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

Charlotte Soneson[1,2,*] 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 charlotte.soneson@uzh.ch or mark.robinson@imls.uzh.ch

Nature Methods: doi:10.1038/nmeth.4612

1 Supplementary Tables

Supplementary Table 1: Experimental single-cell datasets used for the differential expression method evaluation.

dataset	Protocol	Compared cell subsets	Number of cells / group	Used for simulation	Organism	Ref
GSE45719	SMART-Seq (full-length)	16-cell stage blastomere *vs* Mid blastocyst cell (92-94h post-fertilization)	50, 24, 12, 6	yes	mouse	[1]
GSE45719null	SMART-Seq (full-length)	16-cell stage blastomere	24, 12, 6		mouse	[1]
GSE48968-GPL13112	SMARTer C1 (full-length)	BMDC (1h LPS Stimulation) *vs* BMDC (4h LPS Stimulation)	95, 48, 24, 12		mouse	[2]
GSE48968-GPL13112null	SMARTer C1 (full-length)	BMDC (1h LPS Stimulation)	48, 24, 12, 6		mouse	[2]
GSE60749-GPL13112	SMARTer C1 (full-length)	v6.5 mouse embryonic stem cells, culture conditions: 2i+LIF *vs* v6.5 mouse embryonic stem cells, culture conditions: serum+LIF	90, 48, 24, 12	yes	mouse	[3]
GSE60749-GPL13112null	SMARTer C1 (full-length)	v6.5 mouse embryonic stem cells, culture conditions: 2i+LIF	47, 24, 12, 6		mouse	[3]
GSE74596	Smart-Seq2 (full-length)	NKT0 *vs* NKT17	44, 22, 12, 6	yes	mouse	[4]
GSE74596null	Smart-Seq2 (full-length)	NKT0	22, 12, 6		mouse	[4]
EMTAB2805	SMARTer C1 (full-length)	G1 *vs* G2M	96, 48, 24, 12		mouse	[5]
EMTAB2805null	SMARTer C1 (full-length)	G2M	48, 24, 12, 6		mouse	[5]
UsoskinGSE59739	STRT-Seq (UMI)	RT-1.NP1 *vs* RT-1.TH	58, 36, 24, 12		mouse	[6]
UsoskinGSE59739null	STRT-Seq (UMI)	RT-1.NP1	24, 12, 6		mouse	[6]
GSE63818-GPL16791	Tang (full-length)	Primordial Germ Cells, developmental stage: 7 week gestation *vs* Somatic Cells, developmental stage: 7 week gestation	26, 12, 6		human	[7]
GSE62270-GPL17021	CEL-Seq (UMI)	Randomly extracted cells from whole intestinal organoids *vs* Randomly extracted ex vivo isolated 5 day YFP positive cells	400, 240, 120, 48, 24		mouse	[8]
GSE62270-GPL17021null	CEL-Seq (UMI)	Randomly extracted ex vivo isolated 5 day YFP positive cells	200, 120, 48, 24, 12		mouse	[8]
10XMonoCytoT	10X Genomics GemCode (UMI)	CD14 monocytes *vs* cytotoxic T-cells	240, 120, 48, 24		human	[9]
10XMonoCytoTnull	10X Genomics GemCode (UMI)	cytotoxic T-cells	120, 48, 24, 12		human	[9]

Nature Methods: doi:10.1038/nmeth.4612

Supplementary Table 2: Evaluated differential expression methods, together with package versions and the type of input values provided to each of them. Note that "raw counts" here refers to length-scaled TPMs, which are on the scale of the raw counts, but are unaffected by differential isoform usage [10]. CPM values are calculated with edgeR, and Census counts with monocle.

	Short name	Method	Software version	Input	Available from	Reference
	BPSC	BPSC	BPSC 0.99.0/1	CPM	GitHub	[11]
	D3E	D3E	D3E 1.0	raw counts	GitHub	[12]
	DESeq2	DESeq2	DESeq2 1.14.1	raw counts	Bioconductor	[13]
	DESeq2betapFALSE	DESeq2 without beta prior	DESeq2 1.14.1	raw counts	Bioconductor	[13]
	DESeq2census	DESeq2	DESeq2 1.14.1	Census counts	Bioconductor	[13]
	DESeq2nofilt	DESeq2 without the built-in independent filtering	DESeq2 1.14.1	raw counts	Bioconductor	[13]
	DEsingle	DEsingle	DEsingle 0.1.0	raw counts	GitHub	[14]
	edgeRLRT	edgeR/LRT	edgeR 3.19.1	raw counts	Bioconductor	[15–17]
	edgeRLRTcensus	edgeR/LRT	edgeR 3.19.1	Census counts	Bioconductor	[15–17]
	edgeRLRTdeconv	edgeR/LRT with deconvolution normalization	edgeR 3.19.1, scran 1.2.0	raw counts	Bioconductor	[15, 17, 18]
	edgeRLRTrobust	edgeR/LRT with robust dispersion estimation	edgeR 3.19.1	raw counts	Bioconductor	[15–17, 19]
	edgeRQLF	edgeR/QLF	edgeR 3.19.1	raw counts	Bioconductor	[15, 16, 20]
	edgeRQLFDetRate	edgeR/QLF with cellular detection rate as covariate	edgeR 3.19.1	raw counts	Bioconductor	[15, 16, 20]
	limmatrend	limma-trend	limma 3.30.13	log_2(CPM)	Bioconductor	[21, 22]
	MASTcpm	MAST	MAST 1.0.5	log_2(CPM+1)	Bioconductor	[23]
	MASTcpmDetRate	MAST with cellular detection rate as covariate	MAST 1.0.5	log_2(CPM+1)	Bioconductor	[23]
	MASTtpm	MAST	MAST 1.0.5	log_2(TPM+1)	Bioconductor	[23]
	MASTtpmDetRate	MAST with cellular detection rate as covariate	MAST 1.0.5	log_2(TPM+1)	Bioconductor	[23]
	metagenomeSeq	metagenomeSeq	metagenomeSeq 1.16.0	raw counts	Bioconductor	[24]
	monocle	monocle (tobit)	monocle 2.2.0	TPM	Bioconductor	[25]
	monoclecensus	monocle (Negative Binomial)	monocle 2.2.0	Census counts	Bioconductor	[25, 26]
	monoclecount	monocle (Negative Binomial)	monocle 2.2.0	raw counts	Bioconductor	[25]
	NODES	NODES	NODES 0.0.0.9010	raw counts	Author-provided link	[27]
	ROTScpm	ROTS	ROTS 1.2.0	CPM	Bioconductor	[28, 29]
	ROTStpm	ROTS	ROTS 1.2.0	TPM	Bioconductor	[28, 29]
	ROTSvoom	ROTS	ROTS 1.2.0	voom-transformed raw counts	Bioconductor	[28, 29]
	SAMseq	SAMseq	samr 2.0	raw counts	CRAN	[30]
	scDD	scDD	scDD 1.0.0	raw counts	Bioconductor	[31]
	SCDE	SCDE	scde 2.2.0	raw counts	Bioconductor	[32]
	SeuratBimod	Seurat (bimod test)	Seurat 1.4.0.7	raw counts	GitHub	[33, 34]
	SeuratBimodnofilt	Seurat (bimod test) without the internal filtering	Seurat 1.4.0.7	raw counts	GitHub	[33, 34]
	SeuratBimodIsExpr2	Seurat (bimod test) with internal expression threshold set to 2	Seurat 1.4.0.7	raw counts	GitHub	[33, 34]
	SeuratTobit	Seurat (tobit test)	Seurat 1.4.0.7	TPM	GitHub	[25, 33]
	ttest	t-test	stats (R v 3.3)	TMM-normalized TPM	CRAN	[16, 35]
	voomlimma	voom-limma	limma 3.30.13	raw counts	Bioconductor	[21, 22]
	Wilcoxon	Wilcoxon test	stats (R v 3.3)	TMM-normalized TPM	CRAN	[16, 36]

Nature Methods: doi:10.1038/nmeth.4612

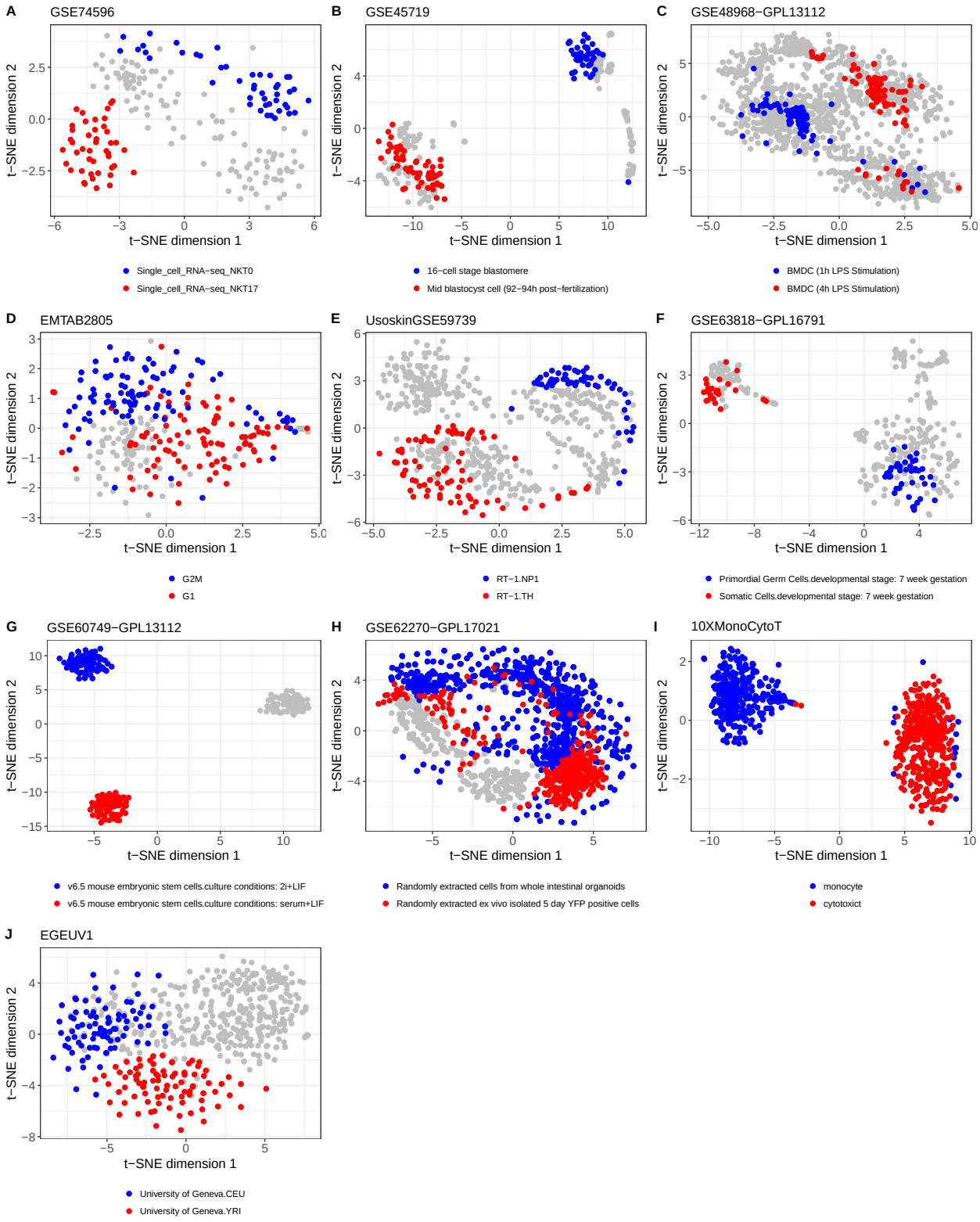
Supplementary Table 3: The processing pipeline used by *conquer* for datasets obtained with full-length transcript protocols (left column) and with 3' end sequencing (UMI) protocols (right column).

Full-length	**UMI**
1. Build a quasi-mapping transcriptome index for Salmon [37] (full-length datasets) or RapMap [38] (UMI datasets) from the combined set of annotated cDNA and ncRNA sequences as well as ERCC spike-in sequences.	
2a. For each cell, find the corresponding SRA run ID(s) and download and merge the respective FASTQ file(s). If requested, trim adapters using cutadapt [39].	
2b. Perform quality control of the FASTQ file(s) using FastQC (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`).	
2c. Estimate transcript abundances (TPMs and estimated counts) using Salmon.	2c. Map reads to the indexed transcriptome using RapMap. 2d. Count the number of unique UMIs assigned to each gene using the umis tool [40].
3. Summarize FastQC and Salmon diagnostics (for full-length datasets) for all cells in the dataset using MultiQC [41].	
4. Read transcript abundances with the tximport R package [10] and create a MultiAssayExperiment object [42] containing gene- and transcript-level estimated counts and TPMs for the entire dataset, as well as phenotypic information obtained from the public repository. For each gene, we include both aggregated transcript counts and length-scaled TPMs [10], as calculated by tximport.	4. Create a MultiAssayExperiment object [42] containing gene- and transcript-level estimated UMI counts for the entire dataset, as well as phenotypic information obtained from the public repository.
5. Perform quality control, exploratory analysis and visualization of the gene-level abundances using the scater Bioconductor package [43] and summarize in a report.	

Nature Methods: doi:10.1038/nmeth.4612

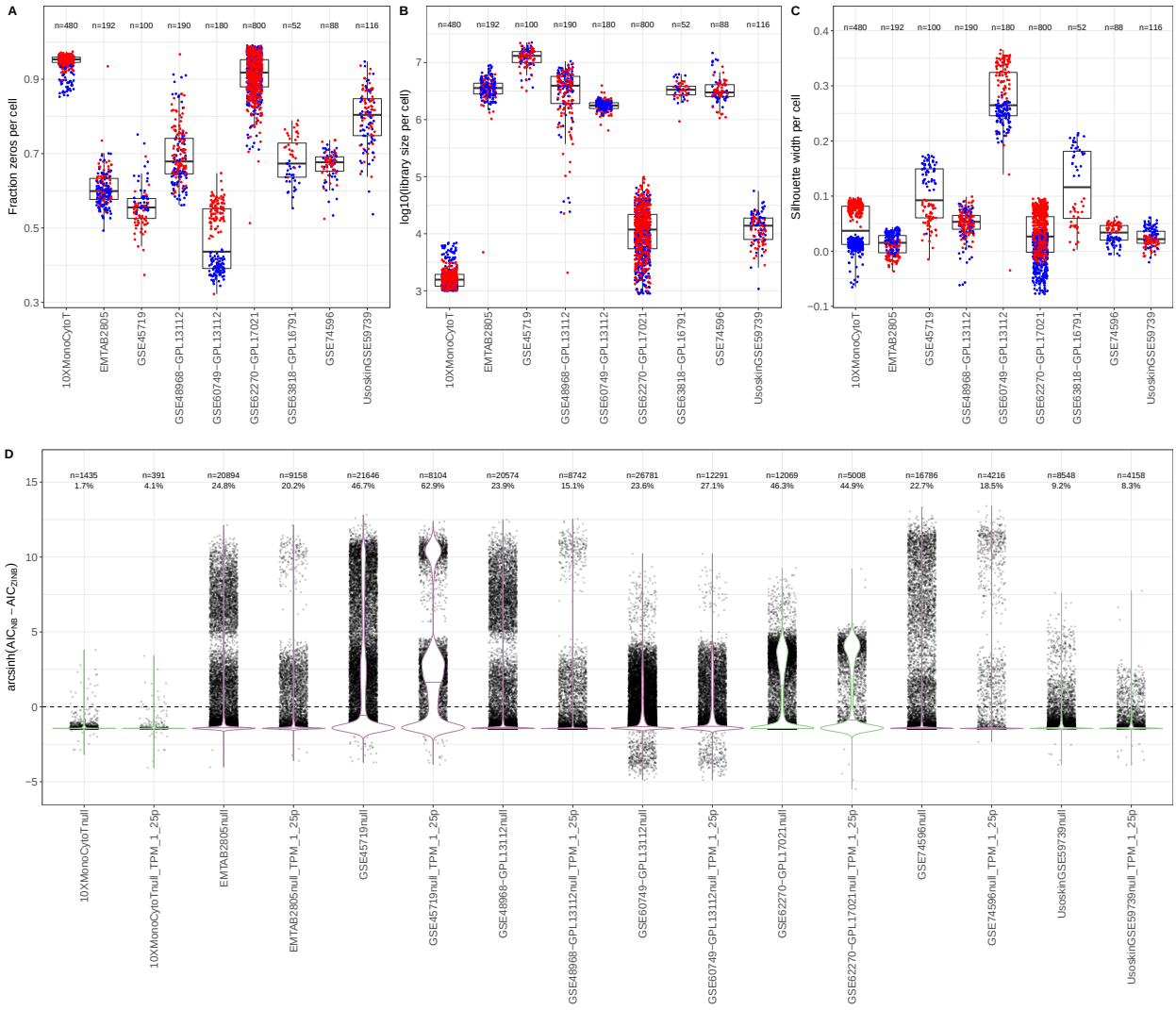
2 Supplementary Figures

Nature Methods: doi:10.1038/nmeth.4612



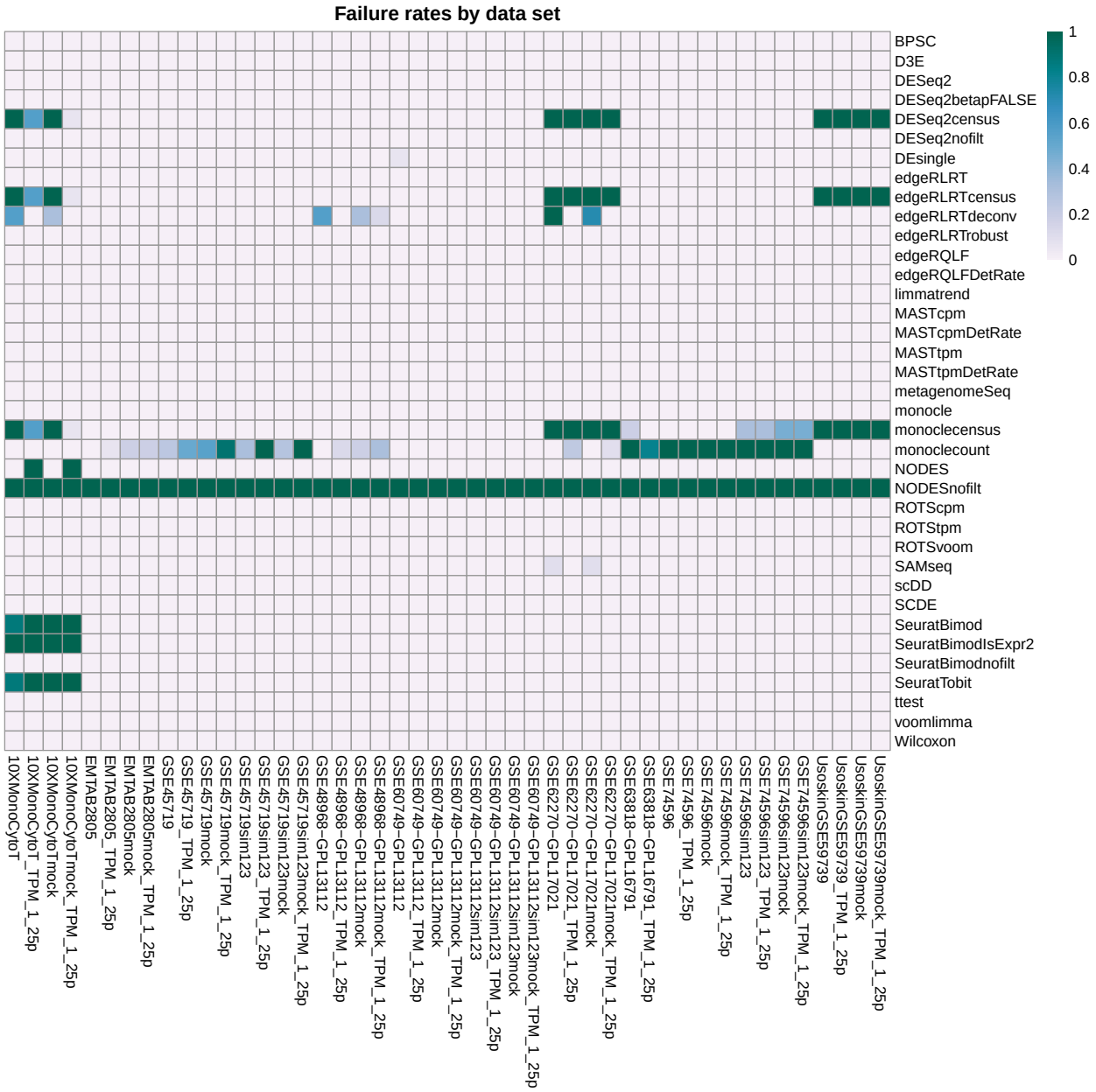
Supplementary Figure 1: t-SNE representations [44] of all real datasets used in this study, calculated using the *scater* R package [43]. Panels A-I represent single-cell RNA-seq dataset, while panel J represents the Geuvadis bulk RNA-seq dataset used for comparison. Each point represents one cell for all datasets except EGEUV1, where it represents a bulk sample. For each dataset, the two cell/sample populations that are selected for the differential expression analysis evaluation are indicated in red and blue, and all others are shown in grey.

Nature Methods: doi:10.1038/nmeth.4612



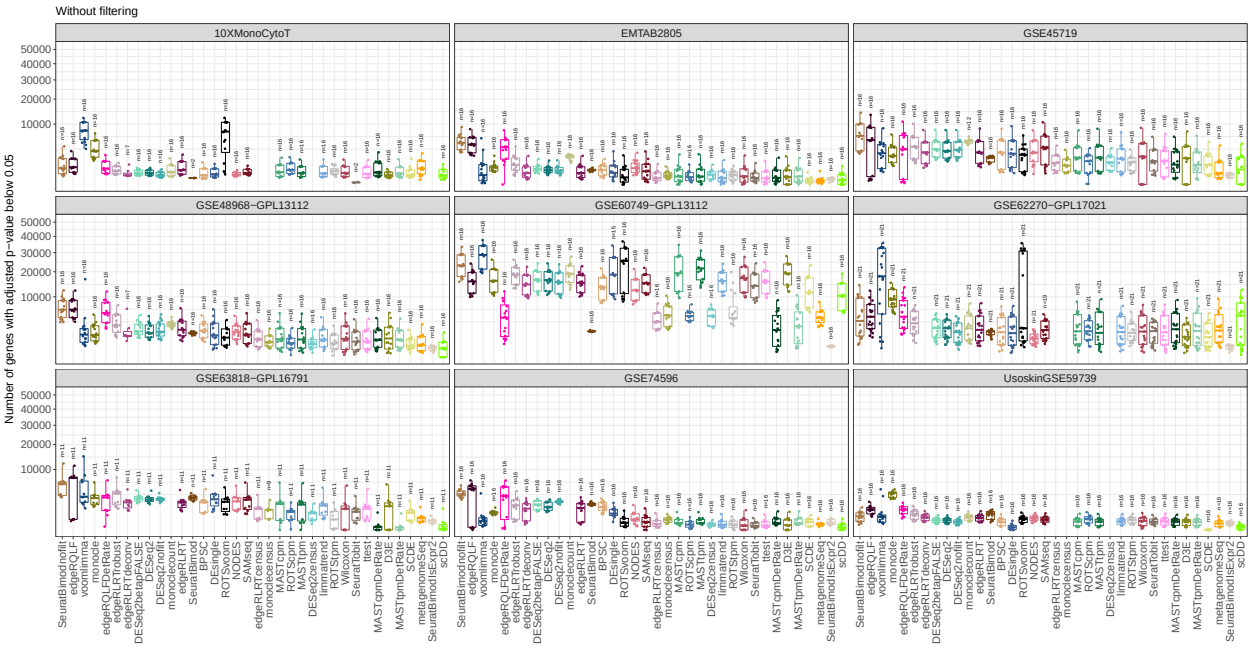
Supplementary Figure 2: A-C. Characteristics of the cells in the maximal size subset of each real scRNA-seq dataset. Each dot corresponds to a cell, and cells from the two different populations selected for the comparison are shown in different colors (cf. Supplementary Figure 1). Center line, median; hinges, first and third quartiles; whiskers, most extreme values within 1.5 IQR from the box; n, number of cells. A. The fraction of genes with an estimated expression of zero. B. The sum of the counts (length-scaled TPMs for the full-length datasets, UMI counts for the UMI datasets) across all genes. C. The silhouette width [45], quantifying how much more similar a cell is to the cells in the same population compared to the cells in the other population. The silhouette widths are based on Euclidean distances calculated from log_2(TPM+1) expression values of all genes. D. Difference between the Akaike Information Criterion (AIC [46]) for a Negative Binomial and a zero-inflated Negative Binomial fit, for each gene in each of the real null datasets, before and after filtering out lowly expressed genes. A large positive value for a gene indicates that a zero-inflated Negative Binomial is preferable to a Negative Binomial distribution for modeling the counts. For improved visibility, the violin plot has been scaled to unit width for each dataset, and the difference between the AIC values is arcsinh-transformed. Thus, the areas inside the violin plots are not comparable across datasets. The numbers above the points indicate the number of genes and the percentage of the genes for which the zero-inflated Negative Binomial results in a lower AIC than the Negative Binomial fit. For comparison, the corresponding number for the Geuvadis bulk RNA-seq dataset is 14%. AIC values were calculated from maximal-size null data sets using the powsim package [47]. Center line, median; n, number of genes.

Nature Methods: doi:10.1038/nmeth.4612

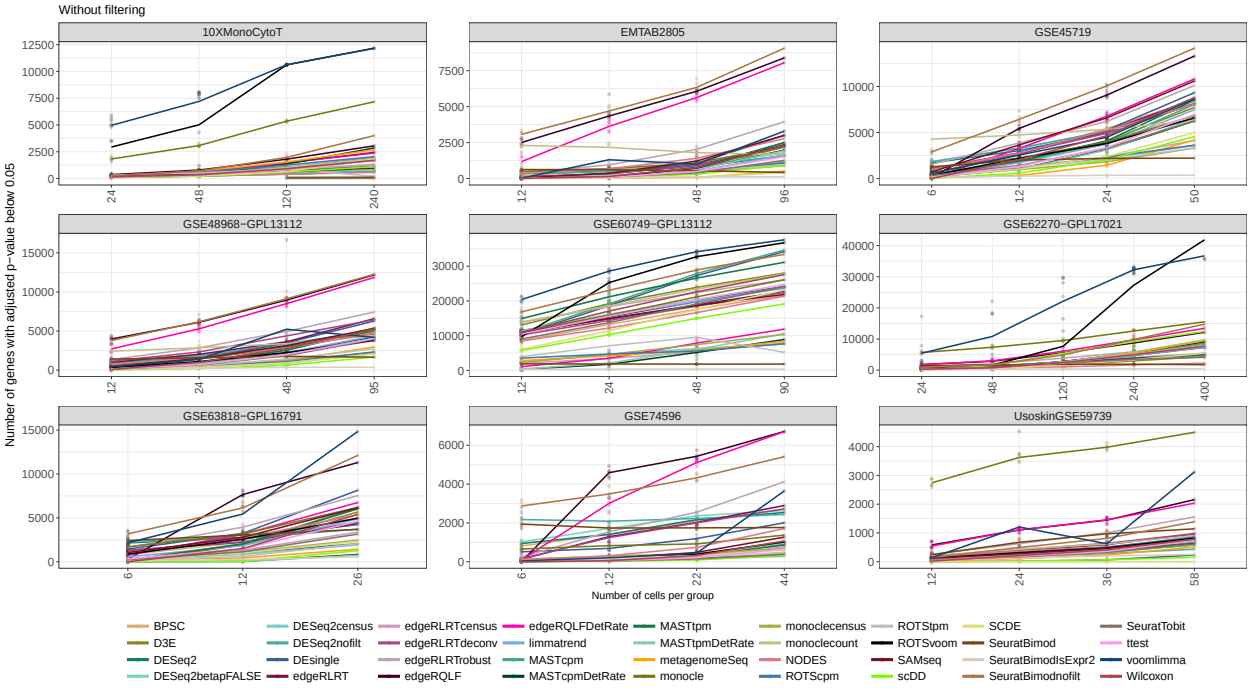


Supplementary Figure 3: The failure rate of each method across all instances of all real and simulated datasets. A method is considered to have "failed" on a dataset instance if no results were generated, regardless of the reason. The reason for the failure varies between methods. The estimation of Census counts fails for many UMI datasets where the TPM estimates (from which the Census counts are derived) are in turn derived from UMI counts by scaling these to sum to 1 million (note that this is also not a recommended workflow in practice). monoclecount fails for certain datasets, where a parametric dispersion fit for the Negative Binomial distribution could not be obtained. For edgeRLRTdeconv, the calculation of valid normalization factors fails for some of the unfiltered datasets with large numbers of zeros per cell. NODES and Seurat, which both by default apply strict internal prefiltering of cells based on the number of detected genes, fail for the 10XMonoCytoT dataset instances where no cell expresses more than 1,000 (Seurat) or 500 (NODES) genes. With the internal filtering disabled, NODES did not return valid results for any dataset.

Nature Methods: doi:10.1038/nmeth.4612

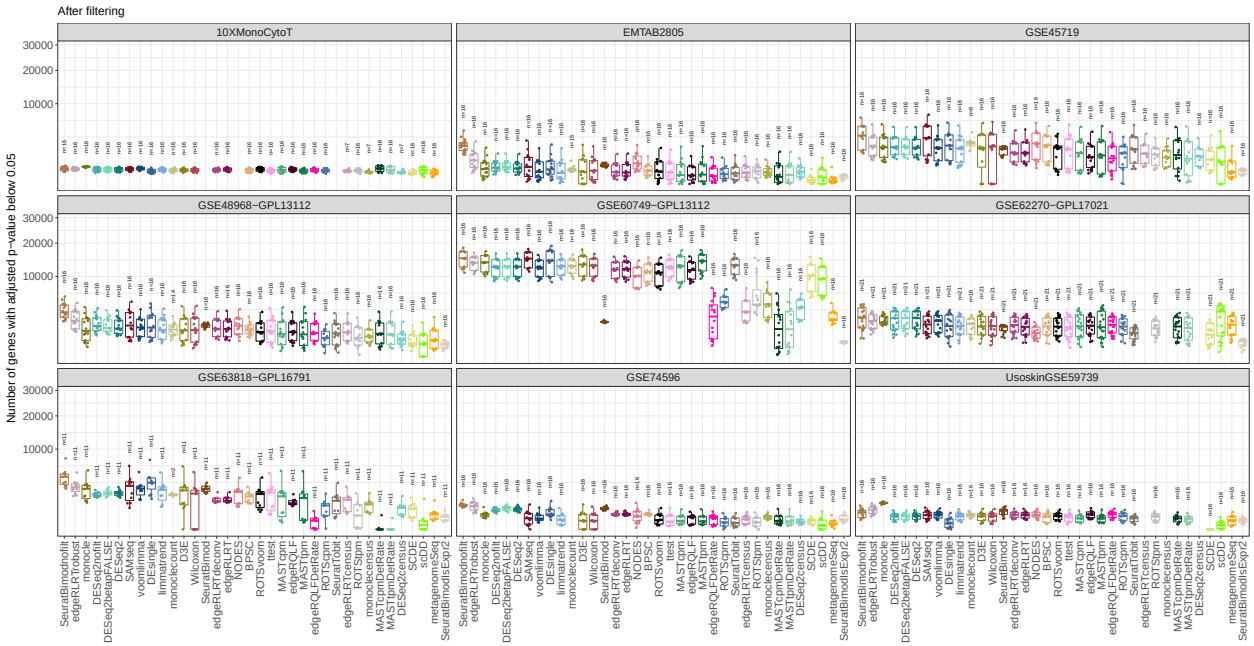


Supplementary Figure 4: The number of genes called significant (adjusted p-value below 0.05) by each differential expression method, stratified by dataset, without gene prefiltering. The order of the methods is such that the median number of significant genes (relative to the largest number of significant genes found in each dataset) across all datasets decreases from left to right. 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.

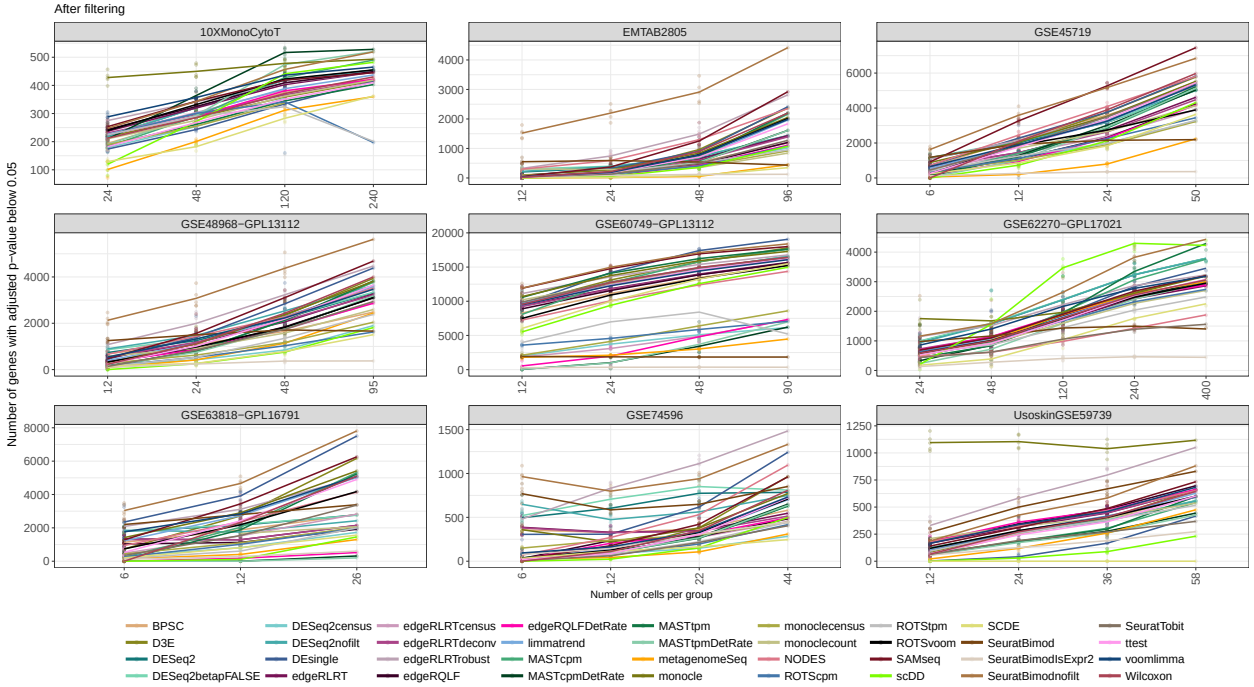


Supplementary Figure 5: The number of genes called significant (adjusted p-value below 0.05) by each differential expression method as a function of the number of cells per group, stratified by dataset, without gene prefiltering.

Nature Methods: doi:10.1038/nmeth.4612

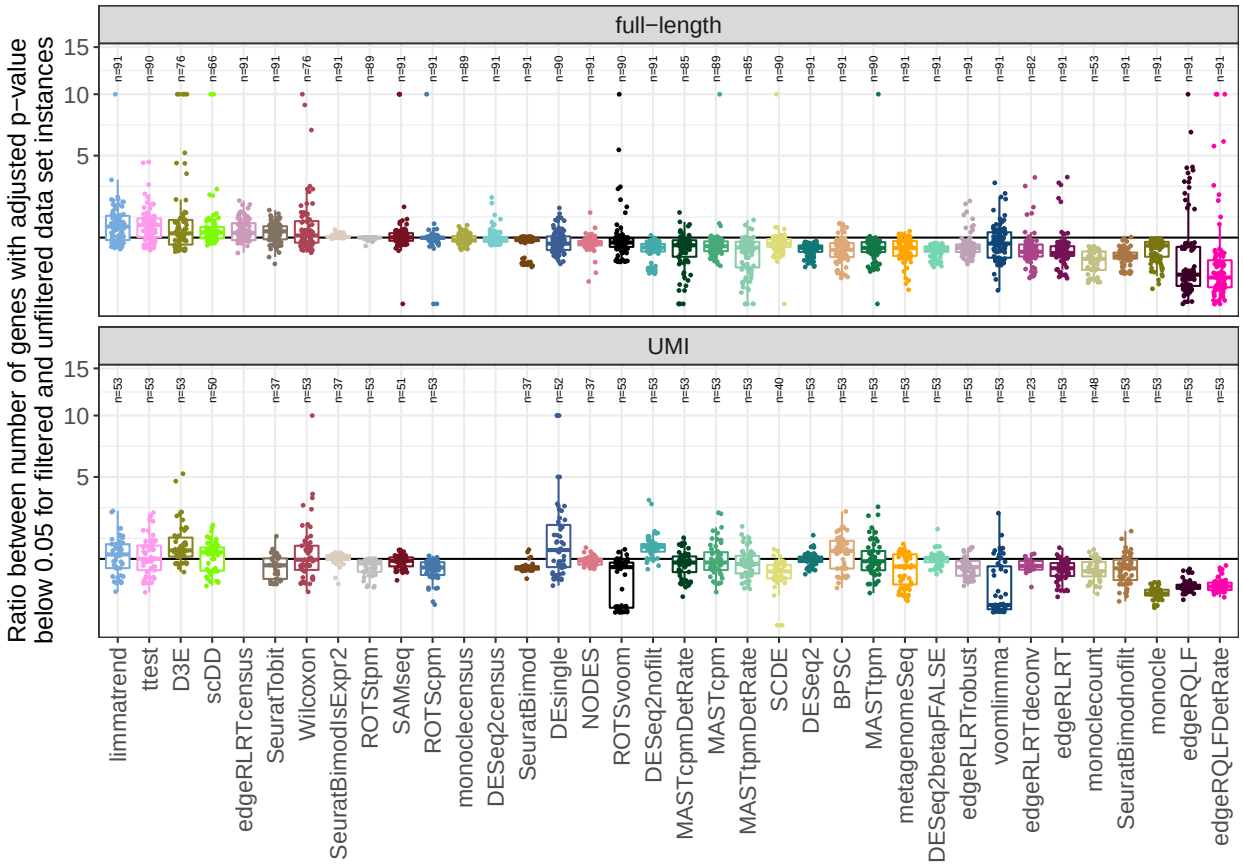


Supplementary Figure 6: The number of genes called significant (adjusted p-value below 0.05) by each differential expression method, stratified by dataset, after gene prefiltering retaining only genes with an estimated expression exceeding 1 TPM in more than 25% of the cells. The order of the methods is such that the median number of significant genes (relative to the largest number of significant genes found in each dataset) across all datasets decreases from left to right. 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.



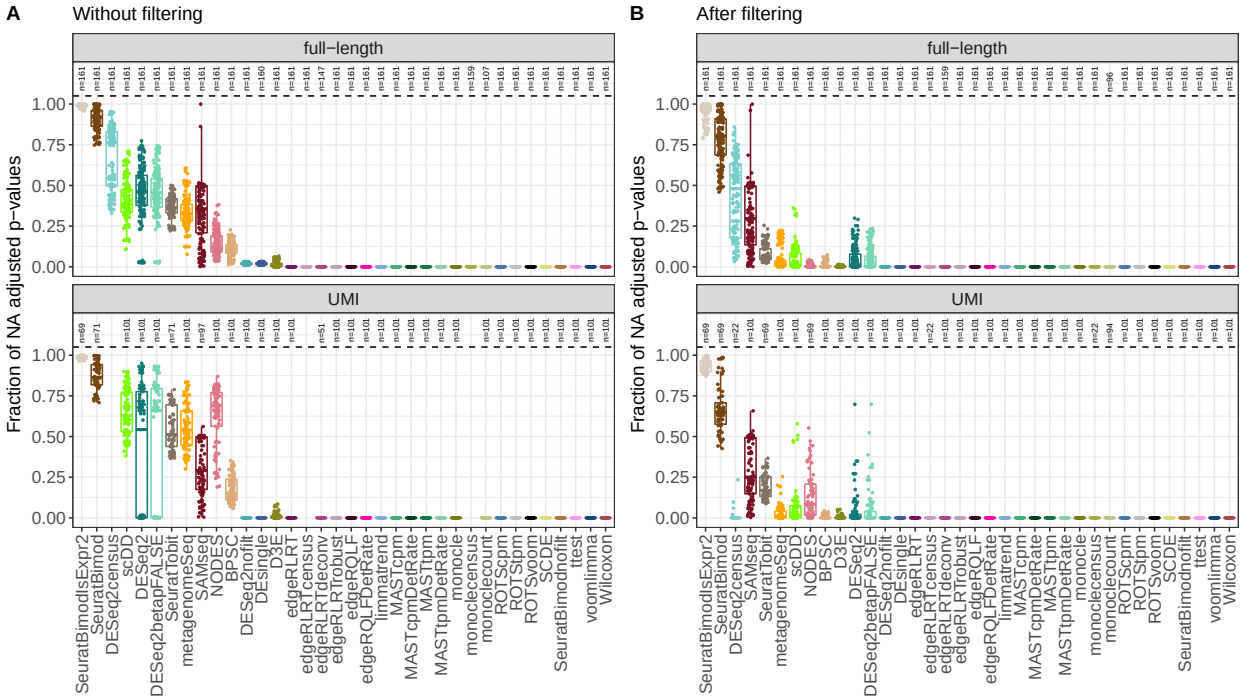
Supplementary Figure 7: The number of genes called significant (adjusted p-value below 0.05) by each differential expression method as a function of the number of cells per group, stratified by dataset, after gene prefiltering retaining only genes with an estimated expression exceeding 1 TPM in more than 25% of the cells.

Nature Methods: doi:10.1038/nmeth.4612



Supplementary Figure 8: The ratio between the number of genes called significant (adjusted p-value below 0.05) after filtering (retaining only genes with an estimated expression exceeding 1 TPM in more than 25% of the cells) and without filtering, for each differential expression method across all real signal datasets, stratified by data type (full-length vs UMI). Ratios above 10 (almost exclusively corresponding to dataset instances where no genes were detected without filtering) are replaced with 10 for increased visibility. Instances where no differentially expressed genes were found either before or after filtering are excluded. Methods are ordered by the median ratio across all dataset instances. The black horizontal line represents a ratio of 1. 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.

Nature Methods: doi:10.1038/nmeth.4612

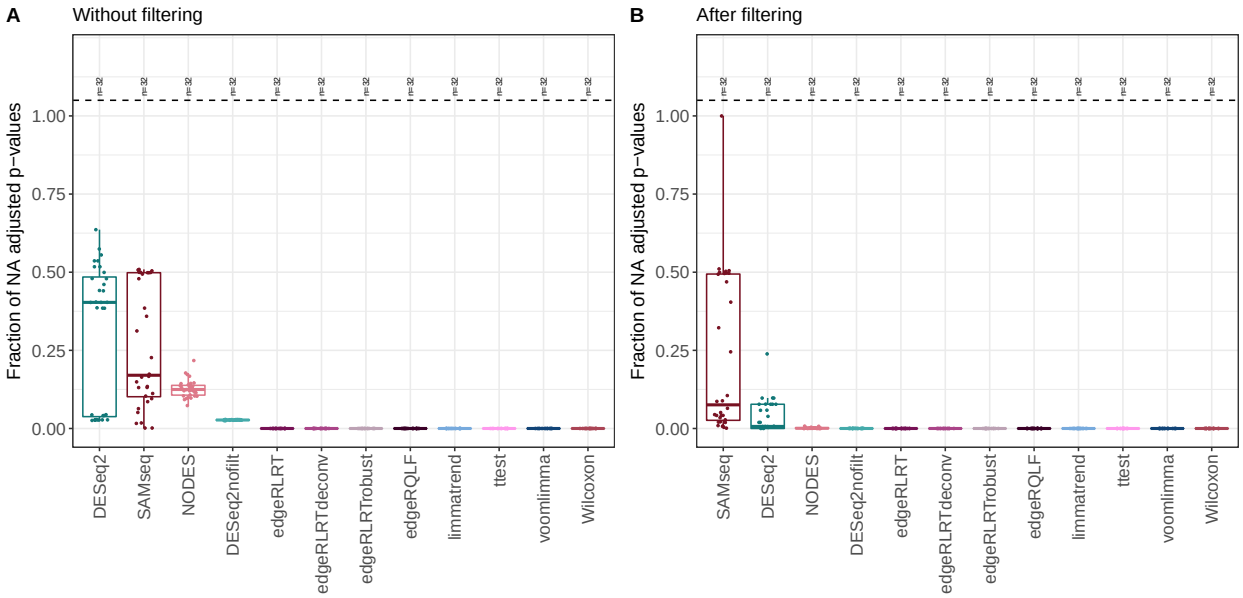


Supplementary Figure 9: Fraction of adjusted p-values reported as NA by the different methods across all instances of the 17 real single-cell datasets. Values are split between full-length and UMI datasets, and the methods are ordered by the median fraction of NA values across all datasets (separately for unfiltered and prefiltered datasets). Note that without gene prefiltering, the Census algorithm failed for the UMI datasets, and thus those results are excluded. 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 exceeding 1 TPM in more than 25% of the cells. For DESeq2 [13], NODES [27], scDD [31] and Seurat, the NA results are due to explicit internal filtering of genes and/or cells, which can be disabled (corresponding to the "nofilt" methods, note however that NODES failed to run when the internal filtering was disabled). Even without internal filtering, DESeq2 returns some NA p-values, corresponding to genes with low expression levels that, when the estimated counts are rounded before being input to DESeq2, have zero (rounded) counts for all cells. The same is true for DEsingle [14], which is also applied to rounded count estimates. For metagenomeSeq, the NA results arise predominantly from genes with 0 or 1 non-zero expression values in at least one of the conditions. SAMseq [30] only reports genes with adjusted p-values strictly less than 1, and the remaining genes are excluded from the result list. For BPSC [11] and D3E [12], the non-reported adjusted p-values result from convergence problems for a subset of the genes. After filtering, DESeq2, and in particular Seurat, still return a large number of NA test results, suggesting that their internal filtering is stricter than the fixed prefiltering criterion that we employed. For both filtered and unfiltered data, DESeq2 excludes more genes when applied to Census transcript counts than when applied to approximate read counts, likely since the former are much lower. 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.

Nature Methods: doi:10.1038/nmeth.4612

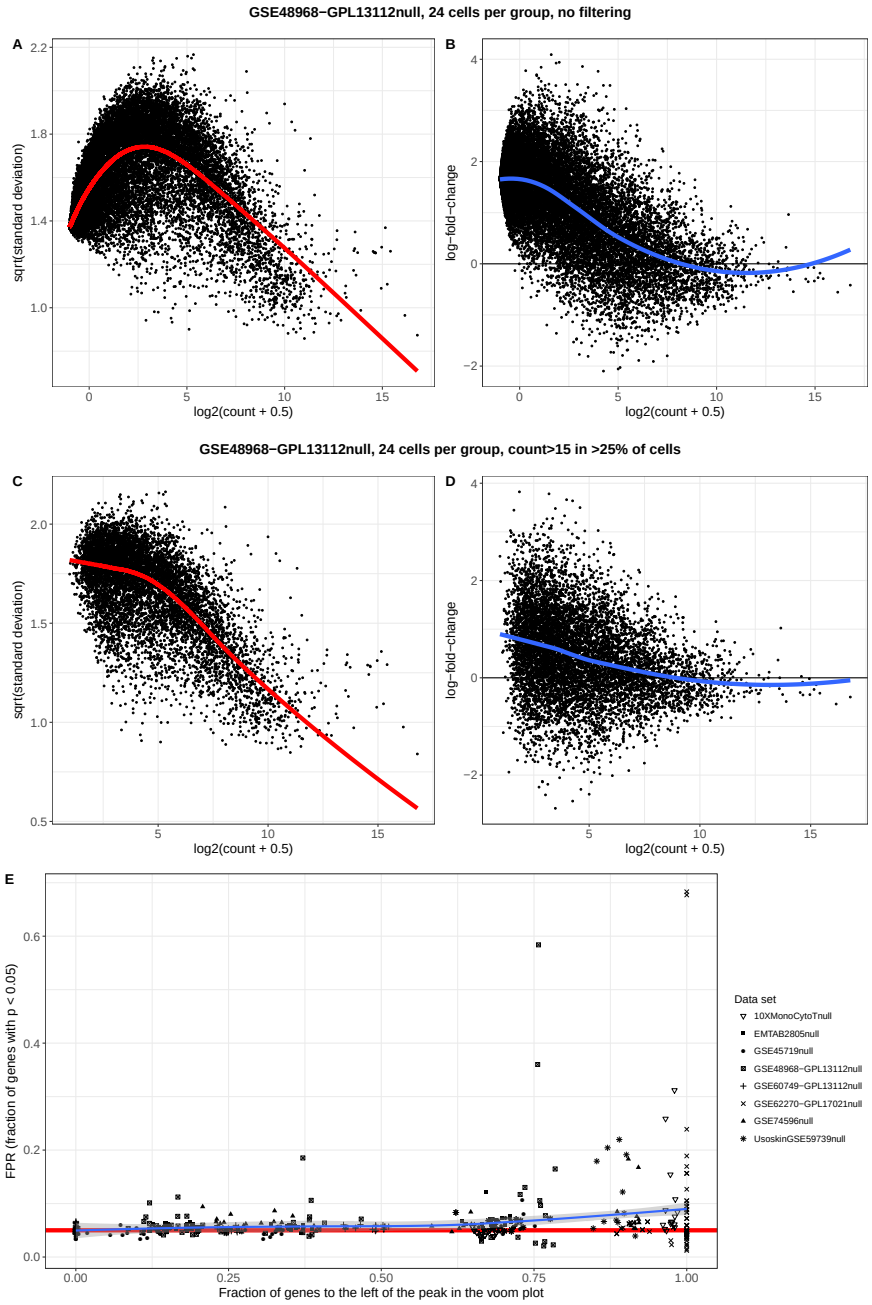
Supplementary Figure 10: Fraction of adjusted p-values reported as NA across all instances for each of the 17 signal and null scRNA-seq datasets. Note that some methods failed to return any results for some datasets (Supplementary Figure 3), and thus those results are excluded. Methods are ordered by the median fraction of NA adjusted p-values across all dataset instances. 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.

Nature Methods: doi:10.1038/nmeth.4612



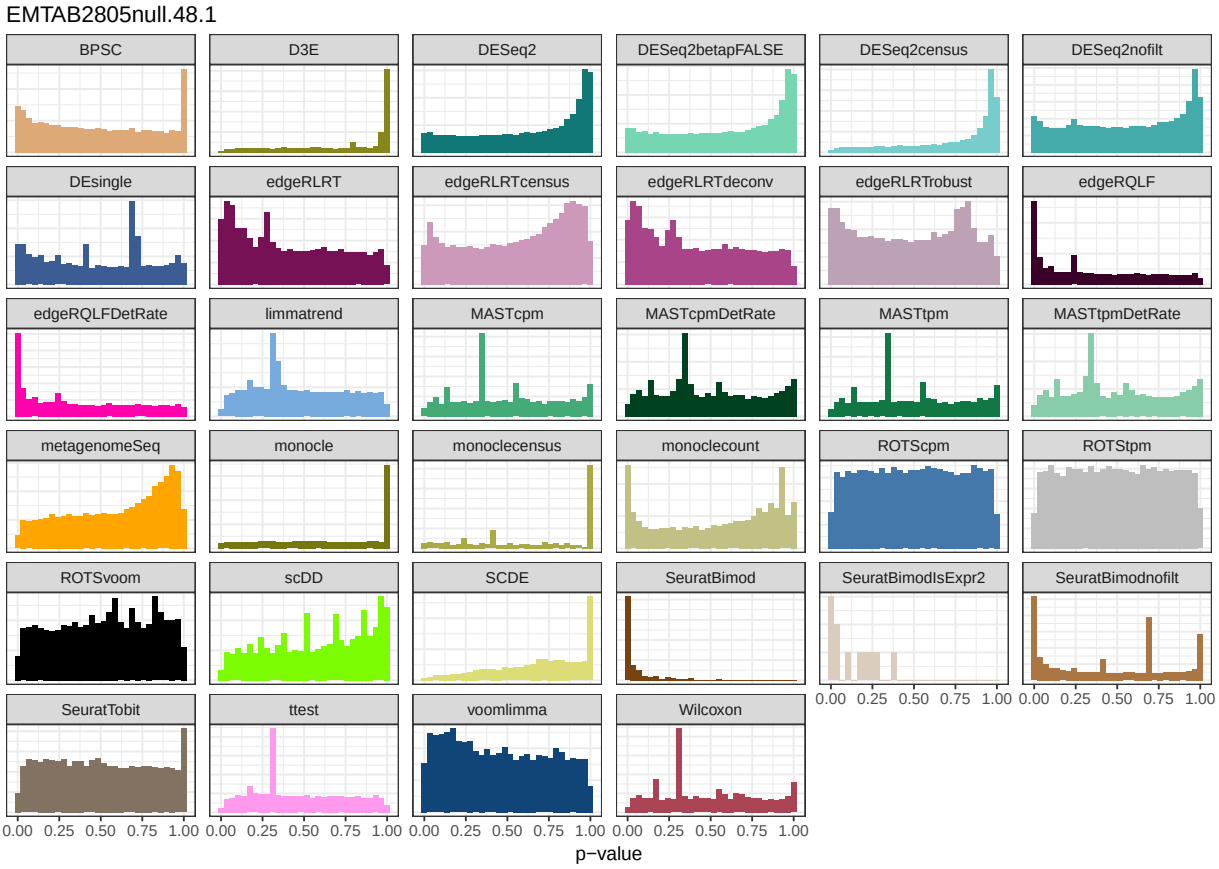
Supplementary Figure 11: Fraction of adjusted p-values reported as NA for a subset of the differential expression methods when applied to the Geuvadis bulk RNA-seq dataset. Center line, median; hinges, first and third quartiles; whiskers, most extreme values within 1.5 IQR from the box; n, number of data set instances.

Nature Methods: doi:10.1038/nmeth.4612



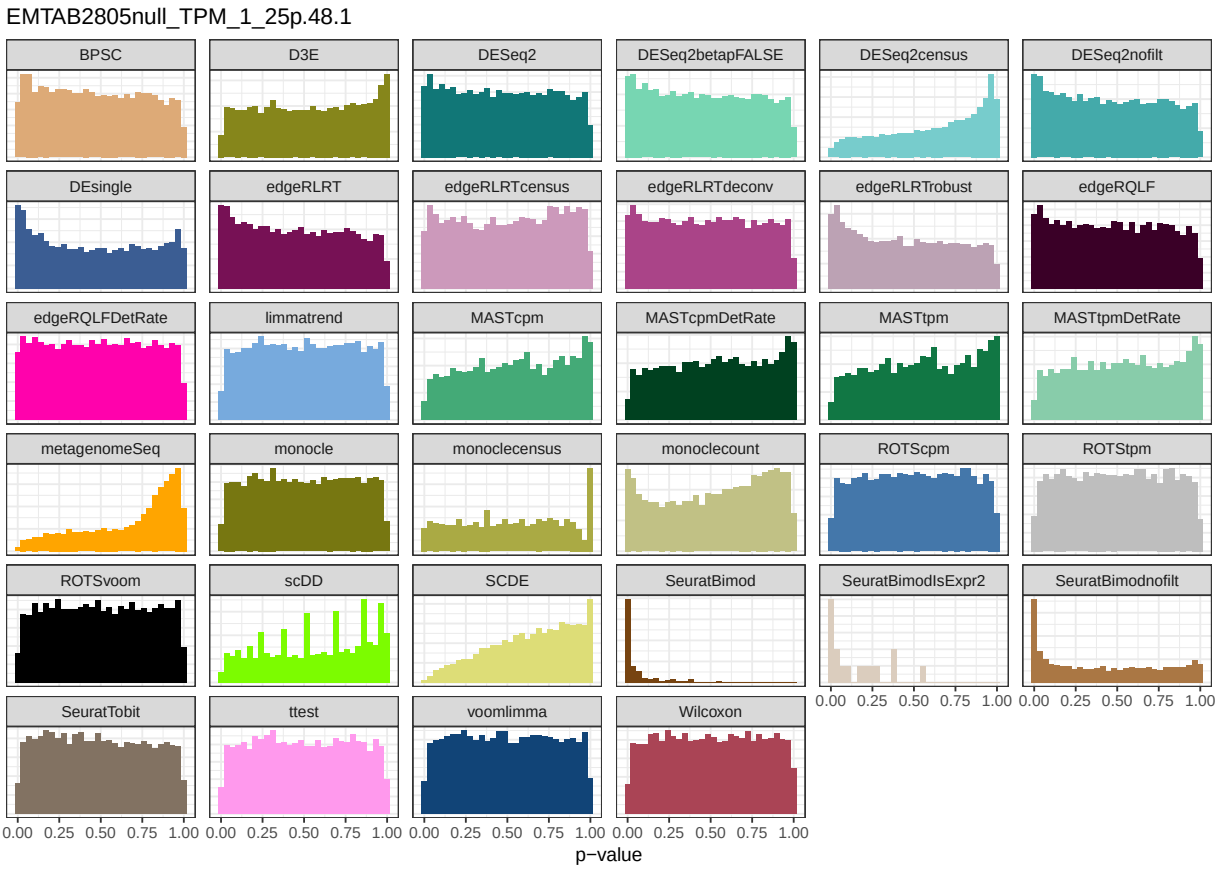
Supplementary Figure 12: Investigation of the variable FPR observed with voom/limma. A. voom plot, illustrating the mean-variance relationship of the voom-transformed counts from a GSE48968-GPL13112 null dataset instance with 24 cells per groups, without filtering. B. MA-plot, indicating the relationship between the average expression and the log-fold-change between the two groups. The MA-plot is not centered around zero log-fold-change, which induces a large number of false detections. C-D. As A-B, but for data where only genes with more than 15 reads in at least 25% of the cells are retained. The fitted mean-variance relationship in panel C is now monotonically decreasing, and the vertical shift of the MA-plot in D is considerably reduced. E. Association between the fraction of genes with expression below the point corresponding to the "peak" (the maximum standard deviation) of the fitted curve in the voom plot and the observed FPR, across multiple datasets and a wide range of filtering levels. If the fitted curve in the voom plot is monotonically decreasing (i.e., none of the genes have expression to the left of the peak), the FPR is around 5%, as expected. As the number of genes with low expression increases, the observed FPR shows an increasing trend. In all panels, the curves represent smooth fits to the data points using loess [48].

Nature Methods: doi:10.1038/nmeth.4612



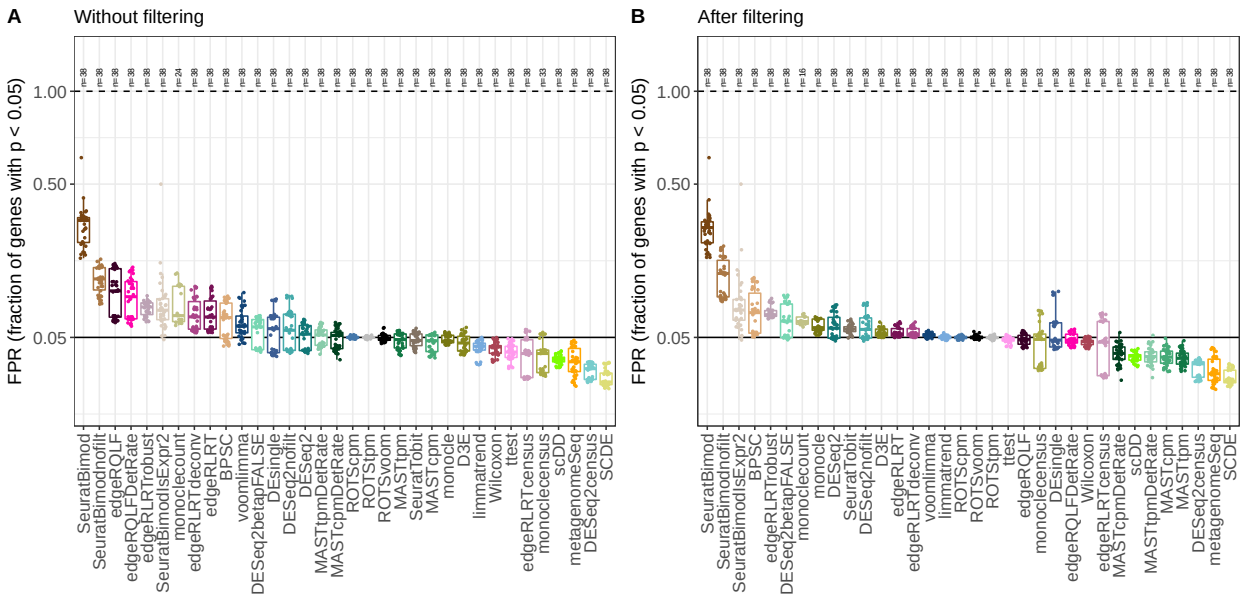
Supplementary Figure 13: Representative p-value histograms for all methods returning nominal p-values applied to one of the scRNA-seq null dataset instances, without gene prefiltering.

Nature Methods: doi:10.1038/nmeth.4612



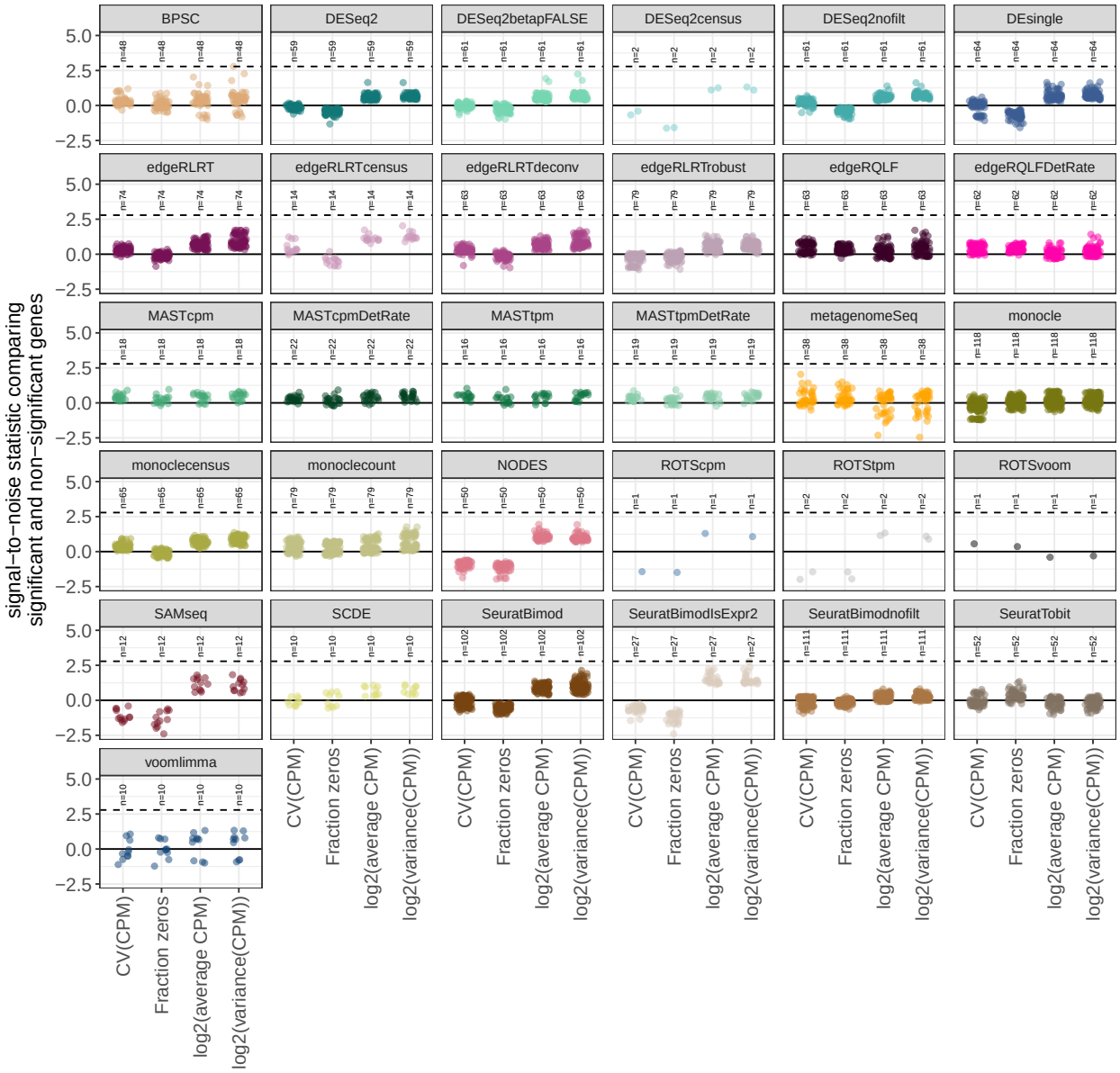
Supplementary Figure 14: Representative p-value histograms for all methods returning nominal p-values applied to one of the scRNA-seq null dataset instances, after gene prefiltering retaining only genes with an estimated expression exceeding 1 TPM in more than 25% of the cells.

Nature Methods: doi:10.1038/nmeth.4612



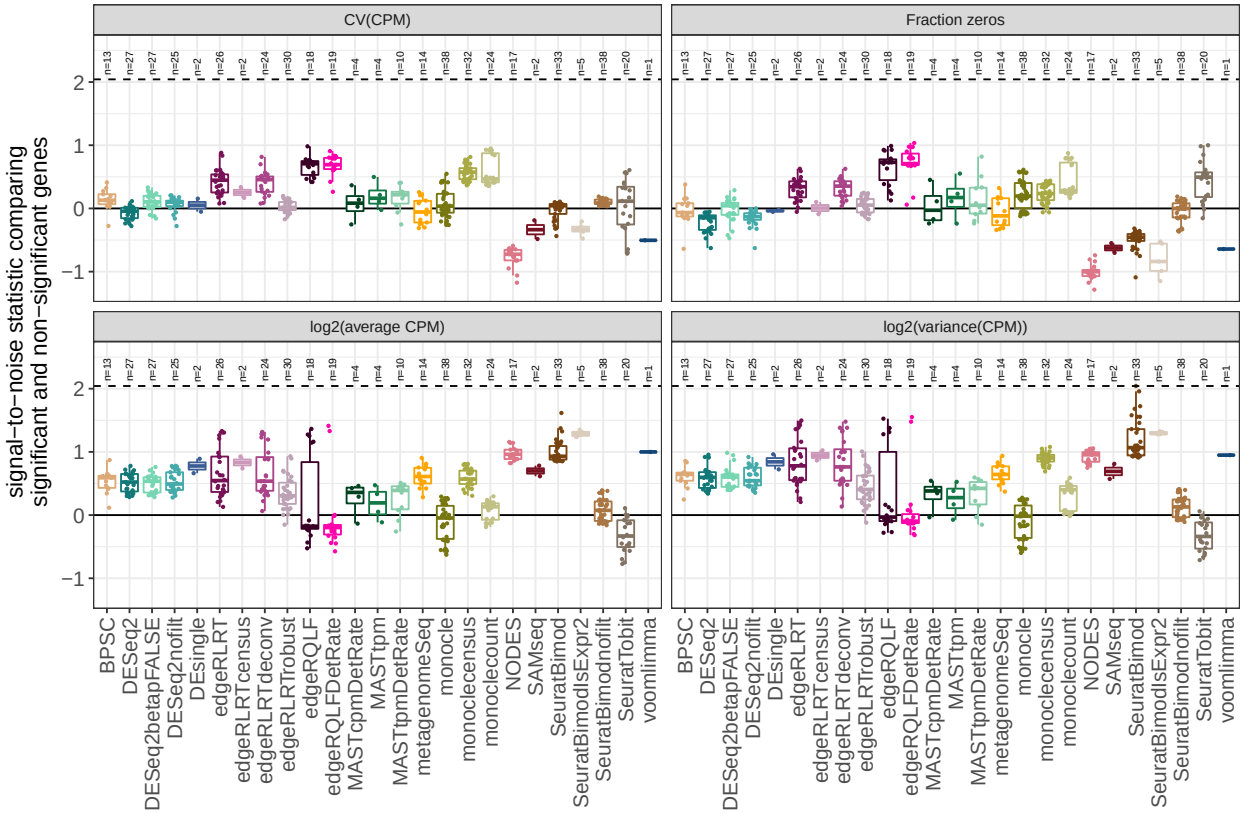
Supplementary Figure 15: Type I error control across all instances of the three simulated single-cell null datasets, with a range of sample sizes. (A) Without any prefiltering of genes. (B) After filtering, retaining only genes with an estimated expression exceeding 1 TPM in more than 25% of the cells. 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.

Nature Methods: doi:10.1038/nmeth.4612



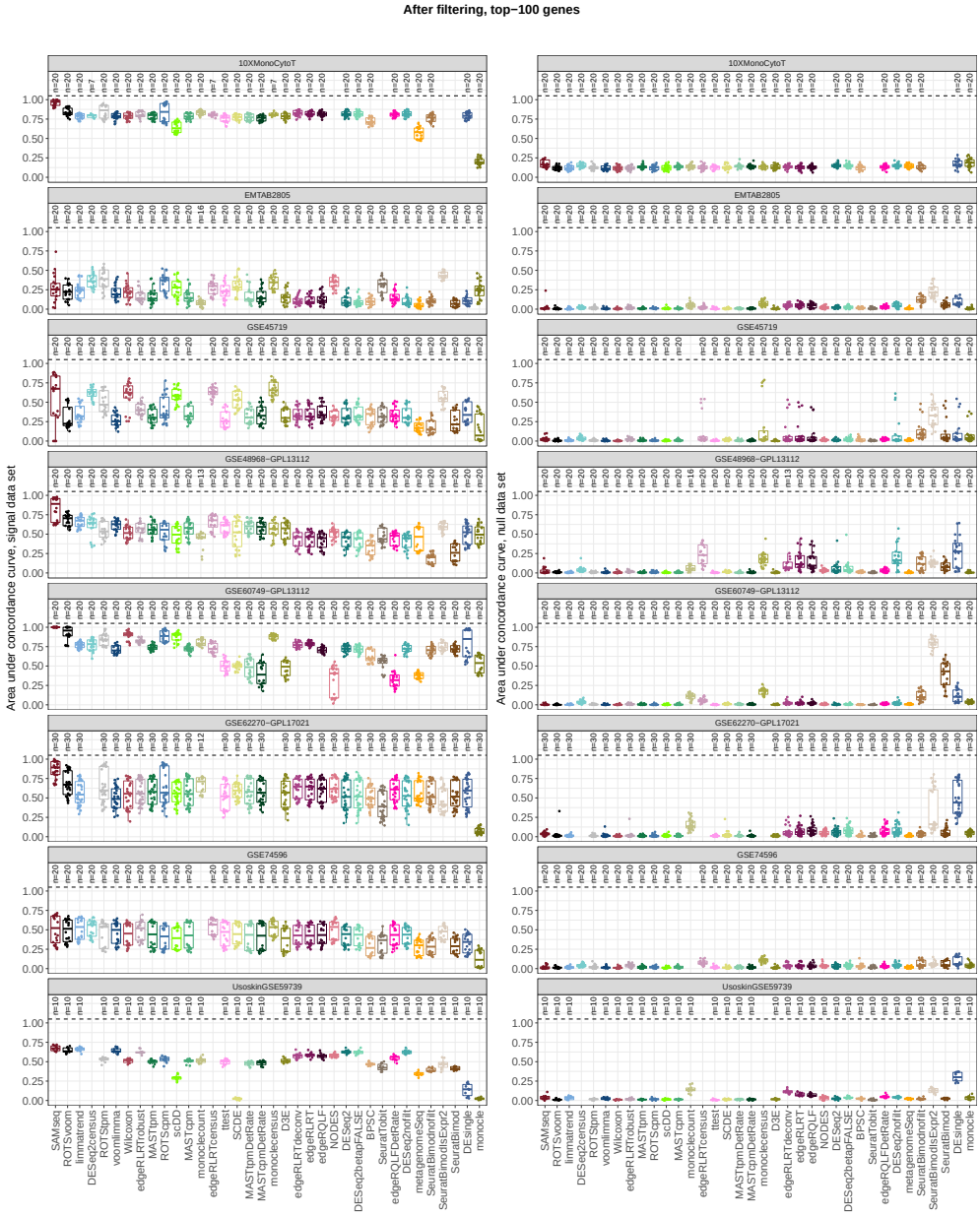
Supplementary Figure 16: Characteristics of genes falsely called significant by each of the evaluated methods in the real scRNA-seq data. For each instance of the real scRNA-seq null datasets, we record characteristics of each gene (average CPM, variance and coefficient of variation of CPM, fraction zeros across all samples) 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. n, number of data set instances with enough false positives to perform the comparison.

Nature Methods: doi:10.1038/nmeth.4612



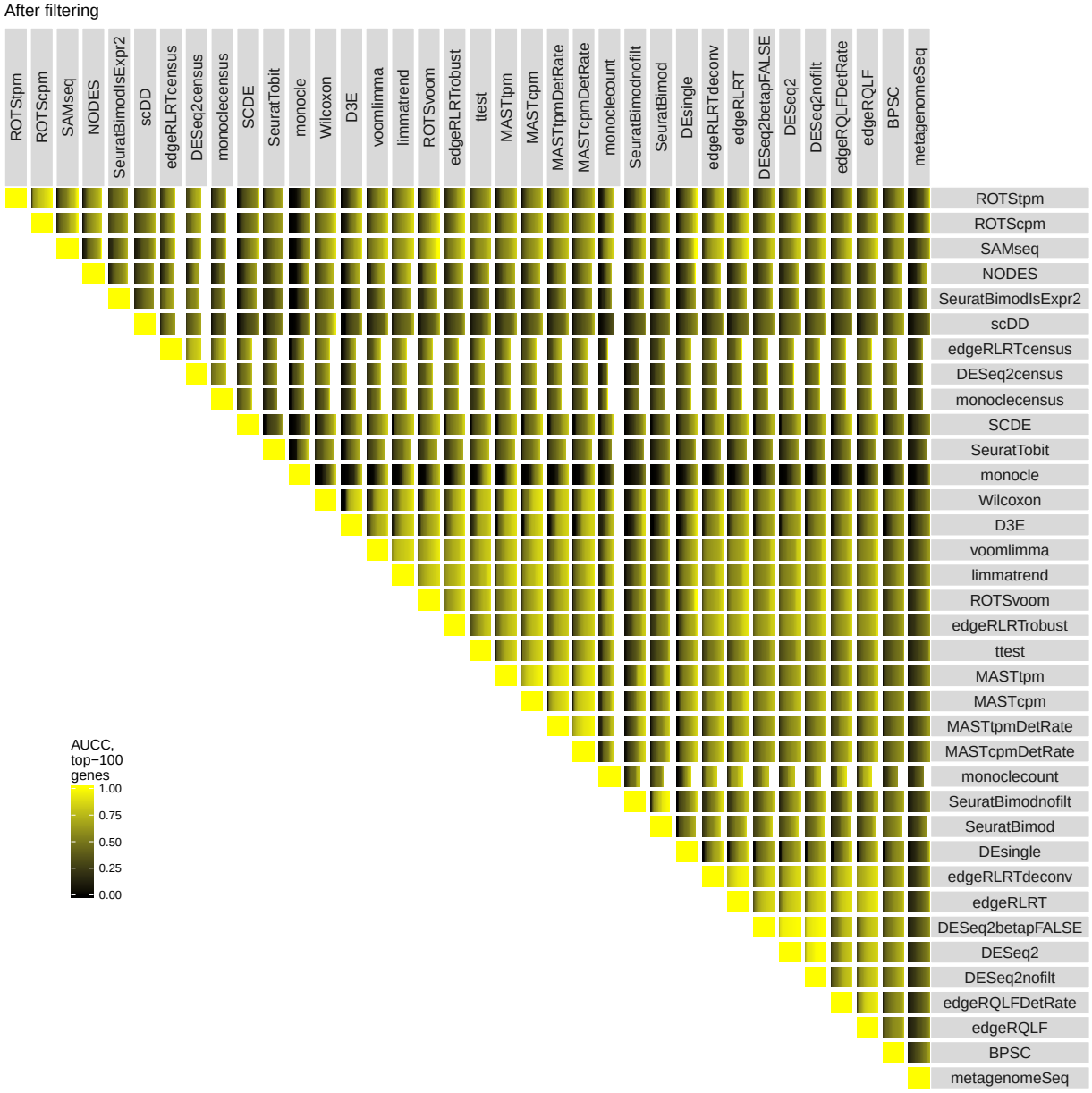
Supplementary Figure 17: Characteristics of genes falsely called significant by each of the evaluated methods in the simulated data. For each instance of the three simulated scRNA-seq null datasets, we record characteristics of each gene (average CPM, variance and coefficient of variation of CPM, fraction zeros across all samples) 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. 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.

Nature Methods: doi:10.1038/nmeth.4612



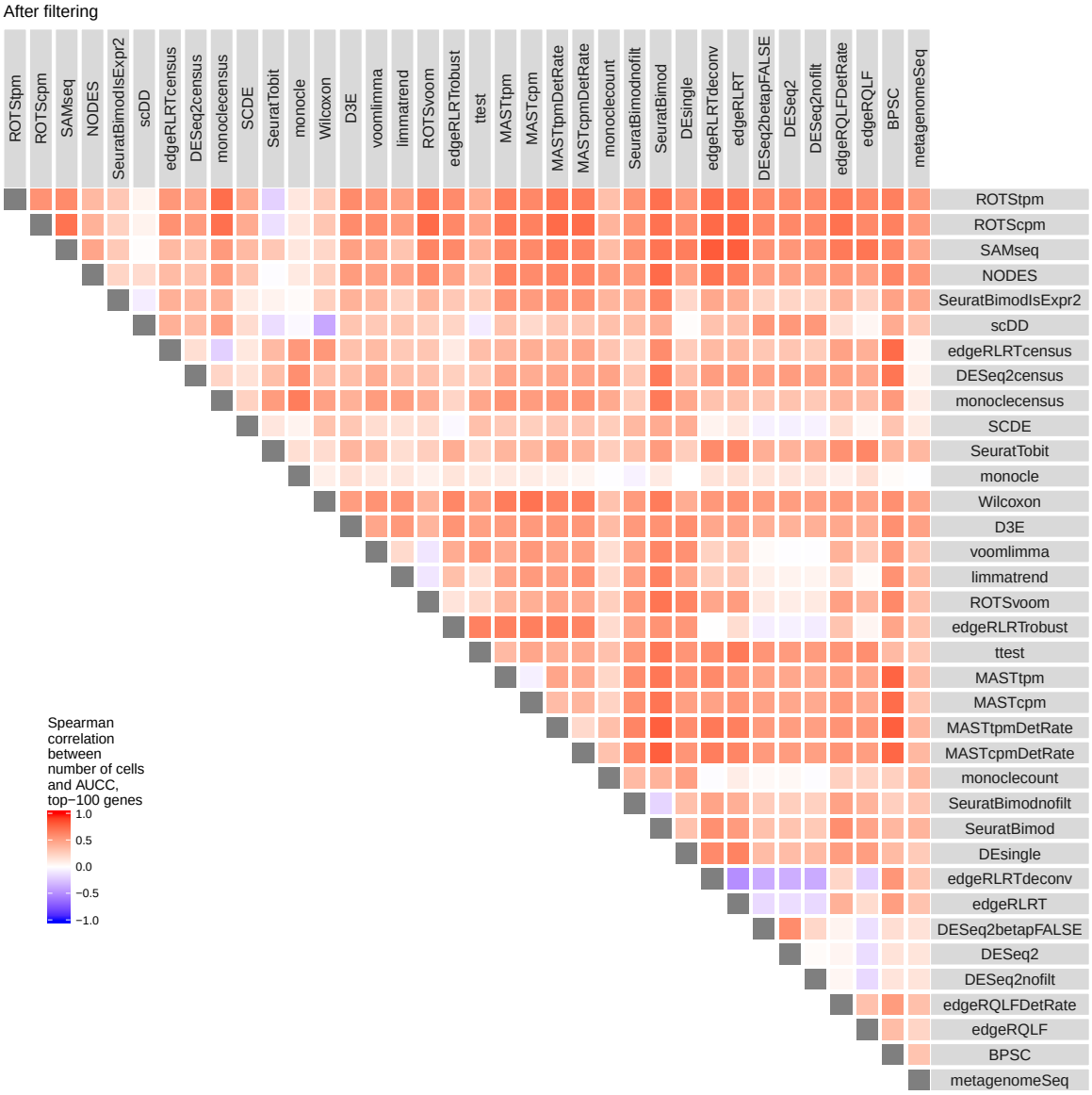
Supplementary Figure 18: Concordance scores for the individual real scRNA-seq signal and null datasets. A concordance curve is constructed between the rankings of each method obtained for each pair of dataset instances with the same sample size, and the concordance score is defined as the partial area under the concordance curve, until $K = 100$, divided by the maximal possible value of $K^2/2$ to obtain relative areas between 0 and 1. Left panel: Area under the concordance curve for the signal datasets. Right panel: Area under the concordance curve for the null datasets. All methods are generally more self-consistent for the signal datasets than for the null datasets, but the differences between the methods are often small and as expected, the concordances depend strongly on the dataset. Most methods achieve the best performance for the GSE60749-GPL13112 and 10XMonoCytoT datasets where the signal is very strong, and the smallest difference between signal and null datasets is seen for the EMTAB2805 dataset, where the signal is much less pronounced (Supplementary Figures 1-2). Methods are ordered by decreasing median difference between signal and null concordance scores across all datasets. Interestingly, non-parametric methods (Wilcoxon test, SAMseq) and methods based on log-like transformations of data (limma-trend, ROTSvoom, voom-limma) show a slight advantage over many count-based methods in terms of the median concordance difference between signal and null datasets, which may indicate that count-based methods are more sensitive to small perturbations of the scRNA-seq data. 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 pairs.

Nature Methods: doi:10.1038/nmeth.4612



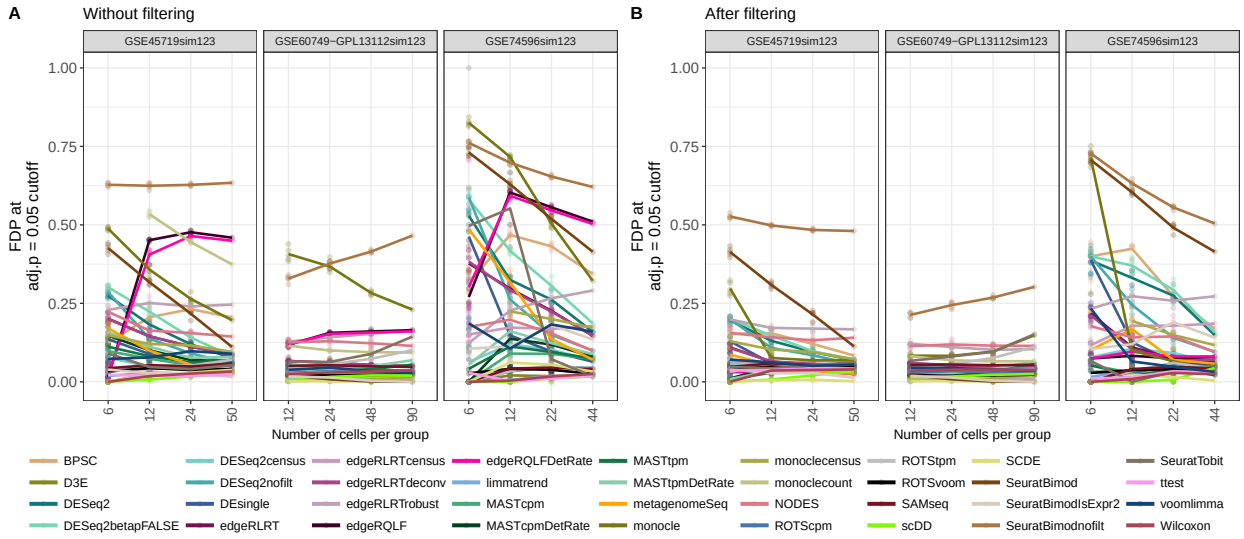
Supplementary Figure 19: Distribution of the area under the partial concordance curve (until $K = 100$) for each pair of methods, across all dataset instances. Dark colors indicate low concordance and yellow indicates high concordance, and each cell shows the concordance scores across all considered dataset instances.

Nature Methods: doi:10.1038/nmeth.4612



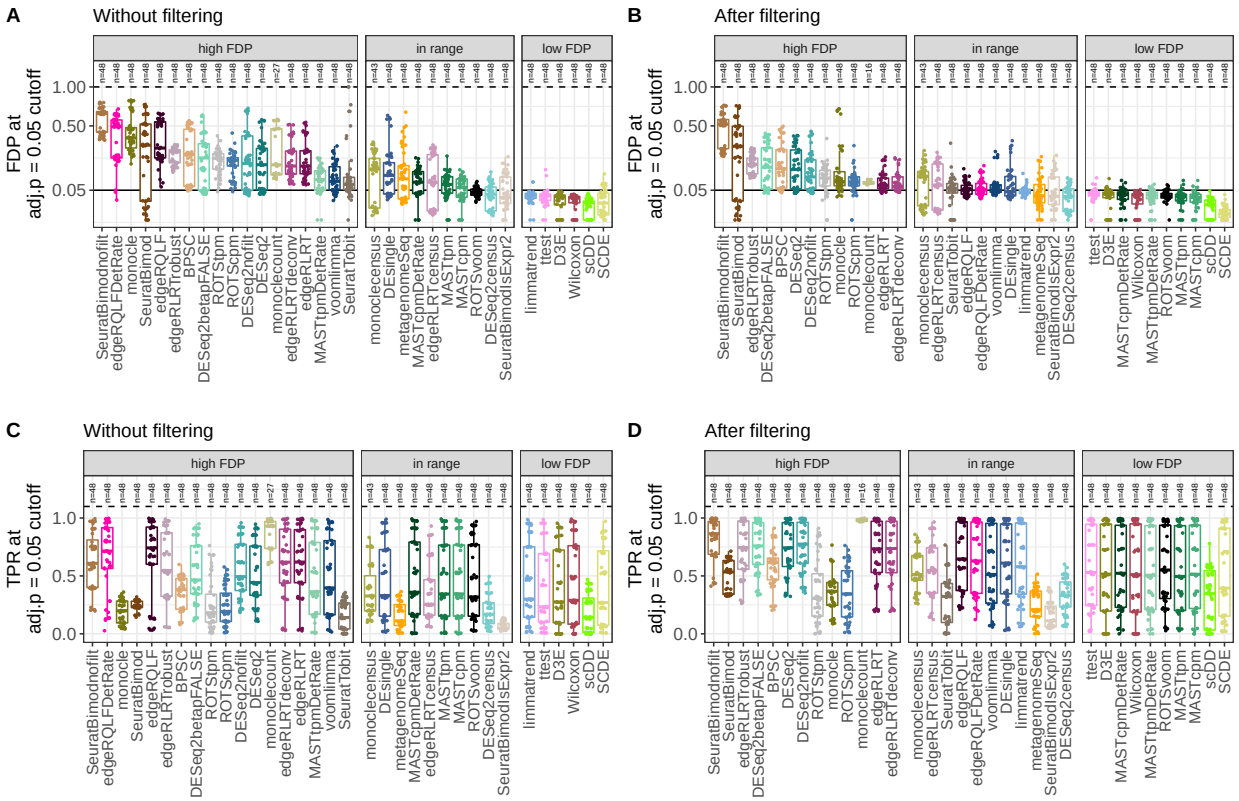
Supplementary Figure 20: Spearman correlation between the number of cells per group and the partial area under the concordance curve (until $K = 100$) for each pair of methods, across all instances of the real signal scRNA-seq datasets.

Nature Methods: doi:10.1038/nmeth.4612

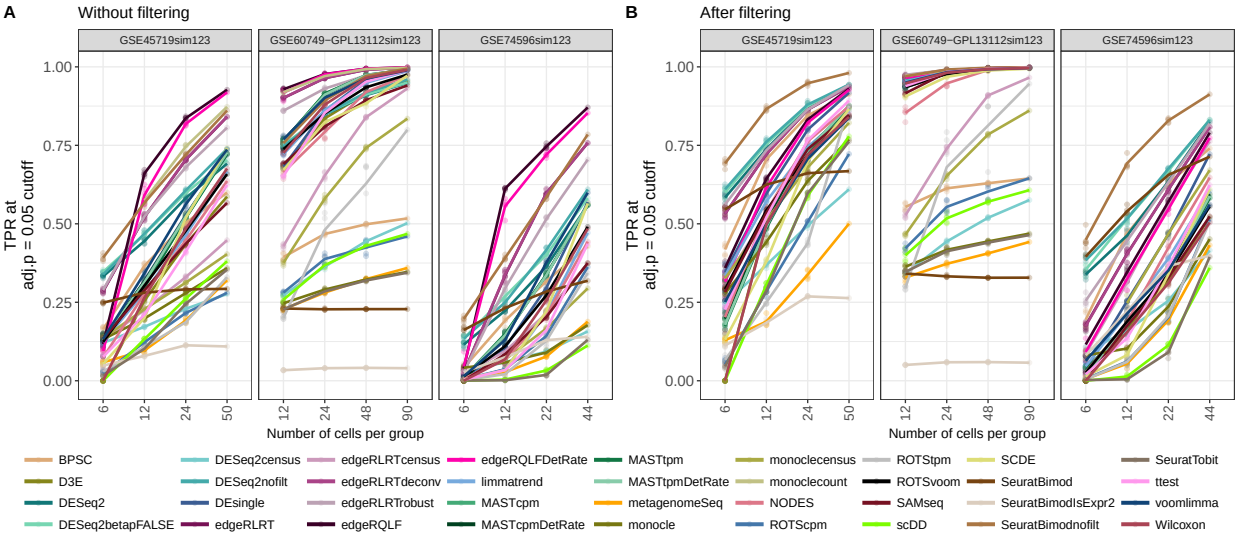


Supplementary Figure 21: False discovery proportion (FDP) at an adjusted p-value cutoff at 0.05 across the three simulated datasets, stratified by dataset and shown as a function of the number of cells per group. A. Without gene prefiltering. B. After gene prefiltering, retaining only genes with an estimated expression exceeding 1 TPM in more than 25% of the cells.

Nature Methods: doi:10.1038/nmeth.4612

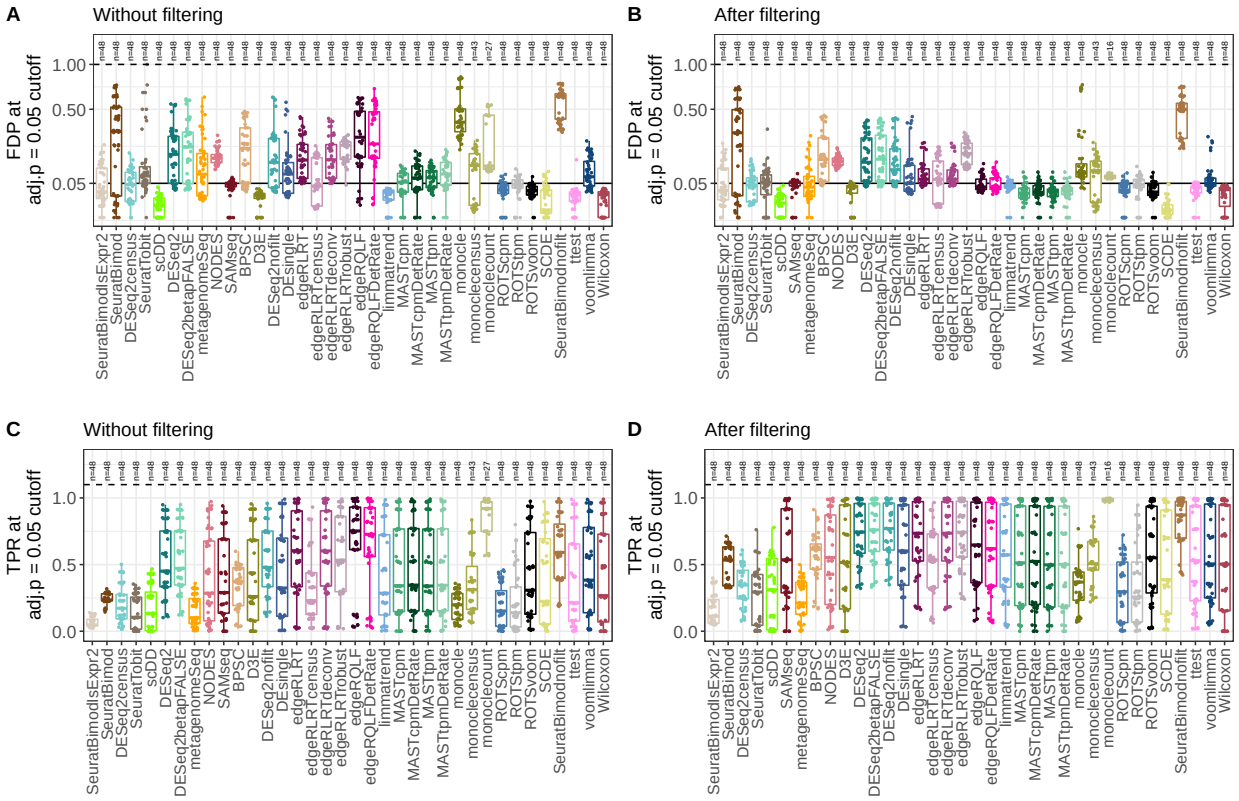


Supplementary Figure 22: Observed FDP and TPR at an adjusted p-value cutoff of 0.05 across the three simulated datasets. Adjusted p-values were calculated from nominal p-values using IHW [49], using the average expression as the covariate. Methods are ordered by decreasing median FDP across all dataset instances. 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.

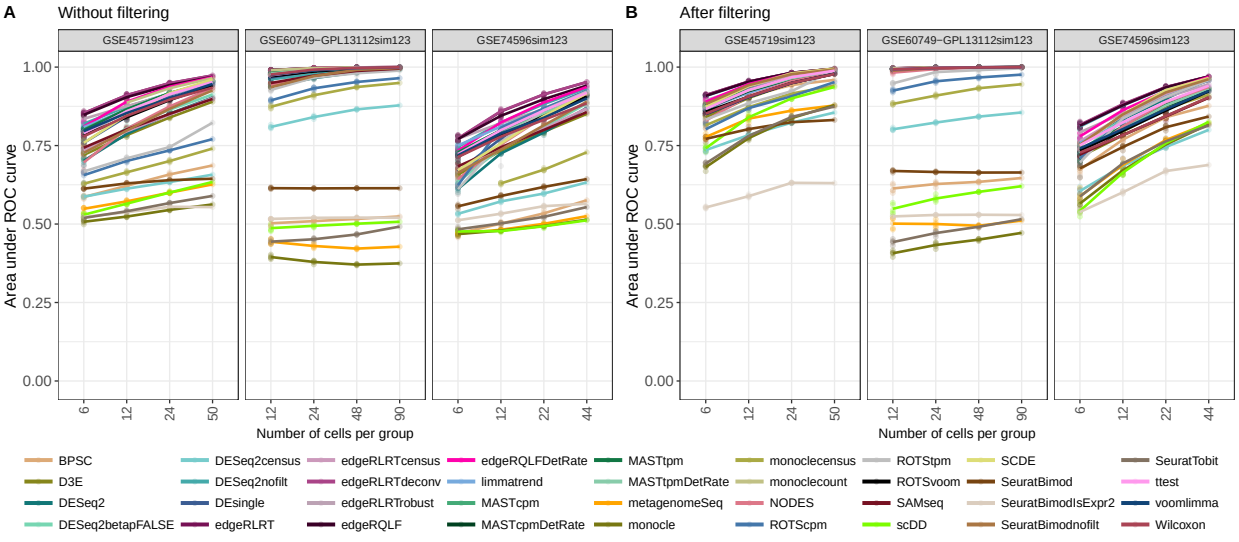


Supplementary Figure 23: Observed TPR at an adjusted p-value cutoff at 0.05 across the three simulated datasets, stratified by dataset and shown as a function of the number of cells per group. A. Without gene prefiltering. B. After gene prefiltering, retaining only genes with an estimated expression exceeding 1 TPM in more than 25% of the cells.

Nature Methods: doi:10.1038/nmeth.4612

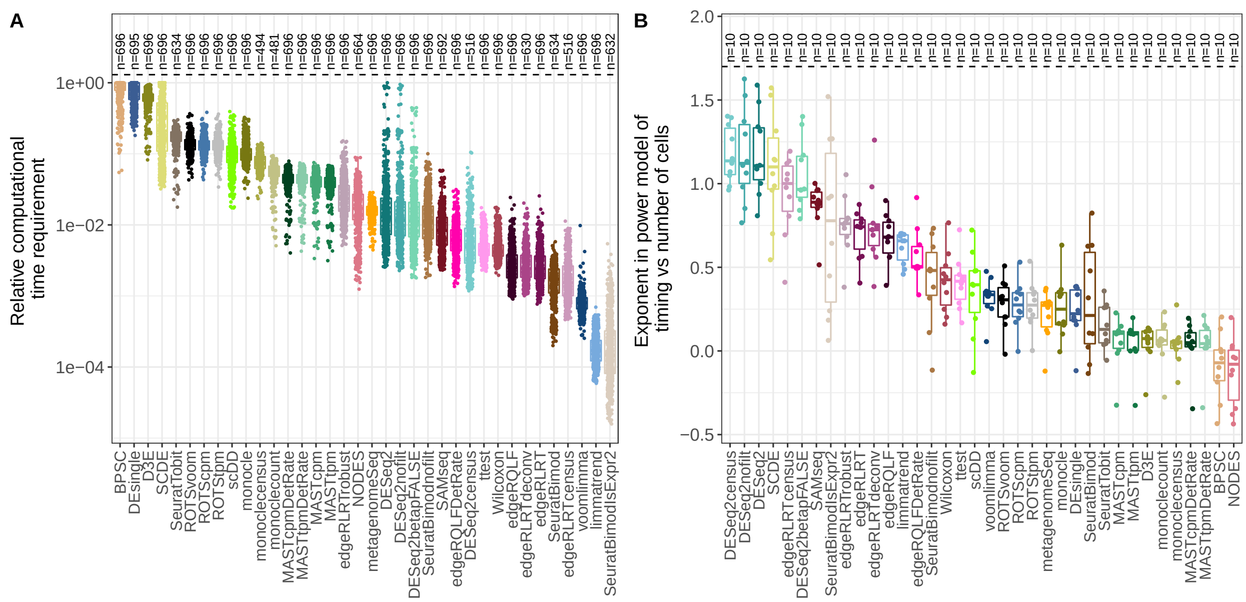


Supplementary Figure 24: Observed FDP and TPR at an adjusted p-value cutoff at 0.05 across the three simulated datasets, with methods ordered by decreasing fraction of NA adjusted p-values from left to right. Center line, median; hinges, first and third quartiles; whiskers, most extreme values within 1.5 IQR from the box; n, number of data set instances.



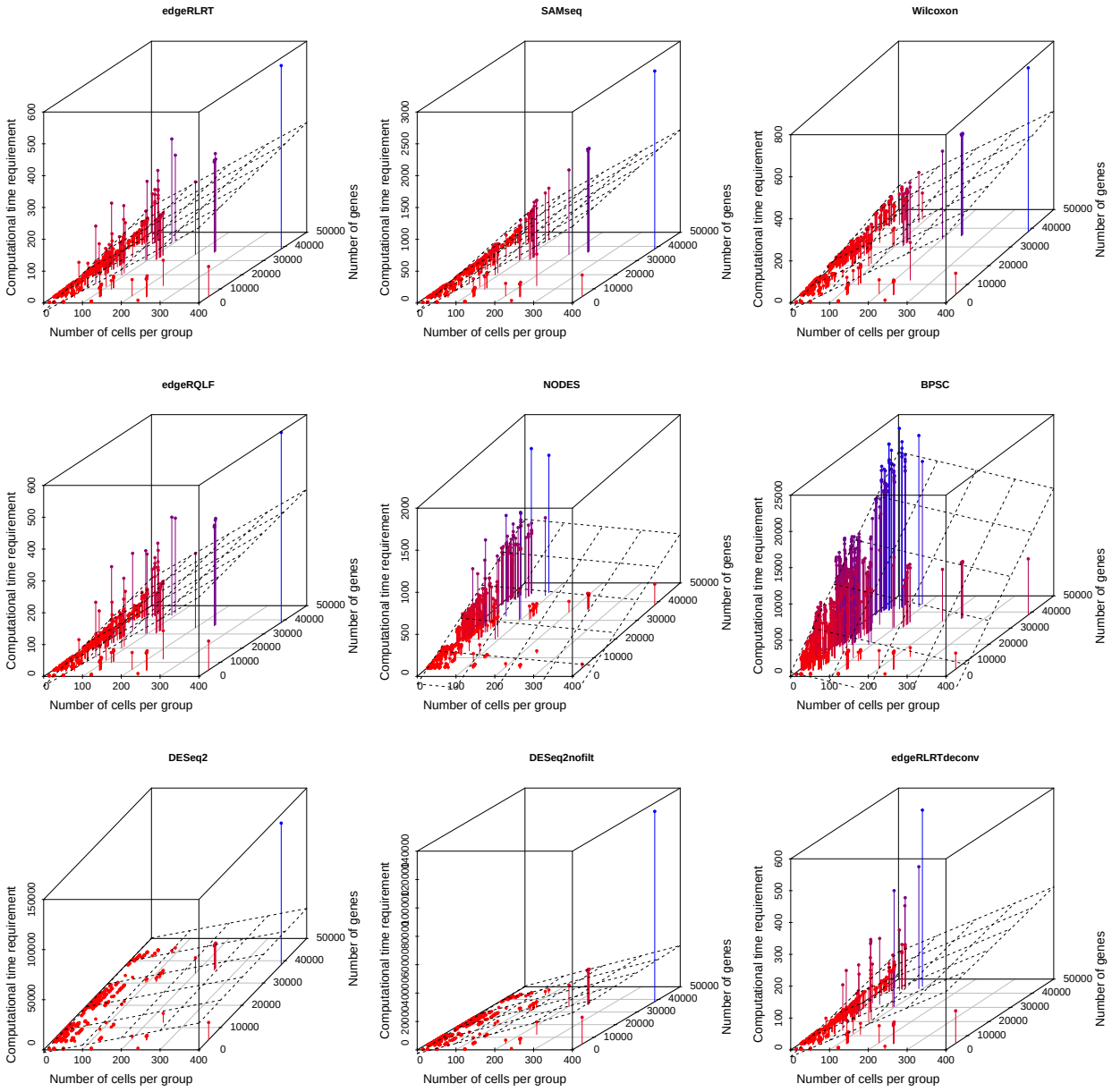
Supplementary Figure 25: Area under the ROC curve across the three simulated datasets, stratified by dataset and shown as a function of the number of cells per group. A. Without gene prefiltering. B. After gene prefiltering, retaining only genes with an estimated expression exceeding 1 TPM in more than 25% of the cells.

Nature Methods: doi:10.1038/nmeth.4612



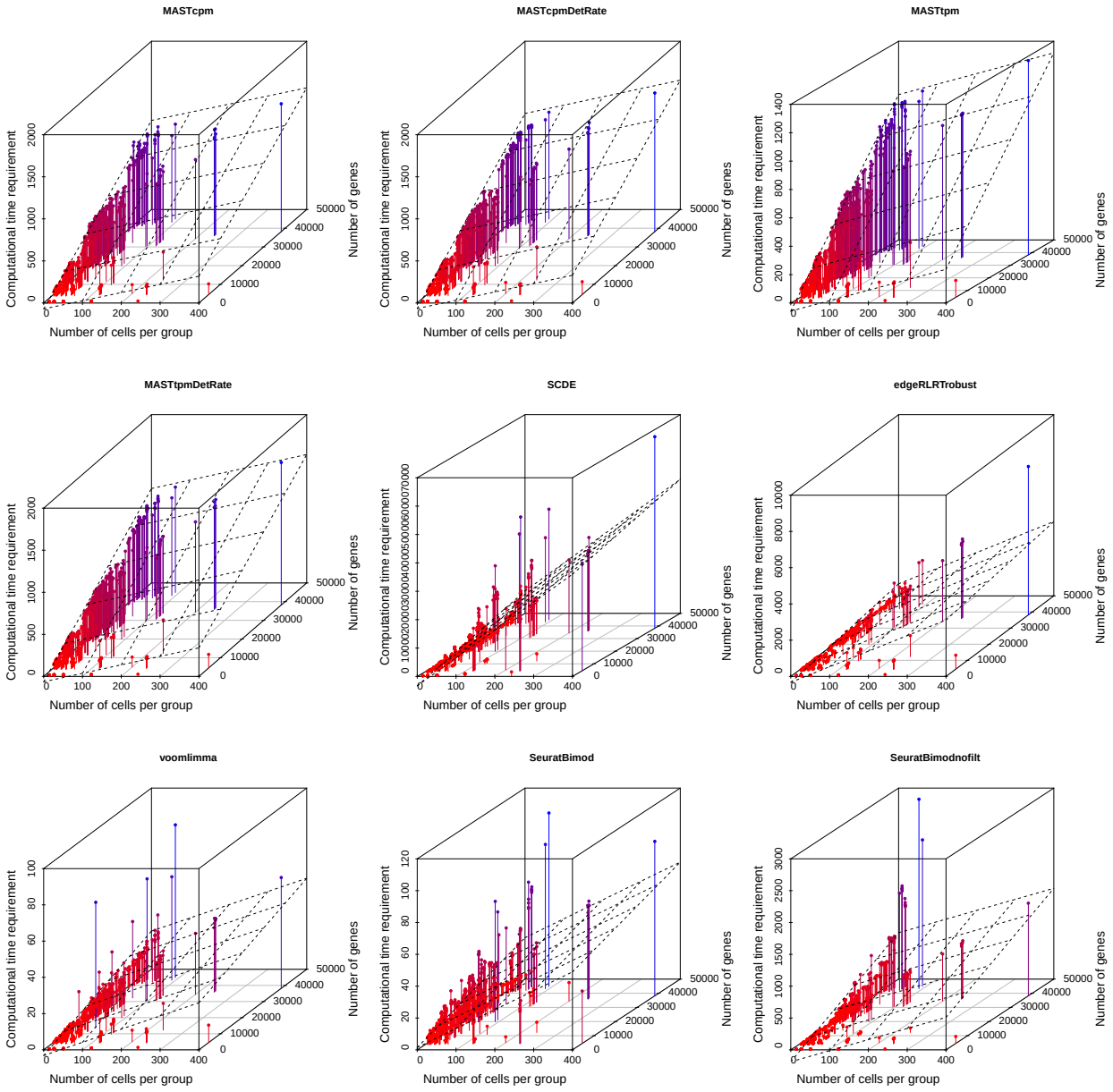
Supplementary Figure 26: Summary of the computational time requirement for all evaluated methods. To make values comparable across datasets with different number of cells and genes, we scale the times for each dataset instance relative to the slowest method for that particular instance to get relative time requirements. All methods were run on a single core. In each panel, the methods are ordered by the median value of the illustrated statistic. A. Relative computational time requirement for the different methods across all scRNA-seq datasets. The y-axis is on a log scale for improved visibility. n, number of data set instances. B. Exponent in a power model relating computational time requirement to the number of cells for datasets instances with similar number of genes. Center line, median; hinges, first and third quartiles; whiskers, most extreme values within 1.5 IQR from the box; n, number of dataset groups (where each group contains dataset instances with similar number of genes)

Nature Methods: doi:10.1038/nmeth.4612



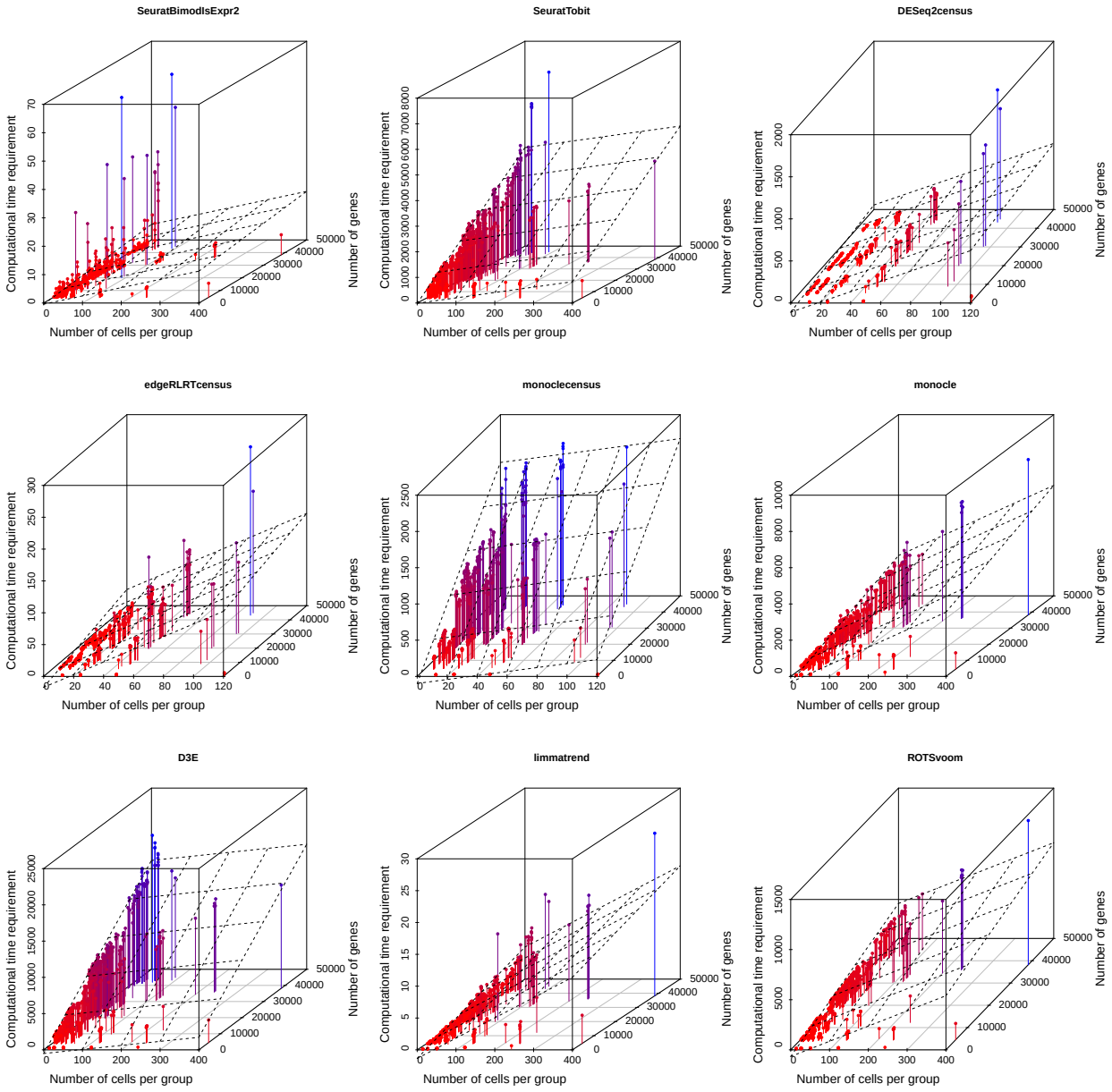
Supplementary Figure 27: Dependence of computational time requirement on number of genes and cells per group for each method, across all instances of the real and simulated scRNA-seq datasets. The grid indicates the best fitting plane for the data points. (continued in the following figures).

Nature Methods: doi:10.1038/nmeth.4612



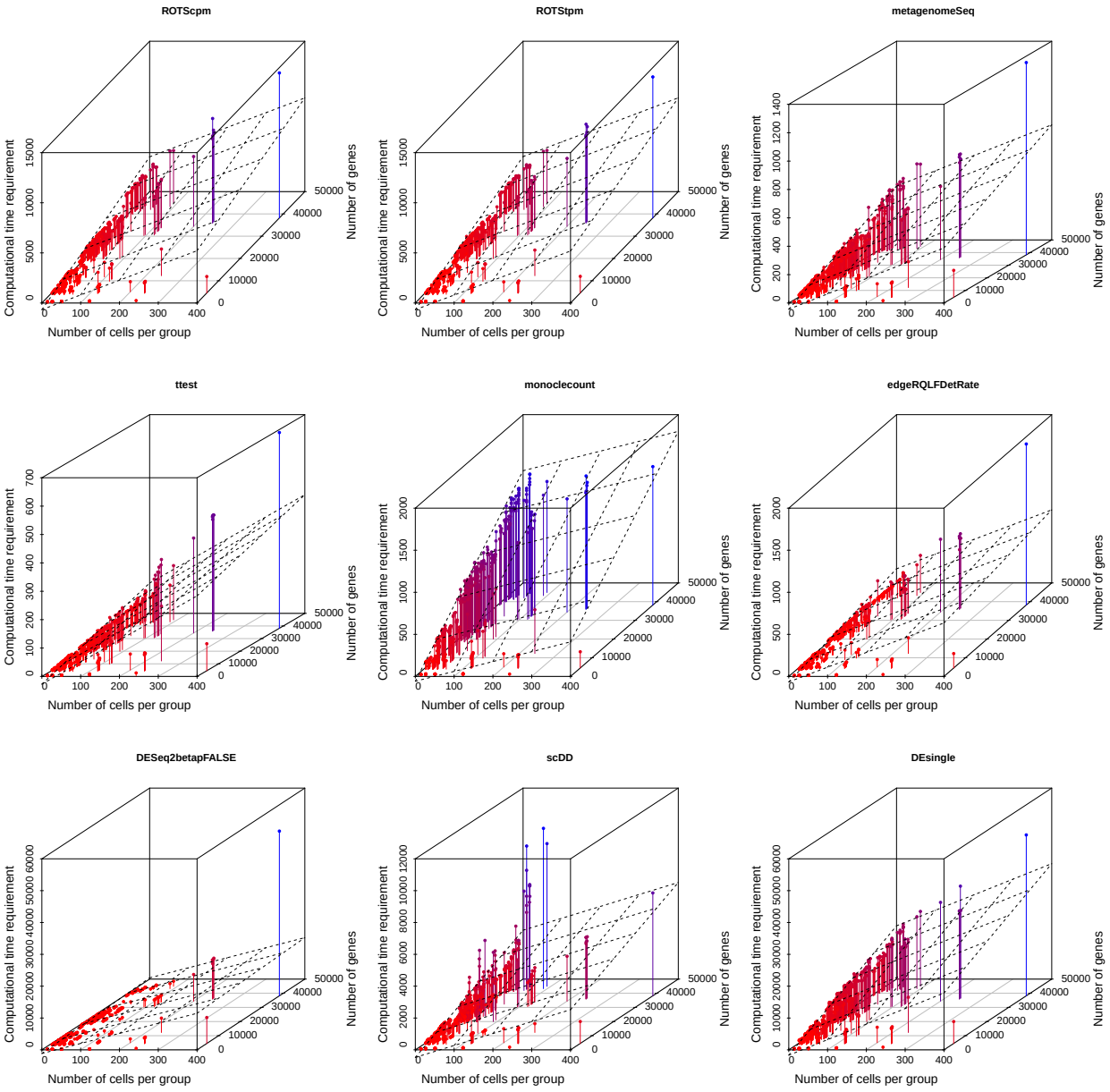
Supplementary Figure 28: Dependence of computational time requirement on number of genes and cells per group for each method, across all instances of the real and simulated scRNA-seq datasets. The grid indicates the best fitting plane for the data points.

Nature Methods: doi:10.1038/nmeth.4612



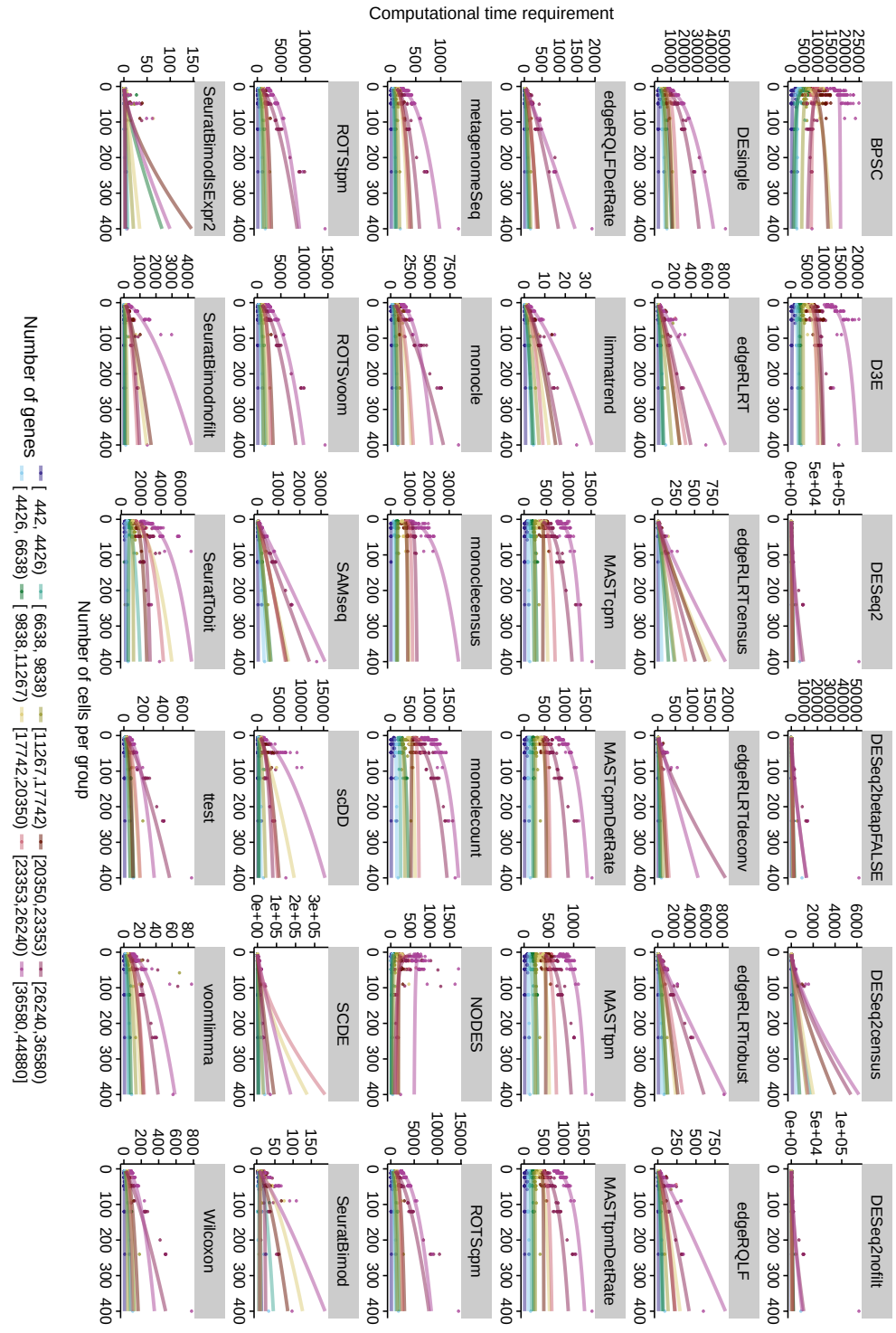
Supplementary Figure 29: Dependence of computational time requirement on number of genes and cells per group for each method, across all instances of the real and simulated scRNA-seq datasets. The grid indicates the best fitting plane for the data points.

Nature Methods: doi:10.1038/nmeth.4612



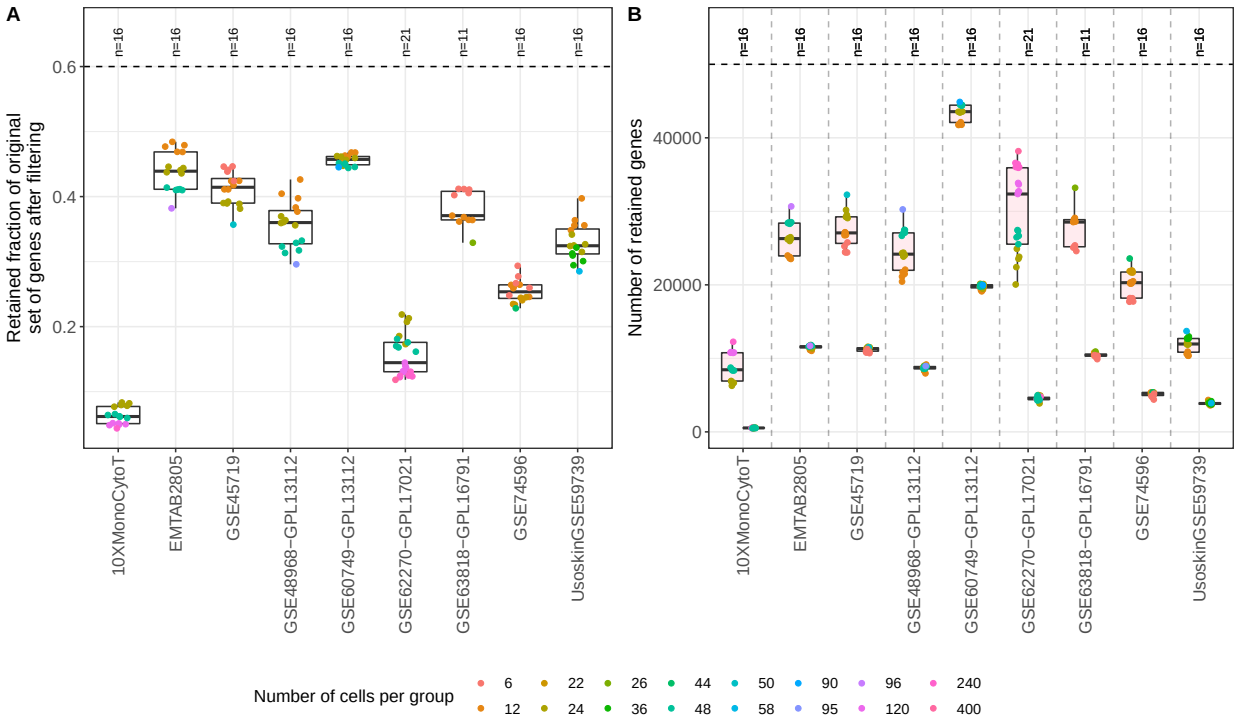
Supplementary Figure 30: Dependence of computational time requirement on number of genes and cells per group for each method, across all instances of the real and simulated scRNA-seq datasets. The grid indicates the best fitting plane for the data points.

Nature Methods: doi:10.1038/nmeth.4612



Supplementary Figure 31: Illustration of the required computational time (T) as a function of the number of cells per group (N), stratified by the approximate number of genes in the dataset. The curves represent fitted models of the form $T = aN^p$.

Nature Methods: doi:10.1038/nmeth.4612



Supplementary Figure 32: A. The fraction of genes remaining in each real scRNA-seq dataset instance after filtering, retaining only genes with an estimated abundance exceeding 1 TPM in more than 25% of the cells. Dataset instances with different number of cells per group are indicated in different colors. B. The number of genes in each dataset, without filtering (left boxplot for each dataset, shaded in light pink) and after filtering (right boxplot for each dataset). 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.

Nature Methods: doi:10.1038/nmeth.4612

References

1. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343,** 193–196 (2014).
2. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **509,** 363–369 (2014).
3. Kumar, R. M. *et al.* Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* **516,** 56–61 (2014).
4. Engel, I. *et al.* Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. *Nat. Immunol.* **17,** 728–739 (2016).
5. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33,** 155–160 (2015).
6. Usoskin, D. *et al.* Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* **18,** 145–153 (2015).
7. Guo, F. *et al.* The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. *Cell* **161,** 1437–1452 (2015).
8. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525,** 251–255 (2015).
9. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8,** 14049 (2017).
10. 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).
11. Vu, T. N. *et al.* Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32,** 2128–2135 (2016).
12. 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).
13. 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).
14. Miao, Z. & Zhang, X. *DEsingle: A new method for single-cell differentially expressed genes detection and classification* 2017.
15. 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).
16. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25 (2010).
17. 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).
18. 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).
19. Zhou, X., Lindsay, H. & Robinson, M. D. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.* **42,** e91 (2014).
20. Lun, A. T. L., Chen, Y. & Smyth, G. K. *Statistical Genomics* (eds Mathé, E. & Davis, S.) 391–416 (Springer New York, 2016).
21. 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).
22. 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).

23. 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).

24. 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).

25. 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).

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

27. Sengupta, D., Rayan, N. A., Lim, M. & Lim, B. Fast, scalable and accurate differential expression analysis for single cells. *bioRxiv doi:10.1101/049734,* 1–9 (2016).

28. 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).

29. 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).

30. 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).

31. Korthauer, K. D. *et al.* A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* **17,** 222 (2016).

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

33. 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).

34. McDavid, A. *et al.* Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29,** 461–467 (2013).

35. Welch, B. A. The generalization of Student's problem when several different population variances are involved. *Biometrika* **34,** 28–35 (1947).

36. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin* **1,** 80–83 (1945).

37. 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).

38. Srivastava, A., Sarkar, H., Gupta, N. & Patro, R. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics* **32,** i192–i200 (2016).

39. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

40. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. en. *Nat. Methods* (2017).

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

42. Ramos, M. *et al.* Software For The Integration Of Multi-Omics Experiments In Bioconductor. *bioRxiv doi:10.1101/144774* (2017).

43. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics,* btw777 (2017).

44. Maaten, L. v. d. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).

45. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20,** 53–65 (1987).

46. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19,** 716–723 (1974).

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

48. Cleveland, W. S. Robust Locally Weighted Regression and Smoothing Scatterplots. *J. Am. Stat. Assoc.* **74,** 829–836 (1979).

49. 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).