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

Charlotte Soneson[1,2] & Mark D Robinson[1,2]

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.

Recent advances have enabled the preparation of sequencing libraries from the transcriptomes of individual cells, making it possible to assess different cell states in tissues at high resolution[1–5]. The number of scRNA-seq data sets is increasing, but, given that the aims of different studies vary widely, public data sets are often processed using very different pipelines. Furthermore, transcript and gene abundances may be represented in different units, and a fraction of the cells and/or genes can be filtered out. This can make the reuse of preprocessed public data sets challenging, particularly comparisons across data sets. To simplify this aspect, we developed conquer, a collection of consistently processed, analysis-ready public scRNA-seq data sets. Each data set 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 data set is suitable for their purposes.

One of the most commonly performed tasks for RNA-seq data is differential gene expression (DE) analysis. Although well-established tools exist for such analysis in bulk RNA-seq data[6–8], methods for scRNA-seq data are just emerging. Given 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 that have been developed for bulk RNA-seq are suitable also for scRNA-seq. A few recent studies suggest that the optimal method may depend on the number of cells and strength of the signal[9] and that methods not initially developed for scRNA-seq analysis can perform well[10]. In this study, we used processed data sets, from conquer and other sources, to evaluate DE methods in scRNA-seq data. Our study expands the number of methods and range of experimental data sets assessed in previous comparisons and includes evaluations based on simulated data. We also investigated the effect of filtering out lowly expressed genes and extended the set of evaluation criteria.

We focused on contrasting two predefined groups of cells, as this setup can be accommodated by all of the considered methods. However, it should be noted that some scRNA-seq data sets 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 cannot 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 data sets: 31 generated with full-length protocols and 5 with 3′-end sequencing (UMI) protocols. We envision that conquer can be useful for a range of applications. Conquer's consistently processed and represented data sets can lower the barriers for evaluating computational methods, for developers as well as end-users, and for teaching and tutorial construction. In addition, conquer can be used to explore the generality of biological hypotheses across data sets from different species and cell types.

Seven data sets from conquer (six full-length and one UMI data set) and two additional UMI count data sets were used for our DE method evaluations (**Supplementary Table 1** and **Supplementary Figs. 1** and **2**). Using two predefined groups of cells from each data set, we generated multiple data set instances with varying numbers of cells. For eight data sets, we generated null data sets (no differential expression expected) by subsampling from a single group. Three data sets were used to simulate data sets with signal (10% of the genes differentially expressed) as well as null data sets. For each instance, we applied 36 DE approaches

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [2]SIB Swiss Institute of Bioinformatics, Zurich, Switzerland. Correspondence should be addressed to C.S. (charlotte.soneson@uzh.ch) or M.D.R. (mark.robinson@imls.uzh.ch).

RECEIVED 6 JUNE 2017; ACCEPTED 16 JANUARY 2018; PUBLISHED ONLINE 26 FEBRUARY 2018; DOI:10.1038/NMETH.4612

(**Supplementary Table 2**). Some methods failed to run for certain data sets (**Supplementary Fig. 3**), and these combinations were excluded from the evaluations.

Number of differentially expressed and non-tested genes

Using all instances of the nine 'signal' scRNA-seq data sets, we compared the number of differentially expressed genes called by the different methods at an adjusted *P* value cutoff of 0.05 (**Supplementary Figs. 4–7**). For full-length data sets, SeuratBimod[14] (without the default internal filtering) detected the largest number of significant genes. edgeR/QLF[7,15] detected many genes if the data set was not prefiltered to remove lowly expressed genes, but showed the largest decrease in the number of significant genes after filtering (**Supplementary Fig. 8**). Conversely, SeuratBimod with nonzero expression threshold, metagenomeSeq[16] and scDD[13] consistently detected few genes. For UMI data sets, the performance of the methods based on the voom transformation[8] was highly variable without gene prefiltering.

Many DE methods implement internal filtering, which means that not all of the quantified genes are actually 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. Although most evaluated methods reported valid results for all genes, some indeed excluded many genes if run with default settings (**Supplementary Figs. 9** and **10**). This is, however, not specific to scRNA-seq data, and similar patterns were seen if a subset of the methods were applied to a large bulk RNA-seq data set[19] (**Supplementary Fig. 11**). If the data sets were filtered before the DE analysis, the fraction of non-reported results decreased, indicating that they mostly correspond to lowly expressed genes.

Type I error control

Using the eight real null data sets, where no truly differential genes are expected, we evaluated the type I error control by recording the fraction of tested genes that were assigned a nominal *P* value of less than 0.05 (**Fig. 1a**). For unfiltered data sets, many methods struggled to correctly control the type I error, and the best performance was obtained by ROTS[20,21] and SeuratTobit. Several of the other methods were too liberal, with SeuratBimod and edgeR/QLF standing out with a large number of false-positive findings. Setting a nonzero expression threshold in Seurat (SeuratBimodIsExpr2) improved the error control, but at the cost of detecting many fewer significant genes (**Supplementary Figs. 4–7**). Conversely, metagenomeSeq, scDD, SCDE[22] and DESeq2 (ref. 6) on Census counts[23] controlled the false-positive rate well below the imposed level. Methods based on voom mostly performed well, but sometimes the number of false positives was very high (**Supplementary Fig. 12**). For UMI data sets, monocle[24] performed best when applied to transcript counts (monoclecount), whereas converting these values to transcripts per million (TPMs) and applying a tobit model (monocle) led to a deterioration in performance. For full-length data sets, however, the TPM values led to a slightly better performance than the read counts. After filtering out lowly expressed genes (**Fig. 1b**) the performance of voom-limma, ROTSvoom and edgeR/QLF stabilized and improved, along with most other methods, whereas SeuratBimod still

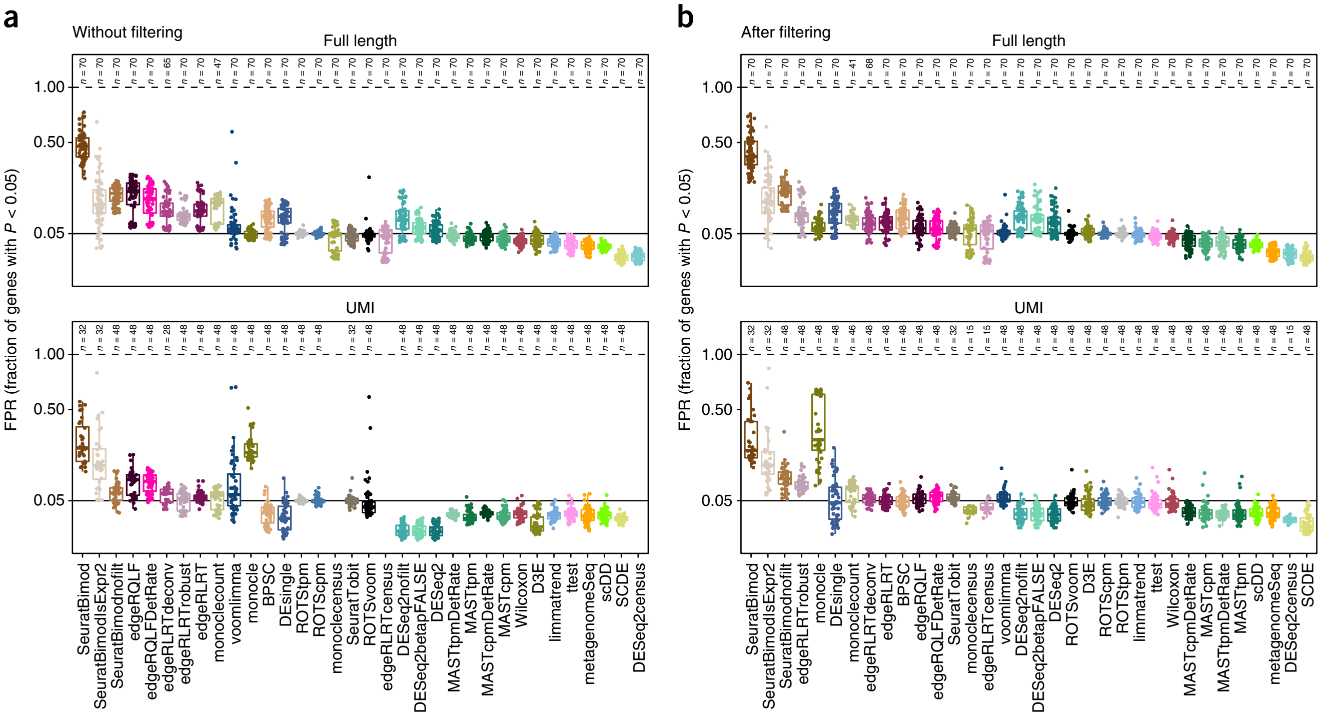

Figure 1 | Type I error control across several instances from eight single-cell null data sets. Values are split between full-length and UMI data sets, and methods are ordered by median FPR across all data sets. (**a**) Without prefiltering of genes (only excluding genes with zero counts across all cells). (**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 were 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 interquartile range (IQR) from the box; *n*, number of data set instances.

assigned low *P* values to a large fraction of the tested genes. *P* value histograms further illustrated that, without filtering, few methods returned uniformly distributed *P* values, whereas the results were considerably improved after the applied filtering (**Supplementary Figs. 13** and **14**). The results were largely similar for the three simulated data sets (**Supplementary Fig. 15**).

Characteristics of false-positive genes

To investigate biases in DE calling, we used the eight unfiltered real null data sets to characterize the set of genes that were (falsely) called significant by the different methods. For each gene in each data set instance, we estimated the average, variance and coefficient of variation of the counts per million (CPM) values across all cells as well as the fraction of cells in which the gene was undetected. For each instance, and for each method calling at least five genes DE, we calculated 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 (**Fig. 2** and **Supplementary Fig. 16**). We observed marked differences between the types of genes detected by the different methods. False positives of NODES[25], ROTS, SAMseq[26] and SeuratBimod had few zeros, high expression and mostly a relatively low coefficient of variation. Conversely, false positives of edgeR/QLF, SeuratTobit, MAST[27] and metagenomeSeq had relatively many zeros. The same evaluation performed on the simulated data sets showed largely similar results (**Supplementary Fig. 17**).

Between-method similarity

Using the nine real scRNA-seq 'signal' data sets, we quantified the concordance between gene rankings returned by different methods (for within-method consistency, see **Supplementary Fig. 18**). For each data set, 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 data sets and clustering based on the resulting similarities (**Fig. 3**) revealed, for example, that although the four MAST modes gave overall similar rankings, the inclusion of the detection rate as a covariate had 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 clustered together, as did some of the general non-parametric methods (the Wilcoxon test and D3E[28]), which were 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 gave similar rankings. The degree of similarity between any given pair of methods varied widely across the data set instances (**Supplementary Fig. 19**), but for most method pairs, it was somewhat positively associated with the number of cells per group (**Supplementary Fig. 20**).

FDR control and power

Using the simulated data sets, we evaluated the false discovery rate (FDR) control and statistical power of the methods. Several methods, such as voom/limma, ROTStpm, MAST, the methods

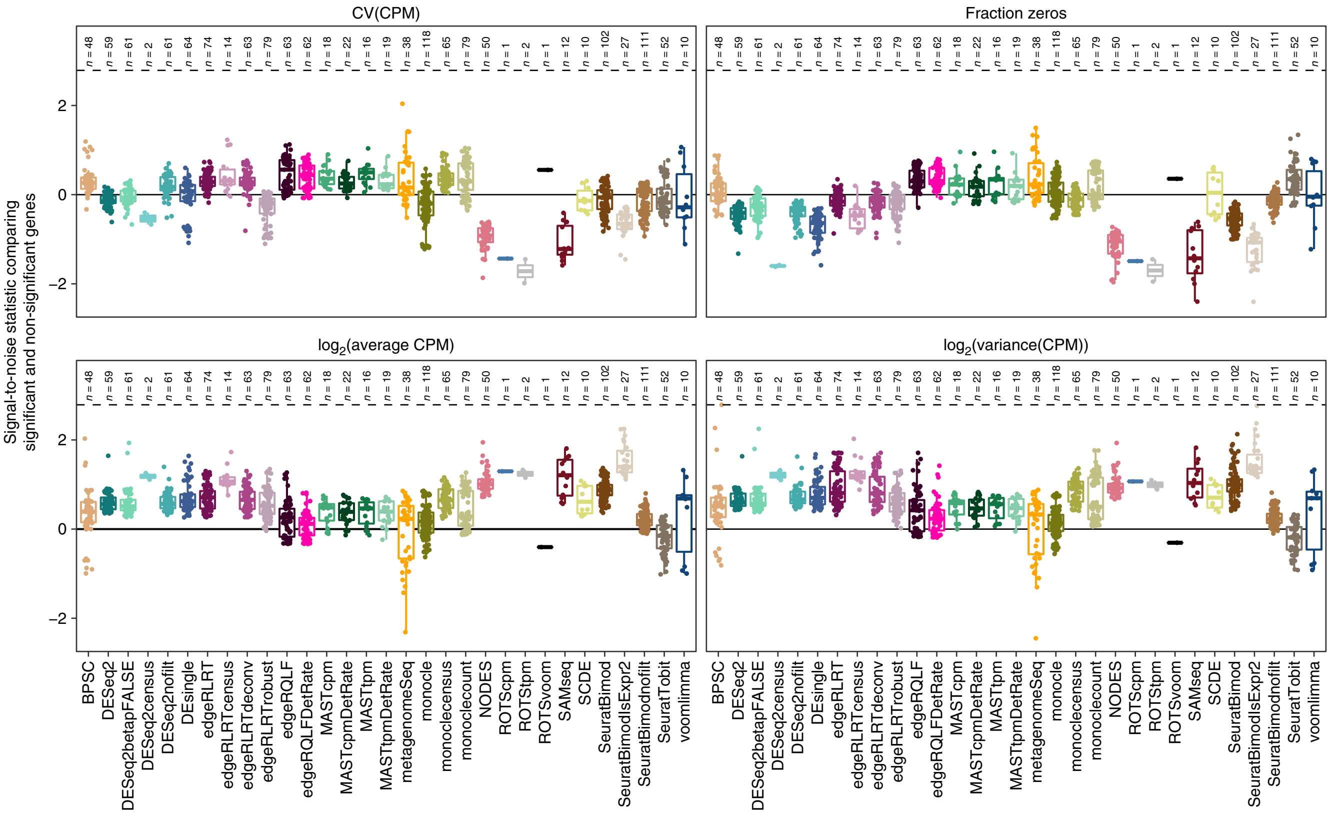

Figure 2 | Characteristics of genes falsely called significant by DE methods. A signal-to-noise statistic compares average CPM, variance and coefficient of variation of CPM, and fraction of zeros across all cells between genes called significant and all other genes for each instance of eight real scRNA-seq null data sets. A positive statistic indicates that the corresponding characteristic is more pronounced in the set of genes called significant. ROTSvoom, D3E, limma-trend, the *t* test and the Wilcoxon test did not return enough false-positive findings to be included. 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.

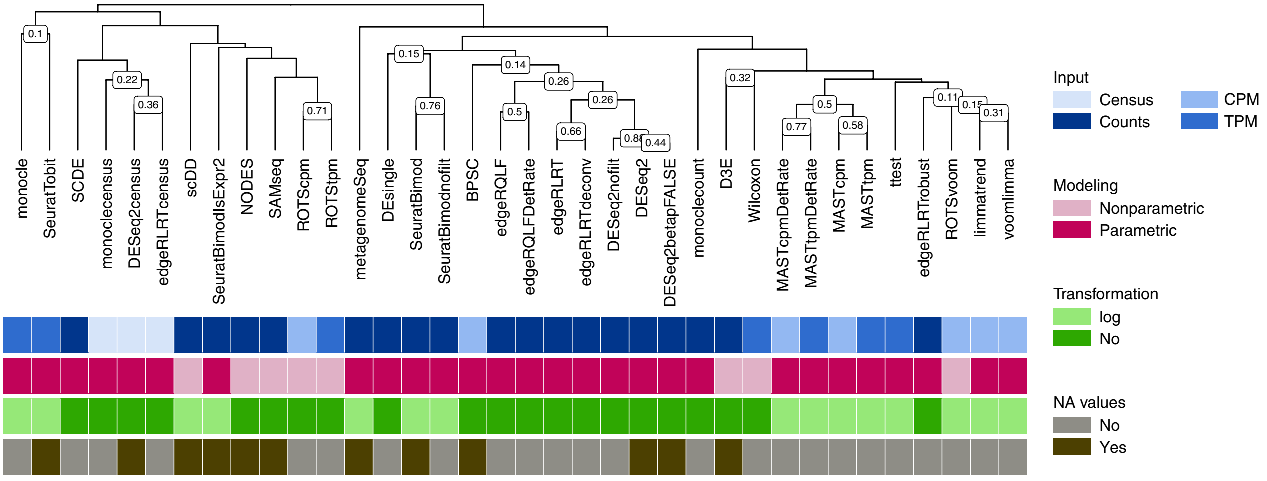

Figure 3 | Average similarities between gene rankings obtained by the evaluated DE methods. The dendrogram was obtained by complete-linkage hierarchical clustering based on the matrix of average AUCC values across all data sets. The labels of the internal nodes represent their stability across data sets (fraction of instances where they are observed). Only nodes with stability scores of at least 0.1 are labeled. Colored boxes represent method characteristics.

applied to Census counts, SeuratTobit, SeuratBimod with nonzero expression cutoff and SAMseq robustly controlled the FDR close to the imposed level (**Fig. 4a**). SCDE, scDD, the *t* test, D3E, limma-trend[8,29], the Wilcoxon test and the other variants of ROTS controlled the FDR at a lower level than imposed. The worst FDR control for the unfiltered data was obtained by monocle, SeuratBimod and edgeR/QLF. After filtering, edgeR/QLF improved markedly (**Fig. 4b**), whereas MAST and SCDE yielded even lower false discovery proportions (FDPs). Most methods performed closer to the optimal level for large sample sizes (**Supplementary Fig. 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 had only a minor effect (**Supplementary Fig. 22**).

Practically all of the methods showed increased power with increased sample size (**Fig. 4c,d** and **Supplementary Fig. 23**). Among the methods with good, robust FDR control after filtering, edgeR/QLF, SAMseq, DEsingle[30] and voom-limma achieved high power, whereas for methods such as metagenomeSeq, SeuratTobit, SeuratBimodIsExpr2 and the methods applied to Census counts, the FDR control came at the cost of reduced power. The power to detect true differences was weakly related to the fraction of genes that were excluded by internal filtering procedures (**Supplementary Fig. 24**). However, DESeq2 and NODES achieved high power despite strong filtering. The area under the ROC curve (AUROC), which indicates whether the methods are able to rank truly differentially expressed genes ahead of truly non-differential ones, showed favorable performance of edgeR, followed by MAST, limma (voom and trend), SCDE, DEsingle, DESeq2 and SeuratBimod without filtering and the non-parametric methods (**Fig. 4e**). After prefiltering, the rankings of most methods were improved (**Fig. 4f**) and the AUROC was typically higher for data sets with more cells (**Supplementary Fig. 25**).

Other aspects

As the number of cells that are studied in a data set increases, computational efficiency becomes important for method selection. For comparative purposes, we ran all of the 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, were the slowest for most data sets, whereas the bulk methods (edgeR, DESeq2 and especially the limma variants) were generally faster (**Supplementary Fig. 26a**). Most single-cell methods (with the exception of SCDE) scaled well with increasing number of cells, whereas the computational time required for the bulk RNA-seq methods was more sample size dependent (**Supplementary Figs. 26b** and **27–31**).

Although the evaluations in this study were centered on the simplest experimental situation, comparing two groups of cells, many real studies require a more complex experimental design, which not all of the 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, whereas 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,

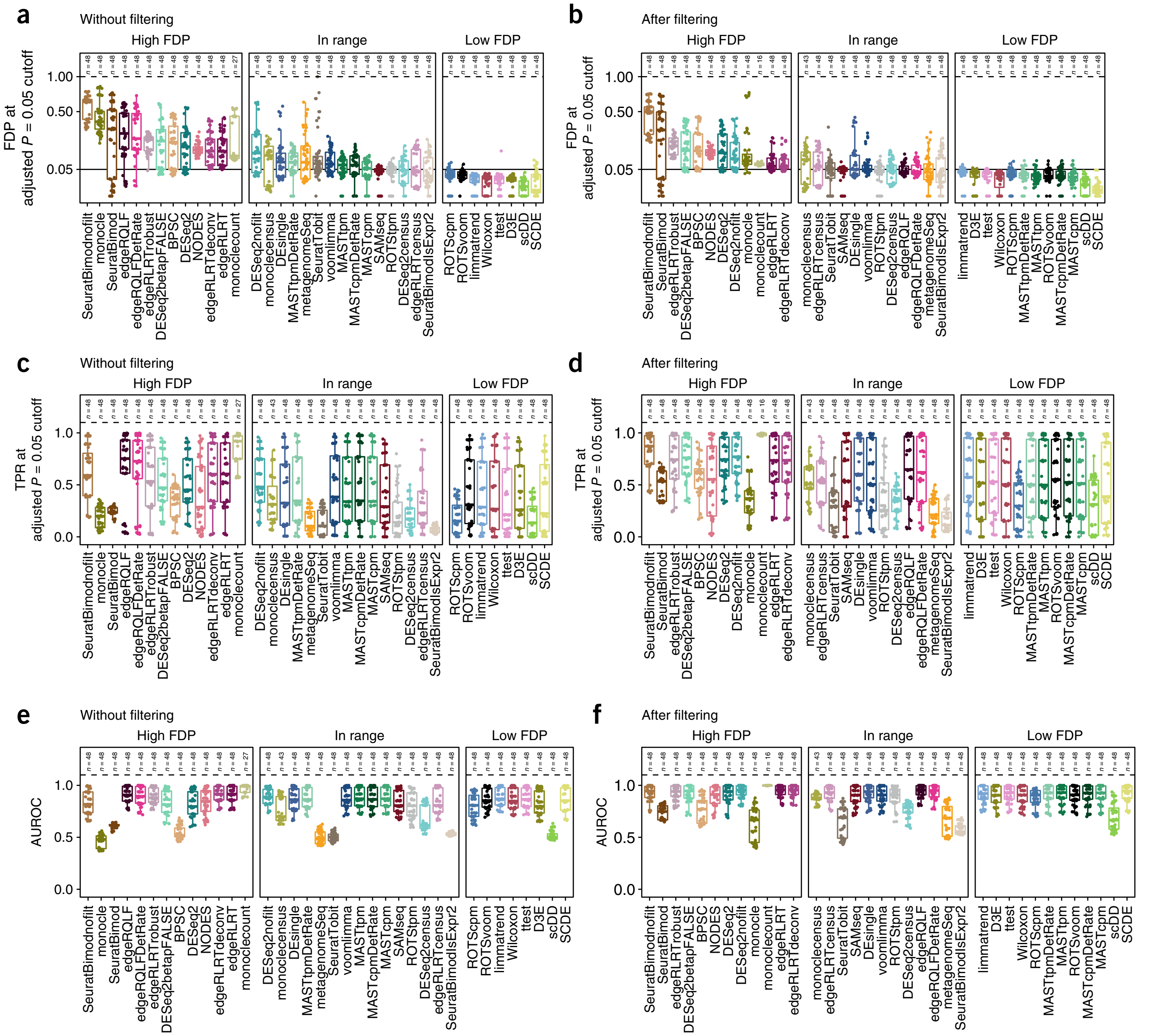

Figure 4 | Differential expression detection performance, summarized across all instances of the three simulated data sets. Methods were stratified by their ability to control the FDR at 0.05 across data sets (high FDP, >75% of observed FDPs are >0.05 or median FDP is >0.15; low FDP, >75% of observed FDPs <0.05 or median FDP is <0.0167). (**a**,**b**) Observed FDP at an adjusted *P* value cutoff of 0.05. Horizontal line represents the target FDR of 0.05 and *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 AUROC. 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.

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 leveraged consistently processed public data in conquer to carry out an extensive comparison of methods for DE analysis of scRNA-seq data. The fact that conquer provides gene expression estimates in multiple units allowed us to compare methods requiring different types of input and to investigate the effect of using different input values for the same method. We found that prefiltering of genes is essential for obtaining good, robust performance for several of the evaluated methods. Most notably, edgeR/QLF tended to call lowly expressed genes with many zeros significant if these were present in the data, but otherwise performed well, and voom-limma also performed more robustly after filtering out lowly expressed genes.

The DE methods that we tested varied greatly in the number of genes called differential, as well as in the ability to control the type I error rate and the FDR. 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, whereas appropriate error control was associated with a lack of power for many other methods.

We also found that some DE methods were biased in the types of genes that they preferentially detected as differential, which can

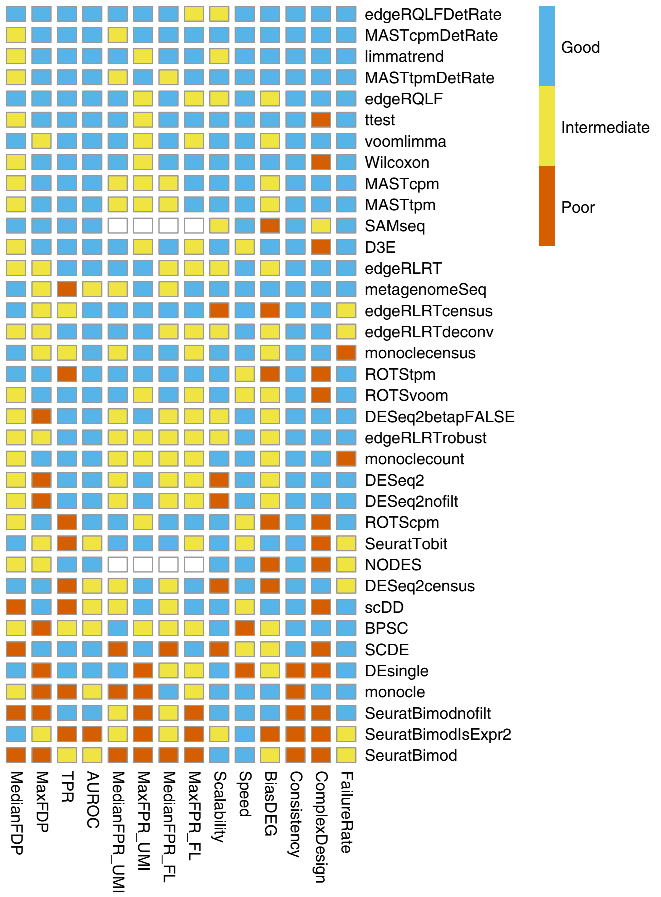

Figure 5 | Summary of DE method performance across all major evaluation criteria. Criteria and cutoff values for performance categories are available in the Online Methods. Methods are ranked by their average performance across the criteria, with the numerical encoding good = 2, intermediate = 1, poor = 0. NODES and SAMseq do not return nominal *P* values and were therefore not evaluated in terms of the FPR.

have important implications in practical applications. In agreement with previous evaluations, methods developed for bulk RNA-seq analysis did not perform worse than those specifically developed for scRNA-seq data, but sometimes showed a stronger dependence on data prefiltering.

Figure 5 summarizes performance across the main evaluation criteria in our study. For each evaluation aspect, each method was classified as 'good', 'intermediate' or 'poor' (Online Methods). Although 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 data sets. Although these are relatively small numbers compared with the thousands of cells that can be sequenced in an actual experiment, DE analysis is typically used to compare sets of homogeneous cells (for example, 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.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors 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 read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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.* **18**, 735–743 (2017).
11. Lun, A.T.L. & Marioni, J.C. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics* **18**, 451–464 (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. USA* **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. *IEEE/ACM 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. & Prabhakar, S. Fast, scalable and accurate differential expression analysis for single cells. Preprint available at https://www.biorxiv.org/content/early/2016/04/22/049734 (2016).
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 sequencing 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**, e3 (2004).
30. Miao, Z. & Zhang, X. DEsingle: a new method for single-cell differentially expressed genes detection and classification. Preprint available at https://www.biorxiv.org/content/early/2017/09/08/173997 (2017).
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).

ONLINE METHODS

Conquer. The conquer pipeline processes scRNA-seq data sets using the steps outlined in **Supplementary Table 3**, including quality control, abundance estimation, exploratory analysis and summarization.

Many of the processed data sets 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 data set 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 data sets. Information about the underlying reference is also included as metadata in the processed data sets 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 data sets can be used to compare computational methods that require different types of input, with minimal bias. The processed data sets and the resulting reports can be browsed and downloaded from http://imlspenticton.uzh.ch:3838/conquer/, and the underlying code used to process all data sets is available from https://github.com/markrobinsonuzh/conquer.

Evaluation of differential expression methods. *Experimental and simulated data.* Seven of the real data sets from conquer, with a large number of cells, are selected as the basis for the evaluation of DE analysis methods. For each of the data sets, 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 data sets span a wide spectrum of signal strengths and population homogeneities (**Supplementary Figs. 1** and **2**). For each data set, 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 data set instances, and thus each original data set contribute 11–21 separate instances, depending on the number of different sample sizes (**Supplementary Table 1**). Moreover, for each data set with enough cells we generate null data sets 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 data sets (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 data sets across a range of aspects are provided as **Supplementary Data 1**. As for the original, experimental data sets, we subsample data set instances with varying number of cells per group, and further generate null data sets by random sampling from one of the simulated groups. In each simulated data set, 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 data sets using edgeR[7]. For each of the three data sets, the rounded length-scaled TPMs for all genes with at least two nonzero counts are used as input to the simulator, and a data set 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 data sets from conquer, we downloaded and processed two additional UMI data sets. First, the UMI counts corresponding to the GEO entry GSE59739 (ref. 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 data set, as well as for the UMI data set 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 data sets.

For comparison, we also downloaded a bulk RNA-seq data set 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 data sets. For this data set, we perform DE analysis using a subset of the methods applied to the single-cell RNAseq data sets, comparing samples from the CEU and YRI populations generated at the University of Geneva.

For each real and simulated data set, we perform the DE analysis evaluation both on the full, 'unfiltered', data set (excluding only genes with 0 counts in all considered cells) and on a filtered data set, where we retain only genes with an estimated TPM above 1 in more than 25% of the considered cells. Depending on the data set and the number of considered cells, between 4 and 50% of the genes are retained after this filtering (**Supplementary Fig. 32**).

Differential expression analysis methods. For each of the real and simulated scRNA-seq data sets, 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

doi:10.1038/nmeth.4612

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. In addition, 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 data set, would lead to different absolute count values, and thus potentially altered performance, in some of the data sets.

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 nonzero 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 data set, 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 data sets, 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 (for example, 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 μ_S (μ_{NS}) and σ_S (σ_{NS}) represent the mean and s.d. of the gene characteristic among the significant (nonsignificant) genes. Genes with non-interpretable test results (for example, 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 data sets, 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 data set. In a data set

doi:10.1038/nmeth.4612

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 data set 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 data set 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 data set 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 data set with a true underlying signal and a data set 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 (ref. 46). More precisely, for each data set instance and each DE method, we rank the genes by statistical significance (nominal P value or adjusted P value). Then, for each pair of data set instances with the same sample size, for $k = 1,\ldots,K$, 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 (for example, due to different internal filtering criteria), which would cause a problem for example, 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 data set 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' data sets.

Finally, we use the simulated data to evaluate 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 data sets, 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 data set, we normalize all times for a given data set 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 data set instance means that this method was the slowest one for that particular instance, and a value of, for example, 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 data sets ('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 data sets
-Good: no more than 75% of FDPs on one side (above or below) of 0.05 and $0.0167 <$ median FDP < 0.15
-Intermediate: $0.15 \leq$ 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 data sets
-Good: maximal FDP < 0.15
-Intermediate: $0.15 \leq$ maximal FDP < 0.35
-Poor: maximal FDP ≥ 0.35
-*TPR*. Evaluated after filtering, across all simulated signal data set 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 data sets
-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 data sets, separately for full-length and UMI data sets
-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 data sets, separately for full-length and UMI data sets
-Good: maximal FPR < 0.1
-Intermediate: $0.1 \leq$ maximal FPR < 0.25
-Poor: maximal FPR ≥ 0.25

doi:10.1038/nmeth.4612

Scalability. Evaluated based on all data sets
-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 data sets
-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 data sets
-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 data sets 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 data sets for which both signal and data sets are available) are ≥ 0
-Poor: The *t* statistic for any data set 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 data sets
-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 data sets currently available in the conquer repository were processed with Salmon v0.6.0-v0.8.2 (ref. 48), FastQC v0.11.6.devel and MultiQC v0.8 (ref. 49). All analyses for the method evaluation were run in R v3.3 (ref. 50), with Bioconductor v3.4 (ref. 32), except for scDD and DEsingle, which required R 3.4 and Bioconductor v3.5. Performance indices were calculated with iCOBRA v1.2.0 (ref. 51) when applicable, and results were visualized using ggplot2 v2.2.1 (ref. 52). All code used to process the data sets 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/. A snapshot of the two code repositories is available as **Supplementary Software**.

Life Sciences Reporting Summary. Further information on experimental design is available in the **Life Sciences Reporting Summary**.

Data availability. All public data sets included in conquer can be downloaded from http://imlspenticton.uzh.ch:3838/conquer/. The processed abundances for the UsoskinGSE59739 data set were downloaded from http://linnarssonlab.org/drg/ on December 18, 2016. The UMI count matrices for the 10XMonoCytoT data set were downloaded from https://support.10xgenomics.com/single-cell-gene-expression/datasets on September 17, 2017. All processed data sets 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.

33. Aken, B.L. *et al.* The Ensembl gene annotation system. *Database* **2016**, baw093 (2016).
34. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimR: power analysis for bulk and single cell RNA-seq experiments. Preprint available at https://www.biorxiv.org/content/early/2017/06/26/117150 (2017).
35. Soneson, C. & Robinson, M.D. Towards unified quality verification of synthetic count data with countsimQC. *Bioinformatics* https://dx.doi.org/10.1093/bioinformatics/btx631 (2017).
36. Soneson, C., Love, M.I. & Robinson, M.D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (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).
39. Chen, Y., Lun, A.T.L. & Smyth, G.K. Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR. in *Statistical Analysis of Next Generation Sequencing Data* (eds. Datta, S. & Nettleton, D.) 51–74 (Springer International Publishing, 2014).
40. Zhou, X., Lindsay, H. & Robinson, M.D. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.* **42**, e91 (2014).
41. Robinson, M.D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
42. Lun, A.T.L., Bach, K. & Marioni, J.C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
43. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1**, 80–83 (1945).
44. McDavid, A. *et al.* Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29**, 461–467 (2013).
45. Welch, B.L. The generalisation of student's problems when several different population variances are involved. *Biometrika* **34**, 28–35 (1947).
46. Irizarry, R.A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–350 (2005).
47. Svensson, V., Vento-Tormo, R. & Teichmann, S.A. Moore's law in single cell transcriptomics. Preprint available at https://arxiv.org/abs/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* **32**, 3047–3048 (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).

naturereseach

Corresponding author(s): Charlotte Soneson & Mark D Robinson

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

▸ Experimental design

1. Sample size

Describe how sample size was determined.

No sample size calculations were necessary.

2. Data exclusions

Describe any data exclusions.

As described in detail in the Online Methods, we selected large, homogeneous cell groups from the public data sets to include in the method evaluation. Apart from that, no data were excluded.

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

The study does not include experimental findings. All code is available in order to facilitate reproducibility of computational analyses.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No allocation into experimental groups was performed.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No allocation into experimental groups was performed.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)

☒ ☐ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☒ ☐ A statement indicating how many times each experiment was replicated

☒ ☐ The statistical test(s) used and whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ ☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons

☒ ☐ Test values indicating whether an effect is present
Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.

☒ ☐ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)

☒ ☐ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Nature Methods: doi:10.1038/nmeth.4612

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

All code used to build the conquer database and perform the method evaluation is available on GitHub: https://github.com/markrobinsonuzh/conquer, https://github.com/csoneson/conquer_comparison.
All compared methods are implemented as R (v3.3-3.4) or Python packages. Versions are given below, as well as in Supplementary Table 2.
BPSC (v0.99.0/1)
D3E (v1.0)
DESeq2 (v1.14.1)
DEsingle (v0.1.0)
edgeR (v 3.19.1)
scran (v1.2.0)
limma (v.3.30.13)
MAST (v1.0.5)
metagenomeSeq (v1.16.0)
monocle (v2.2.0)
NODES (v0.0.0.9010)
ROTS (v1.2.0)
samr (v2.0)
scDD (v1.0.0)
scde (v2.2.0)
Seurat (v1.4.0.7)
Plotting was done with ggplot2 (v2.2.1)
Other R packages used for the evaluation:
iCOBRA (v1.2.0)
IHW (v1.2.0)
powsim (v1.0)

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

No material or reagents were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No cell lines were used.

b. Describe the method of cell line authentication used.

No cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

No human participants were used.