to rank truly differential genes ahead of truly non-differential ones. For the prefiltered datasets, we limit the evaluation to the genes retained after the filtering.

An interesting aspect, although not strictly related to performance, is the computational time requirement for the different methods. We investigate two aspects of this: first, the actual time required to run each method using a single core. Since this depends on the size of the dataset, we normalize all times for a given dataset instance so that the maximal value across all methods is 1. Thus, a "relative" computational time of 1 for a given method and a given dataset instance means that this method was the slowest one for that particular instance, and a value of, e.g., 0.1 means that the time requirement was 10% of that for the slowest method. Second, we investigate how the computational time requirement scales with the number of cells. This is particularly important for scRNA-seq data, since the number of cells sequenced per study is now increasing rapidly[47]. For this, we consider all instances of all datasets ("signal" and null, as well as simulated data), and divide them into 10 equally sized bins depending on the total number of tested genes. Within each such bin, we model the required time $T$ as a function of the number of cells per group ($N$) as

$$T=aN^p,$$

and record the estimated value of $p$.

## Performance summary criteria

Figure 5 summarizes the performance of the evaluated methods across the range of evaluation metrics. For each metric, the performance of each method is considered either "good", "intermediate" or "poor". Metrics that are mainly

descriptive rather than quantitative are excluded from the summary. Here, we list the criteria used to categorize the methods for each evaluation metric:

**MedianFDP**. Evaluated after filtering, across all simulated signal datasets

- Good: no more than 75% of FDPs on one side (above or below) of 0.05 and 0.0167 < median FDP < 0.15
- Intermediate: 0.15 ≤ median FDP < 0.25 or 0.01 < median FDP ≤ 0.0167, or 0.0167 < median FDP < 0.15 but more than 75% of FDPs on one side of 0.05
- Poor: median FDP ≥ 0.25 or median FDP ≤ 0.01

**MaxFDP**. Evaluated after filtering, across all simulated signal datasets

- Good: maximal FDP < 0.15
- Intermediate: 0.15 ≤ maximal FDP < 0.35
- Poor: maximal FDP ≥ 0.35

**TPR**. Evaluated after filtering, across all simulated signal dataset instances with more than 20 cells

- Good: median TPR > 0.8
- Intermediate: 0.6 < median TPR ≤ 0.8
- Poor: median TPR ≤ 0.6

**AUROC**. Evaluated after filtering, across all simulated signal datasets

- Good: median AUC > 0.8
- Intermediate: 0.65 < median AUC ≤ 0.8
- Poor: median AUC ≤ 0.65

**MedianFPR**. Evaluated after filtering, across all real null datasets, separately for full-length and UMI datasets

- Good: $|\log_2(\text{median FPR}/0.05)| < \log_2(1.5)$
- Intermediate: $\log_2(1.5) \leq |\log_2(\text{median FPR}/0.05)| < 2$
- Poor: $|\log_2(\text{median FPR}/0.05)| \geq 2$

**MaxFPR**. Evaluated after filtering, across all real null datasets, separately for full-length and UMI datasets

- Good: maximal FPR < 0.1
- Intermediate: 0.1 ≤ maximal FPR < 0.25
- Poor: maximal FPR ≥ 0.25

**Scalability**. Evaluated based on all datasets

- Good: median exponent in power model of timing vs number of cells < 0.5
- Intermediate: 0.5 ≤ median exponent in power model of timing vs number of cells < 1
- Poor: median exponent in power model of timing vs number of cells ≥ 1

**Speed**. Evaluated based on all datasets

- Good: median relative computation time requirement (relative to slowest method) < 0.1

- Intermediate: 0.1 ≤ median relative computation time requirement (relative to slowest method) < 0.7
- Poor: median relative computation time requirement (relative to slowest method) ≥ 0.7

**BiasDEG**. Evaluated based on all unfiltered real null datasets

- Good: No false positive genes detected, or |median SNR| < 0.5 for all four SNR statistics (for fraction of zeros, CV(CPM), $\log_2$(average CPM) and $\log_2$(variance(CPM)))
- Intermediate: |median SNR| ≥ 0.5 for at least one statistic, but |median SNR| < 1 for all four statistics
- Poor: |median SNR| ≥ 1 for at least one statistic

**Consistency**. Evaluated after filtering

- Good: The t-statistic of robustness values between signal and null datasets is > 2 for GSE60749-GPL13112 and 10XMonoCytoT, and all t-statistics are ≥ 0
- Intermediate: Any of the t-statistics for GSE60749-GPL13112 or 10XMonoCytoT is ≤ 2, but all t-statistics (across all real datasets for which both signal and datasets are available) are ≥ 0
- Poor: The t-statistic for any dataset is < 0

**ComplexDesign**

- Good: The method allows arbitrary complex (fixed) designs
- Intermediate: The method can accommodate a limited set of designs
- Poor: The method only performs two-group comparisons

**FailureRate**. Evaluated across all datasets

- Good: Average failure rate < 0.01
- Intermediate: 0.01 ≤ Average failure rate < 0.25
- Poor: Average failure rate ≥ 0.25

## Software specifications and code availability

The datasets currently available in the *conquer* repository were processed with Salmon v0.6.0-v0.8.2[48], FastQC v0.11.6.devel and MultiQC v0.8[49]. All analyses for the method evaluation were run in R v3.3[50], with Bioconductor v3.4[32], except for scDD and DEsingle, which required R 3.4 and Bioconductor v3.5. Performance indices were calculated with iCOBRA v1.2.0[51] when applicable, and results were visualized using ggplot2 v2.2.1[52]. All code used to process the datasets for *conquer* can be accessed via GitHub: https://github.com/markrobinsonuzh/conquer. The code used to perform the evaluation of the DE analysis methods is also available from GitHub: https://github.com/csoneson/conquer_comparison. The results of the evaluation can be browsed in a shiny application available at http://imlspenticton.uzh.ch:3838/scrnaseq_de_evaluation/.

## Data availability

All public datasets included in *conquer* can be downloaded from http://imlspenticton.uzh.ch:3838/conquer/. The processed abundances for the UsoskinGSE59739 dataset were downloaded from http://linnarssonlab.org/drg/ on December 18, 2016. The UMI count matrices for the 10XMonoCytoT dataset were downloaded from https://support.10xgenomics.com/single-cell-gene-expression/datasets on September 17, 2017. All processed datasets used for the evaluation (listed in Supplementary Table 1) can be downloaded as a compressed archive from the accompanying website: http://imlspenticton.uzh.ch/robinson_lab/conquer_de_comparison/. Figures 1, 2, 4 and 5 have associated source data.

## References


33. Aken, B. L. *et al.* The Ensembl Gene Annotation System. *Database* baw093 (2016).

34. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsim: Power analysis for bulk and single cell RNA-seq experiments. *bioRxiv* (2017). doi:10.1101/117150

35. Soneson, C. & Robinson, M. D. Towards unified quality verification of synthetic count data with countsimQC. *Bioinformatics* btx631-btx631 (2017). doi:10.1093/bioinformatics/btx631

36. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4,** 1521 (doi: 10.12688/f1000research.7563.1) (2015).

37. Usoskin, D. *et al.* Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat Neurosci* **18,** 145–153 (2015).

38. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40,** 4288–4297 (2012).

863 39. Chen, Y., Lun, A. T. L. & Smyth, G. K. Differential Expression Analysis of
864 Complex RNA-seq Experiments Using edgeR. in *Statistical Analysis of Next*
865 *Generation Sequencing Data* (eds. Datta, S. & Nettleton, D.) 51–74 (Springer
866 International Publishing, 2014).
867 40. Zhou, X., Lindsay, H. & Robinson, M. D. Robustly detecting differential
868 expression in RNA sequencing data using observation weights. *Nucleic Acids*
869 *Res* **42,** e91 (2014).
870 41. Robinson, M. D. & Oshlack, A. A scaling normalization method for
871 differential expression analysis of RNA-seq data. *Genome Biol* **11,** R25 (2010).
872 42. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize
873 single-cell RNA sequencing data with many zero counts. *Genome Biol* **17,** 75
874 (2016).
875 43. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1,**
876 80–83 (1945).
877 44. McDavid, A. *et al.* Data exploration, quality control and testing in single-
878 cell qPCR-based gene expression experiments. *Bioinformatics* **29,** 461–467
879 (2013).
880 45. Welch, B. A. The generalization of Student's problem when several
881 different population variances are involved. *Biometrika* **34,** 28–35 (1947).
882 46. Irizarry, R. A. *et al.* Multiple-laboratory comparison of microarray
883 platforms. *Nat Methods* **2,** 345–350 (2005).
884 47. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Moore's Law in Single
885 Cell Transcriptomics. *arXiv:1704.01379v1* (2017).

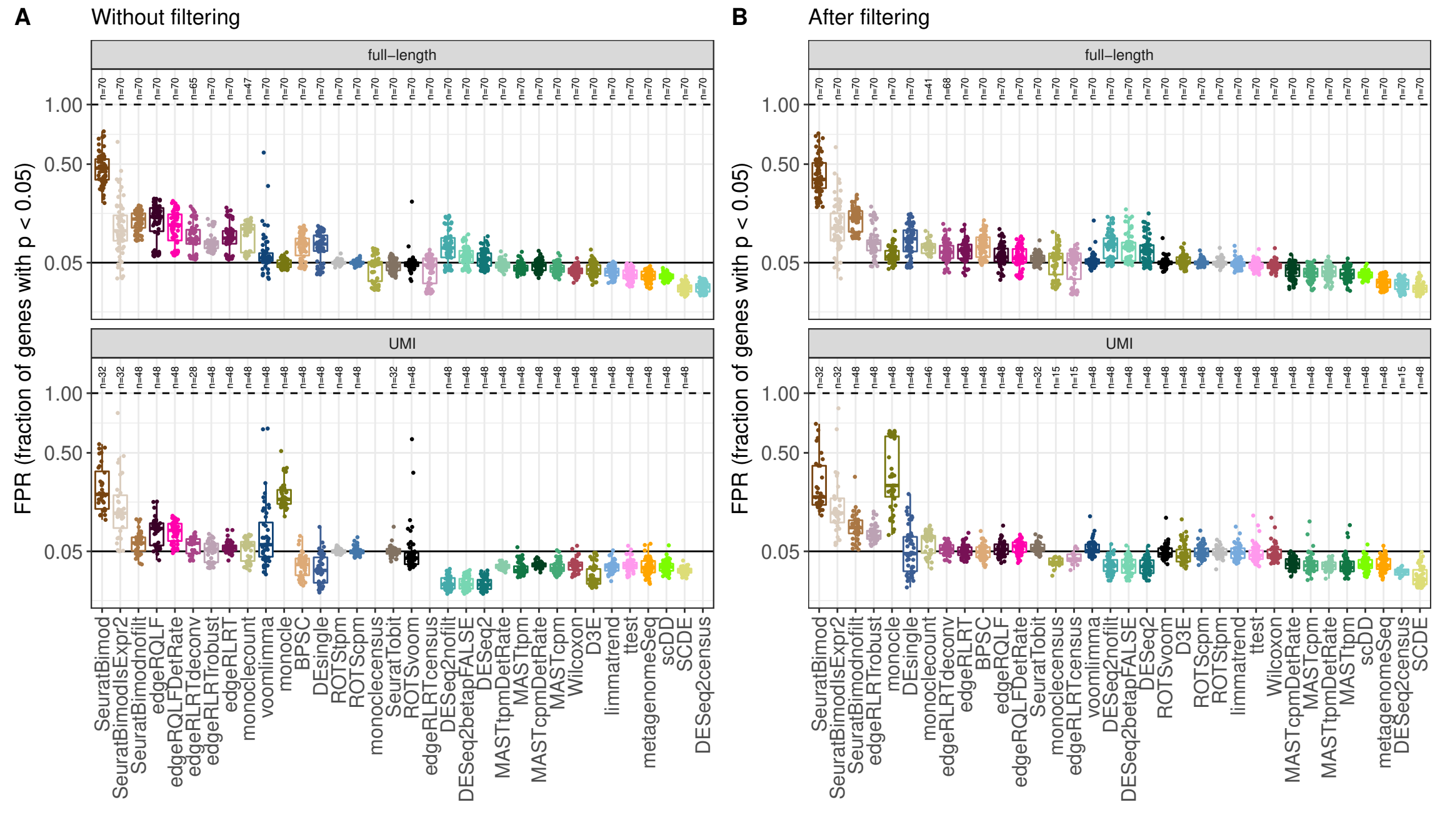
48. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14,** 417–419 (2017).

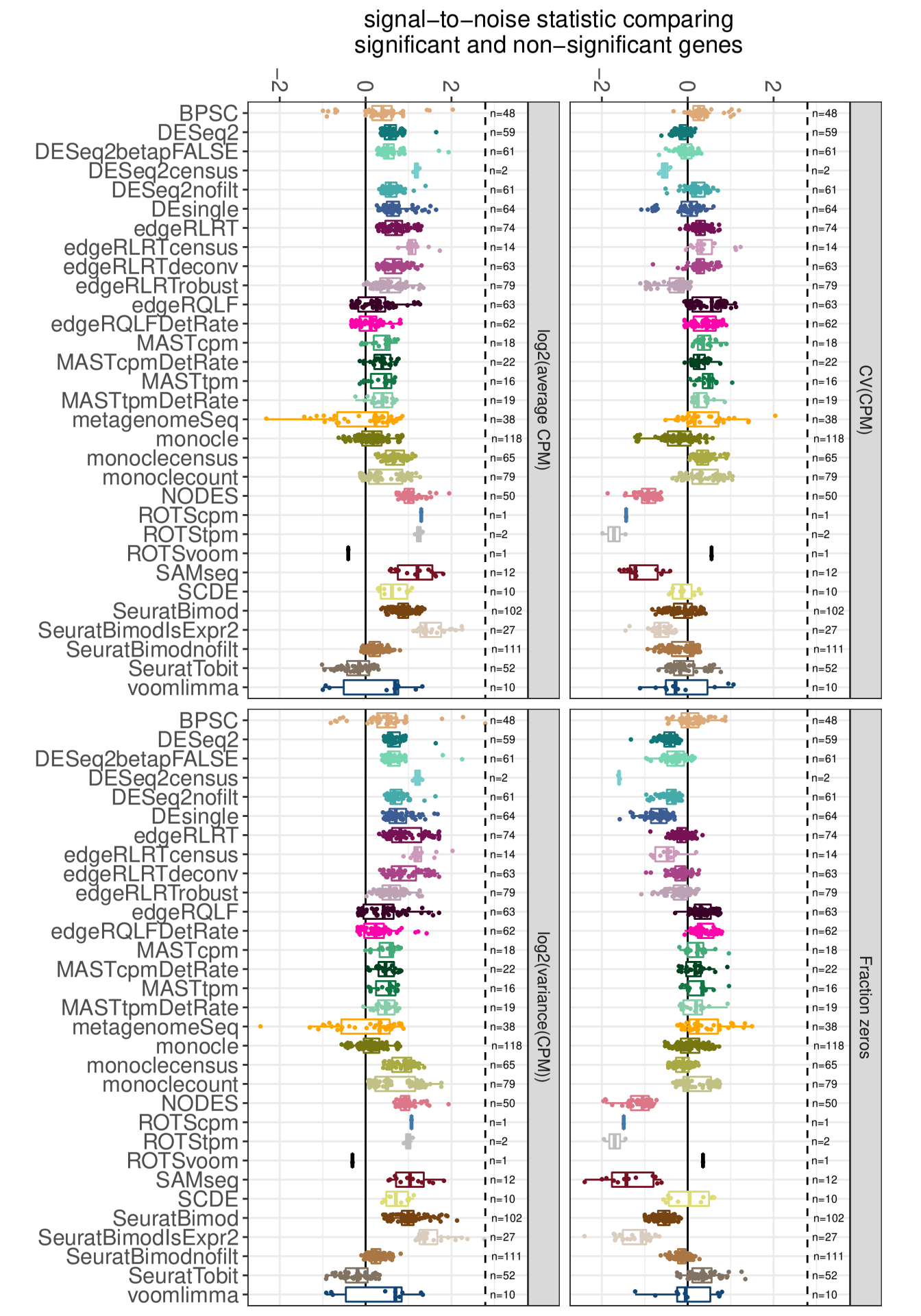
49. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* btw354 (2016).

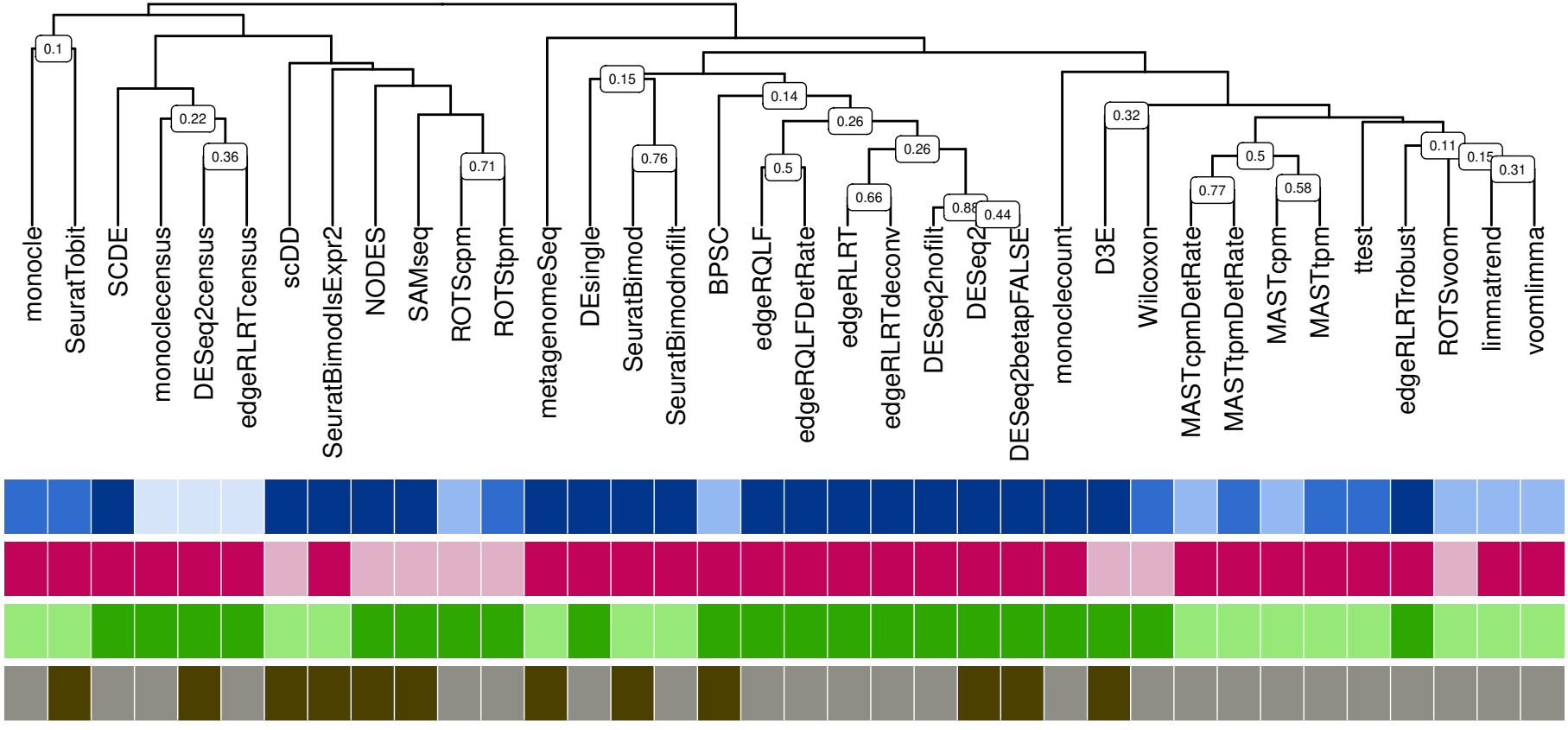
50. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2016).

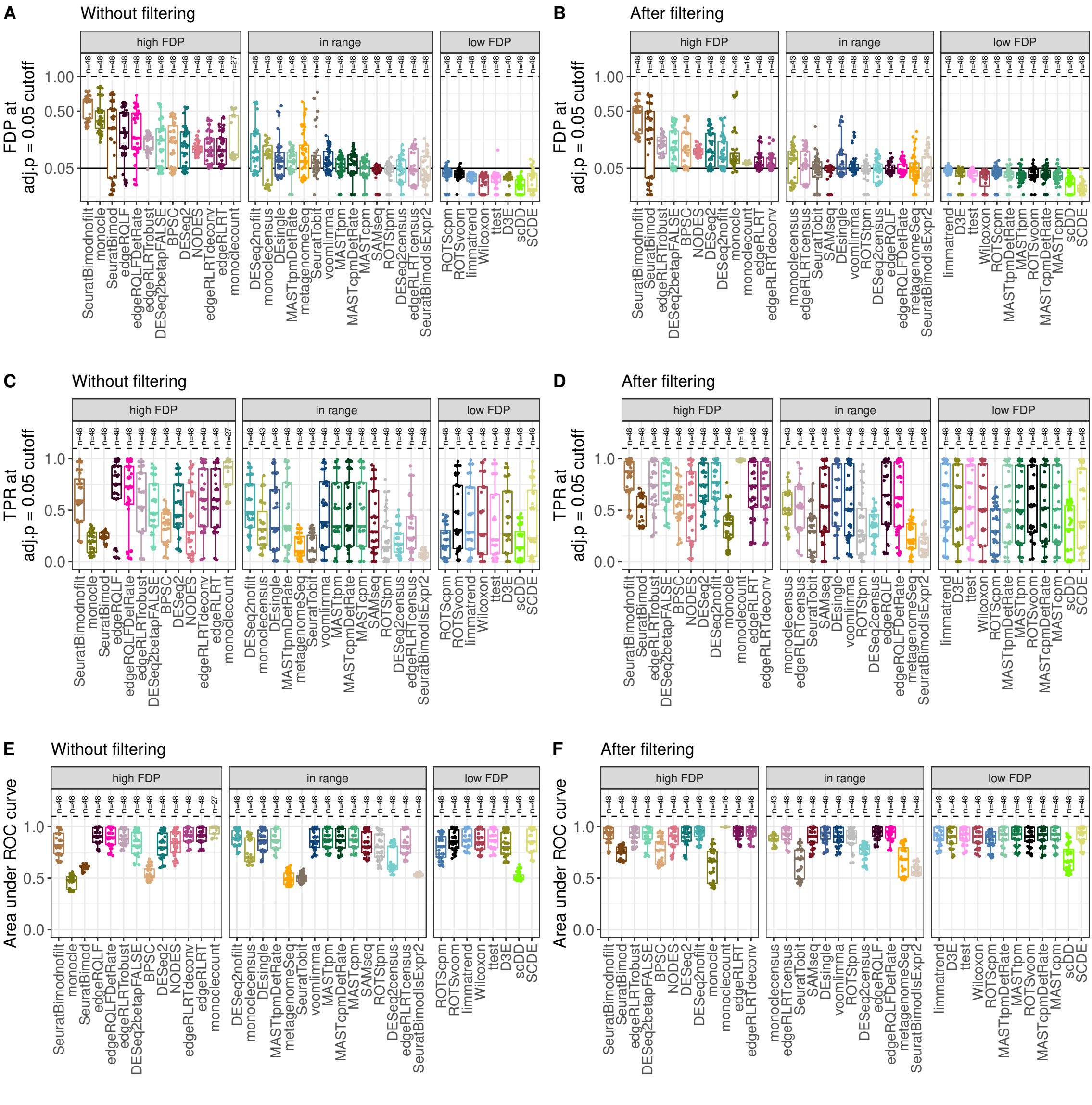
51. Soneson, C. & Robinson, M. D. iCOBRA: open, reproducible, standardized and live method benchmarking. *Nat Methods* **13,** 283 (2016).

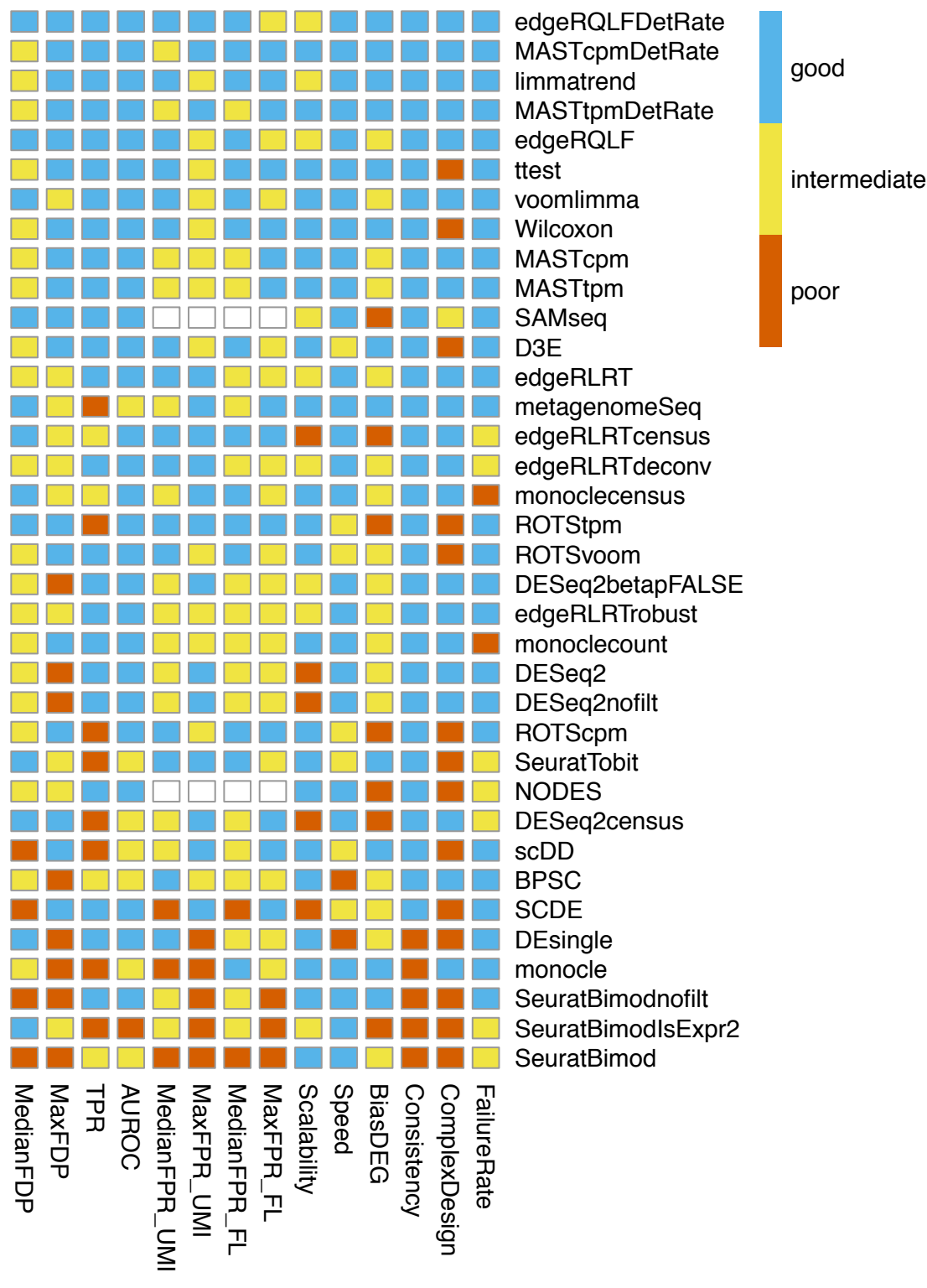
52. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2009).


A
Without filtering
B
After filtering
full−length
UMI
FPR (fraction of genes with p < 0.05)
1.00
0.50
0.05
SeuratBimod
SeuratBimodIsExpr2
SeuratBimodnofilt
edgeRQLF
edgeRQLFDetRate
edgeRLRTdeconv
edgeRLRTrobust
edgeRLRT
monoclecount
voomlimma
monocle
BPSC
DEsingle
ROTStpm
ROTScpm
monoclecensus
SeuratTobit
ROTSvoom
edgeRLRTcensus
DESeq2nofilt
DESeq2betapFALSE
DESeq2
MASTtpmDetRate
MASTtpm
MASTcpmDetRate
MASTcpm
Wilcoxon
D3E
limmatrend
ttest
metagenomeSeq
scDD
SCDE
DESeq2census


signal−to−noise statistic comparing
significant and non−significant genes
−2
0
2
BPSC
DESeq2
DESeq2betapFALSE
DESeq2census
DESeq2nofilt
DEsingle
edgeRLRT
edgeRLRTcensus
edgeRLRTdeconv
edgeRLRTrobust
edgeRQLF
edgeRQLFDetRate
MASTcpm
MASTcpmDetRate
MASTtpm
MASTtpmDetRate
metagenomeSeq
monocle
monoclecensus
monoclecount
NODES
ROTScpm
ROTStpm
ROTSvoom
SAMseq
SCDE
SeuratBimod
SeuratBimodIsExpr2
SeuratBimodnofilt
SeuratTobit
voomlimma
n=48
n=59
n=61
n=2
n=61
n=64
n=74
n=14
n=63
n=79
n=63
n=62
n=18
n=22
n=16
n=19
n=38
n=118
n=65
n=79
n=50
n=1
n=2
n=1
n=12
n=10
n=102
n=27
n=111
n=52
n=10
log2(average CPM)
CV(CPM)
log2(variance(CPM))
Fraction zeros


0.1
0.22
0.36
0.71
0.15
0.14
0.26
0.26
0.5
0.66
0.88
0.44
0.32
0.5
0.77
0.58
0.11
0.15
0.31
0.76
monocle
SeuratTobit
SCDE
monoclecensus
DESeq2census
edgeRLRTcensus
scDD
SeuratBimodIsExpr2
NODES
SAMseq
ROTScpm
ROTStpm
metagenomeSeq
DEsingle
SeuratBimod
SeuratBimodnofilt
BPSC
edgeRQLF
edgeRQLFDetRate
edgeRLRT
edgeRLRTdeconv
DESeq2nofilt
DESeq2
DESeq2betapFALSE
monoclecount
D3E
Wilcoxon
MASTcpmDetRate
MASTtpmDetRate
MASTcpm
MASTtpm
ttest
edgeRLRTrobust
ROTSvoom
limmatrend
voomlimma
Input
Census
CPM
counts
TPM
Modeling
nonparametric
parametric
Transformation
log
no
NA values
no
yes


A
Without filtering
B
After filtering
C
Without filtering
D
After filtering
E
Without filtering
F
After filtering
high FDP
in range
low FDP
FDP at adj.p = 0.05 cutoff
TPR at adj.p = 0.05 cutoff
Area under ROC curve
1.00
0.50
0.05
1.0
0.5
0.0
SeuratBimodnofilt
monocle
SeuratBimod
edgeRQLF
edgeRQLFDetRate
edgeRLRTrobust
DESeq2betapFALSE
BPSC
DESeq2
NODES
edgeRLRTdeconv
edgeRLRT
monoclecount
DESeq2nofilt
monoclecensus
DEsingle
MASTtpmDetRate
metagenomeSeq
SeuratTobit
voomlimma
MASTtpm
MASTcpmDetRate
MASTcpm
SAMseq
ROTStpm
DESeq2census
edgeRLRTcensus
SeuratBimodIsExpr2
ROTScpm
ROTSvoom
limmatrend
Wilcoxon
ttest
D3E
scDD
SCDE


edgeRQLFDetRate
MASTcpmDetRate
limmatrend
MASTtpmDetRate
edgeRQLF
ttest
voomlimma
Wilcoxon
MASTcpm
MASTtpm
SAMseq
D3E
edgeRLRT
metagenomeSeq
edgeRLRTcensus
edgeRLRTdeconv
monoclecensus
ROTStpm
ROTSvoom
DESeq2betapFALSE
edgeRLRTrobust
monoclecount
DESeq2
DESeq2nofilt
ROTScpm
SeuratTobit
NODES
DESeq2census
scDD
BPSC
SCDE
DEsingle
monocle
SeuratBimodnofilt
SeuratBimodIsExpr2
SeuratBimod
good
intermediate
poor
MedianFDP
MaxFDP
TPR
AUROC
MedianFPR_UMI
MaxFPR_UMI
MedianFPR_FL
MaxFPR_FL
Scalability
Speed
BiasDEG
Consistency
ComplexDesign
FailureRate