On the discovery of subpopulation-specific state transitions from multi-sample multi-condition single-cell RNA sequencing data

Helena L. Crowell[1,2], Charlotte Soneson[1,2,3,*], Pierre-Luc Germain[1,4,*], Daniela Calini[5], Ludovic Collin[5], Catarina Raposo[5], Dheeraj Malhotra[5] & Mark D. Robinson[1,2]

[1] *Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland*

[2] *SIB Swiss Institute of Bioinformatics, Zurich, Switzerland*

[3] *Present address: Friedrich Miescher Institute for Biomedical Research and SIB Swiss Institute of Bioinformatics, Basel, Switzerland*

[4] *D-HEST Institute for Neuroscience, Swiss Federal Institute of Technology, Zurich, Switzerland*

[5] *F. Hoffmann-La Roche Ltd, Pharma Research and Early Development, Neuroscience, Ophthalmology and Rare Diseases, Roche Innovation Center Basel, Basel, Switzerland*

[*] *These authors contributed equally.*

Abstract

Single-cell RNA sequencing (scRNA-seq) has quickly become an empowering technology to profile the transcriptomes of individual cells on a large scale. Many early analyses of differential expression have aimed at identifying differences between subpopulations, and thus are focused on finding subpopulation markers either in a single sample or across multiple samples. More generally, such methods can compare expression levels in multiple sets of cells, thus leading to cross-condition analyses. However, given the emergence of replicated multi-condition scRNA-seq datasets, an area of increasing focus is making sample-level inferences, termed here as differential state analysis. For example, one could investigate the condition-specific responses of cell subpopulations measured from patients from each condition; however, it is not clear which statistical framework best handles this situation. In this work, we surveyed the methods available to perform cross-condition differential state analyses, including cell-level mixed models and methods based on aggregated "pseudobulk" data. We developed a flexible simulation platform that mimics both single and multi-sample scRNA-seq data and provide robust tools for multi-condition analysis within the muscat R package.

Introduction

A fundamental task in the analysis of single-cell RNA-sequencing (scRNA-seq) data is the identification of systematic transcriptional changes using differential expression analysis[1]. Such analyses are a critical step toward a deeper understanding of molecular responses that occur in development, after a perturbation or in disease states[2,3,4,5]. Most current scRNA-seq differential expression methods are designed to test one set of cells against another (or more generally, multiple sets together), and can be used to compare cell subpopulations (e.g., for identifying marker genes) or across conditions (cells from one condition versus another)[6]. In such statistical models, the cells are the experimental units and thus represent the population that inferences will extrapolate to.

Given the rise of multi-sample multi-group scRNA-seq datasets, where measurements are made on hundreds to thousands of cells per sample, the goal shifts to making sample-level inferences (i.e., experimental units are samples), in order to account for sample-to-sample as well as cell-to-cell variability and make conclusions that extrapolate to the samples rather than cells. We refer to this generally as differential state (DS) analysis, whereby a given subset of cells (termed hereafter as subpopulation) is followed across a set of samples (e.g., individuals) and experimental conditions (e.g., treatments), in order to identify subpopulation-specific responses, i.e., changes in cell state. DS analysis: i) should be able to detect changes that only affect a single cell subpopulation, a subset of subpopulations or even a subset of cells within a single subpopulation; ii) is intended to be an orthogonal analysis to clustering or cell subpopulation assignment; and, iii) can be considered a separate analysis to the search for differential abundance of subpopulations across conditions.

We intentionally use the term *subpopulation* to be more generic than cell *type*[7,8], which itself is meant to represent a discrete and stable molecular signature; however, the precise definition of cell type is widely debated[2,3]. In our framework, a subpopulation is simply a set of cells deemed to be similar enough to be considered as a group and where it is of interest to interrogate such sets of similarly-defined cells across multiple samples and conditions. Therefore, cells from a scRNA-seq experiment are first organized into subpopulations, e.g., by integrating the multiple samples together[9] and clustering or applying a subpopulation-level assignment algorithm[10] or cell-level prediction[11]; clustering and manual annotation is also an option. Regardless of the mode or the uncertainty in subpopulation assignment, the discovery framework we describe provides a basis for biological interpretation and a path to discovering interesting expression patterns *within* subpopulations across samples. Even different subpopulation assignments of the same data could be readily interpretable. For example, T cells could be defined as a single (albeit diverse) cell subpopulation or could be divided into discrete subpopulations, if sufficient information to categorize the cells at this level of resolution is available. In either case, the framework presented here would focus on the subpopulation of interest and look for expression changes *across* conditions. This naturally introduces an interplay with the definition of cell types and states themselves (e.g., discrete states could be considered as types) and thus with the methods used to computationally or manually classify cells. Overall, our goal here is to explore the space of scRNA-seq datasets with several subpopulations and samples, in order to understand the fidelity of methods to discover cell state changes.

It is worth noting that extensive workflows for DS analysis of high-dimensional cytometry data have been established[12,13,14,15], along with a rich set of visualization tools and differential testing methods[16,17,13,18], and applied to, for example, unravel subpopulation-specific responses to immunotherapy[19]. Notably, aggregation-based methods (e.g., representing each sample as the median signal from all cells of a given subpopulation) compare favorably in (cytometry) DS analysis to methods that run on full cell-level data[17]; however, in the cytometry case, only a limited range of cell-level and aggregation approaches were tested, only simplistic regimes of differential expression were investigated (e.g., shifts in means), and the number of features measured with scRNA-seq is considerably higher (with typically fewer cells).

In scRNA-seq data, aggregating cell-level counts into sample-level "pseudobulk" counts for differential expression is not new; pseudobulk analysis has been applied to discover cell-type specific responses of lupus patients to IFN-β stimulation[20] and in mitigating plate effects by summing read counts in each plate[21]. In these cases, pseudobulk counts were used as input to bulk RNA-seq differential engines, such as edgeR[22], DESeq2[23] or limma-voom[24,25]. Also, non-aggregation methods have been proposed, e.g., mixed models were previously used on cell-level scRNA-seq expression data[26] to separate sample and batch effects, and variations on such a mixed model could be readily applied for the sample-level inferences that are considered here. Various recent related developments have taken place: a compositional model was proposed to integrate cell type information into differential analysis, although replication was not considered[27]; a multivariate mixed effects model was proposed to extend univariate testing regimes[28]; and, a tool called PopAlign was introduced to estimate low-dimensional mixtures and look for state shifts from the parameters of the mixture distributions[29]. Ultimately, there is scope for alternative methods to be applied to the discovery of interesting single-cell state changes.

In existing comparison studies of scRNA-seq differential detection methods[30,6,31], analyses were limited to comparing groups of cells and had not explicitly considered sample-level inferences or aggregation approaches. The rapid uptake of new single-cell technologies has driven the collection of scRNA-seq datasets across multiple samples. Thus, it remains to be tested whether existing methods designed for comparing expression in scRNA-seq data are adequate for such cross-sample comparisons, and in particular, how sensitive aggregation methods are to detect subpopulation-level responses.

In this study, we developed a simulation framework, which is anchored to a reference dataset, that mimics various characteristics of scRNA-seq data and used it to evaluate 16 DS analysis methods (see **Supplementary Table 1**) across a wide range of simulation scenarios, such as varying the number of samples, the number of cells per subpopulation, and the magnitude and type of differential expression pattern introduced. We considered two conceptually distinct representations of the data for each subpopulation, cell-level or sample-level, and from these, made sample-level inferences. On cell-level data, we applied: i) mixed models (MM) with a fixed effect for the experimental condition and a random effect for sample-level variability; ii) approaches comparing full distributions (e.g., K-sample Anderson-Darling test[32]); and, as a reference point, we applied well-known scRNA-seq methods, such as scDD[33] and MAST[34], although these methods were not specifically intended for the across-sample situation. Alternatively, we assembled sample-level data by aggregating measurements for each subpopulation (for each sample) to obtain pseudobulk data in several ways; we then leveraged established bulk RNA-seq analysis frameworks to make sample-level inferences.

All methods tested are available within the muscat R package and a Snakemake[35] workflow was built to run simulation replicates. Since discovery of state changes in cell subpopulations is an open area of research, anchor datasets are openly available via Bioconductor's ExperimentHub, to facilitate further bespoke method development.

Using existing pipelines for integrating, visualizing, clustering and annotating cell subpopulations from a replicated multi-condition dataset of mouse cortex, we applied pseudobulk DS analysis to unravel subpopulation-specific responses within brain cortex tissue from mice treated with lipopolysaccharide.

Results

Simulation framework. To explore the various aspects of DS analysis, we developed a straightforward but effective simulation framework that is anchored to a labeled multi-sample multi-subpopulation scRNA-seq reference dataset, and exposes parameters to modulate: the number of subpopulations and samples simulated, the number of cells per subpopulation (and sample), and the type and magnitude of a wide range of patterns of differential expression. Using (non-zero-inflated) negative binomial (NB) as the canonical distribution for droplet scRNA-seq datasets[6,36], we first estimate subpopulation- and sample-specific means, dispersion and library size parameters from the reference data set (see **Figure 1a**). Baseline multi-sample simulated scRNA-seq data can then be simulated also from a NB distribution, by sampling from the subpopulation/sample-specific empirical distributions of the mean, dispersion and library size. To this baseline, genes can be selected as subpopulation-specific (i.e., mean different in one subpopulation versus the others), or as a state gene (differential expression introduced in the samples from one condition), or neither (equal relative expression across all samples and subpopulations). To introduce changes in expression that represent a change in cell state, we follow the differential distribution approach of Korthauer *et al.*[33], adding changes in the mean expression (DE), changes in the proportions of low and high expression-state components (DP), differential modality (DM) or changes in both proportions and modality (DB). Genes that are not subject to state changes are either equivalently expressed (EE), or expressed at low and high expression-states by an equal proportion (EP) of cells in both conditions; see **Figure 1b**. Here, the changes are added to samples in a condition-specific manner, thus mimicking a subpopulation-specific state change amongst replicates of one condition.

As reference datasets, we used i) scRNA-seq data of Peripheral Blood Mononuclear Cells (PBMCs) from 8 lupus patients measured before and after 6h-treatment with IFN-β (16 samples in total)[20], where cells were already annotated into various immune subpopulations; and, ii) single-nuclei RNA-seq data of brain cortex tissue from 8 mice split into a vehicle and lipopolysaccharide treatment group. In order to introduce known state changes, simulations were based only on control and vehicle samples, respectively. Importantly, our simulation framework is able to reproduce important characteristics of individual scRNA-seq datasets (e.g., mean-dropout and mean-variance relationships) from a $\mathrm{countsimQC}$[37] analysis (see **Supplementary File 1**) as well as sample-to-sample variability, as illustrated by pseudobulk-level dispersion-mean trends (**Supplementary Fig. 1a**). By varying the proportion of subpopulation-specific and DS genes, we are able to generate multiple subpopulations that are distinct but proximal, and clearly separated from one another in lower-dimensional space (**Fig. 1c**); in particular, parameters control the distinctness of each subpopulation and of the group-wise state changes. Subpopulation-specific log-fold-changes (logFCs) further allow modulating differential expression to be of equal magnitude across all subpopulations, or such that a given subpopulation exhibits a weakened (logFC < 2), amplified (logFC > 2), or null (logFC $= 0$) differential signal relative to the default (logFC $= 2$; see **Figure 1c**). Taken together, we constructed a simulation that replicates aspects of individual scRNA-seq datasets, mimics sample-to-sample variability and offers a high level of flexibility to introduce subpopulation-specific identities (e.g., via marker genes) as well as condition-specific state changes.

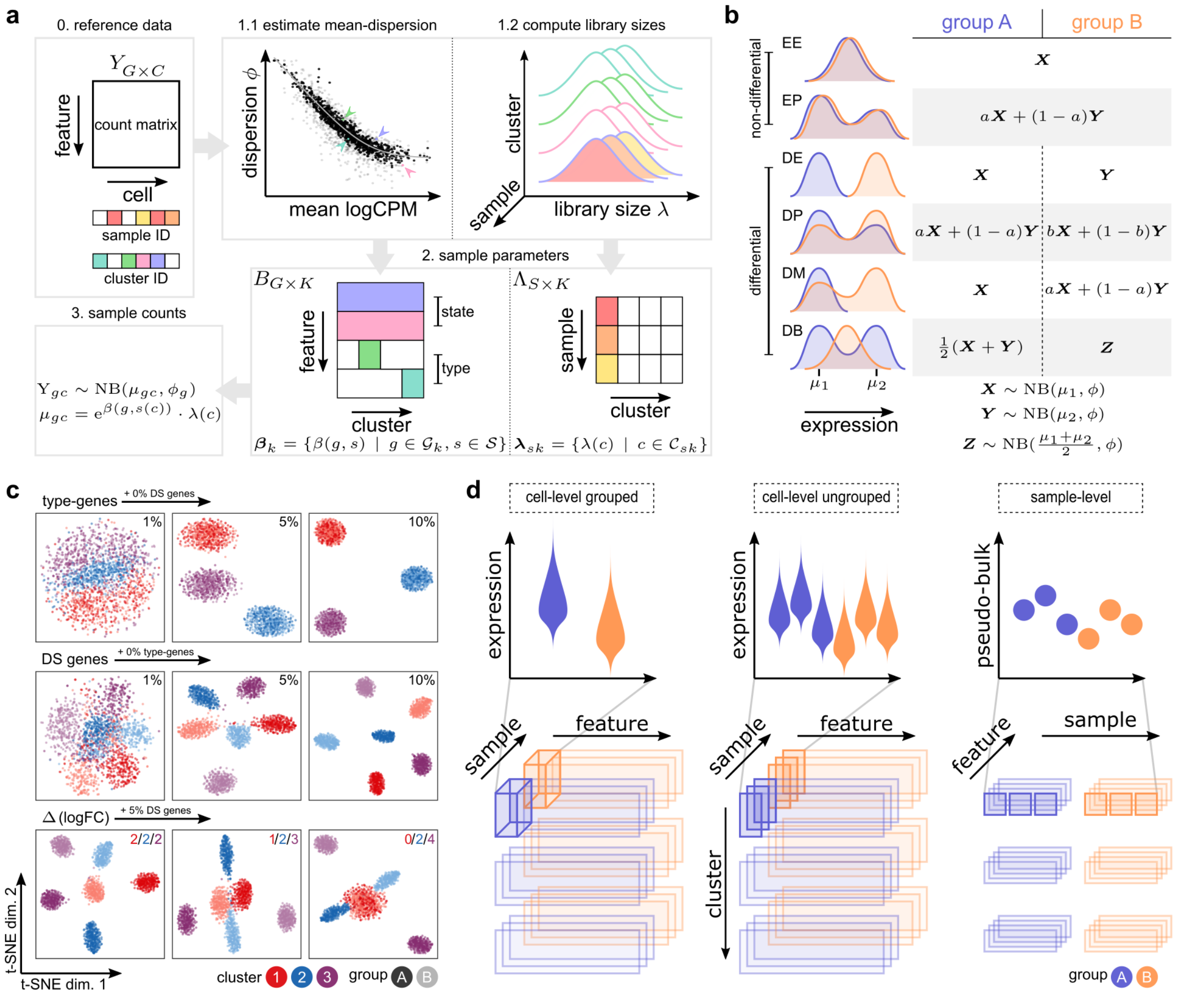

Figure 1: Schematic overview of muscat's simulation framework. (**a**) Given a count matrix of features by cells and, for each cell, pre-determined cluster (subpopulation) identifiers as well as sample labels (0), dispersion and sample-wise means are estimated from a negative binomial distribution for each gene (for each subpopulation) (1.1); and library sizes are recorded (1.2). From this set of parameters (dispersions, means, library sizes), gene expression is sampled from a negative binomial distribution. Here, genes are selected to be "type" (subpopulation-specifically expressed; e.g., via marker genes), "state" (change in expression in a condition-specific manner) or equally expressed (relatively) across all samples (2). The result is a matrix of synthetic gene expression data (3); (**b**) Differential distributions are simulated from a NB distribution or mixtures thereof, according to the definitions of random variables $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$. (**c**) t-SNE plots for a set of simulation scenarios with varying percentage of "type" genes (top), DS genes (middle), and difference in the magnitude (logFC) of DS between subpopulations (bottom). (**d**) Schematic overview of cell- and sample-level approaches for DS analysis. Top panels show a schematic of the data distributions or aggregates across samples (each violin is a group or sample; each dot is a sample) and conditions (blue or orange). The bottom panels highlight the data organization into sub-matrix slices of the original count table.

Aggregation versus non-aggregation methods. The starting point for a differential state analysis is a (sparse) matrix of gene expression, either as counts (with library or size factors) or normalized data (log-transformed expression values, residuals[38,39]), where each row is a gene and each column a cell. Each cell additionally has a subpopulation (cluster) label as well as a sample label; metadata should be linked to samples, such that they can be organized into comparable groups with sample-level replicates (e.g., via a design matrix). The data processing aspect, depending on whether to aggregate data to the subpopulation-sample level, is described in the schematic in **Figure 1d**. The methods presented here are modular and thus the subpopulation label could originate from an earlier step in the analysis, such as clustering[40,41,42] after integration[43,9] or after inference of cell-type labels at the subpopulation-[10] or cell-level[11]. The specific details and suitability of these various preprocessing steps is an active area of current research and a full evaluation of them is beyond the scope of the current work; a comprehensive review was recently made available[44].

For aggregation-based methods, we considered various combinations of input data (log-transformed expression values, residuals, counts), summary statistics (mean, sum), and methods for differential testing (limma-voom, limma-trend, edgeR) that are sensible from a methodological perspective. For example, limma-voom and edgeR operate naturally on pseudobulk counts, while we have also used limma-trend on the mean of log-transformed library-size-normalized counts (logcounts). MAST[34] was run on logcounts; Anderson-Darling (AD) tests[32] and scDD[33] on both logcounts and standardized residuals (vstresiduals)[38]. For the AD tests, we considered two distinct approaches to test for equal distributions, with alternative hypotheses having samples different either sample-wise or group-wise (see **Supplementary Table 1** and Methods).

Performance of differential state detection. First, we generated null simulations where no genes are truly differential (across conditions), to evaluate the ability of methods to control error rates (3 replicates in each of 2 conditions, $K = 2$ subpopulations). While various methods show mild departures from uniform (**Supplementary Fig. 2a**), the Anderson-Darling tests, regardless of whether they were run comparing groups or samples, deviated the furthest from uniform and were the most unstable across replicates.

To compare the ability of methods to detect DS genes, we simulated $S_1 = S_2 = 3$ samples across 2 conditions. To retain the empirical distribution of library sizes, we simulated the same number of genes as in the reference dataset, and selected a random subset of $G = 4{,}000$ genes for further analysis to reduce runtimes. We simulated $K = 3$ subpopulations and introduced 10% of genes with DS, with equal magnitude of differential expression across subpopulations ($\mathbb{E}[\text{logFC}] = 2$) and randomly assigned to genes across the range of expression strength.

To ensure that method performances are comparable and do not suffer from low cell numbers, we simulated an average of 200 cells per subpopulation-sample instance, amounting to a total of $\sim 200 \times (S_1 + S_2) \times K \approx 3{,}600$ cells per simulation. Each simulation and method was repeated 5 times per scenario, and performances were averaged across replicates.

In the context of DS analysis, each of the G genes is tested independently in each of K subpopulations, resulting in a total of $\sim G \times K$ differential tests (occasionally, a small number of genes are filtered out due to low expression). Multiple testing correction could thus, in principle, be performed globally, i.e., across all tests ($n = G \times K$), or locally, i.e., on each of the subpopulation-level tests ($n = G$). We compared overall False Discovery Rate (FDR) and True Positive Rate (TPR) estimates computed from both locally and globally adjusted p-values. Global p-value adjustment led to a systematic reduction of both FDRs and TPRs (**Fig. 2a**; stratified also by the type of DS) and is therefore very conservative.

Moreover, detection performance is related to expression level, with differences in lowly expressed genes especially difficult to detect (**Supplementary Fig. 4**). On the basis of these observations, for the remainder of this study, all method performances were evaluated using locally adjusted p-values, after exclusion of genes with a simulated expression mean below 0.1.

In general, all methods performed best for genes of the DE category, followed by DM, DP, and DB (**Fig. 2a**). This level of difficulty by DS type is to be expected, given that genes span the range of expression levels and imposing mixtures of expression changes (DM, DP) dampens the overall magnitude of change compared to DE. In particular, DB, where the means are not different in the two conditions, is particularly difficult to detect, especially at low expression; therefore, several methods, including most of those that analyze full distributions (Anderson-Darling, scDD), underperform in this situation. For example, the Anderson-Darling tests on vstresiduals show good sensitivity, but also result in unacceptably high FDRs. For DE, DM and DP, there is a set of methods that perform generally well, including most of the pseudobulk approaches and cell-level MM models. Aggregation- and MM-based methods also performed fairly consistent across simulation replicates, while other methods were generally more erratic in their performance (**Supplementary Fig. 3**).

Comparison of simulated and estimated logFC highlighted that MM-based methods and $\mathrm{limma\text{-}trend}$ applied to mean-logcounts systematically underestimate logFCs, with estimates falling close to zero for a large fraction of gene-subpopulation combinations (**Supplementary Fig. 5a**). Although the differential detection performance does not seem to be compromised, applying the logarithm transformation (with an offset to avoid zero) to the rather low counts of cell-level data attenuates the scale and thus the magnitude of the estimated logFCs. For the

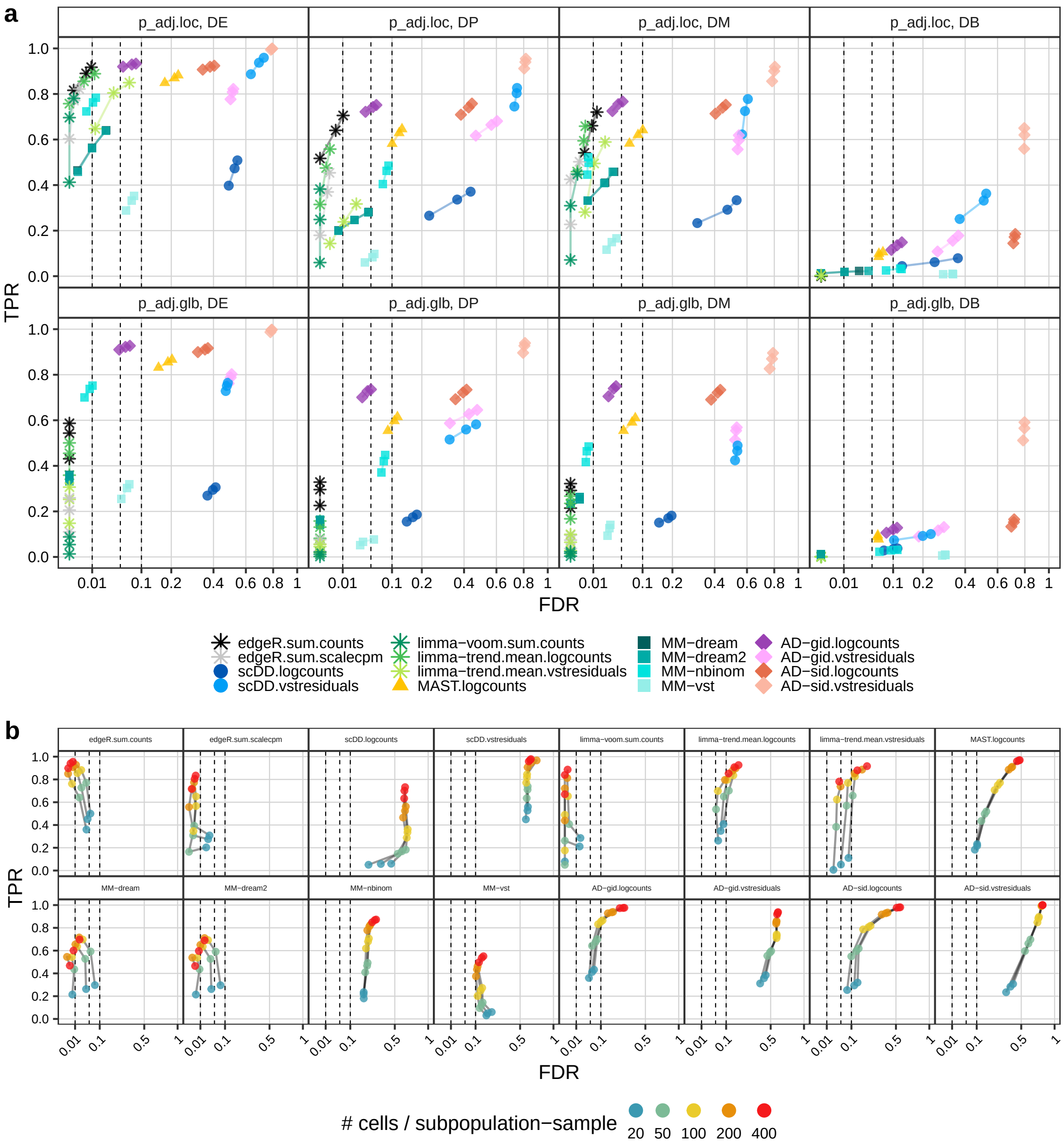

Figure 2: DS method performance across p-value adjustment types, differential distribution categories, and subpopulation-sample cell counts. All panels show observed overall true positive rate (TPR) and false discovery rate (FDR) values at FDR cutoffs of 1%, 5%, and 10%; dashed lines indicate desired FDRs (i.e., methods that control FDR at their desired level should be left of the corresponding dashed lines). For each panel, performances were averaged across 5 simulation replicates, each containing 10% of DS genes (of the type specified in the panel labels of **(a)**, and 10% of DE genes for **(b)**; see **Figure 1b** for further details). (**a**) Comparison of locally and globally adjusted p-values, stratified by DS type. Performances were calculated from subpopulation-level (locally) adjusted p-values (top row) and cross-subpopulation (globally) adjusted p-values (bottom row), respectively. (**b**) Performance of detecting DS changes according to the number of cells per subpopulation-sample, stratified by method.

remainder of methods, simulated and estimated logFC showed high correspondence across all gene categories.

To investigate the effect of subpopulation size on DS detection, we ran methods on simulations containing 10% of DE genes using subsets of 20 to 400 cells per subpopulation-sample (**Fig. 2b**). For most methods, FDR control varies drastically with the number of cells, while TPRs improve for more cells across all methods. For aggregation-based methods, ~ 100 cells were sufficient to reach decent performance; in particular, there is a sizable gain in performance in going from 20 to 100 cells (per subpopulation per sample), but only a moderate gain in deeper sampling of subpopulations (e.g., 200 or 400 cells per subpopulation per sample). Except for edgeR on pseudobulk summed scaled CPM, unbalanced sample and group sizes had no effect on method performances (**Supplementary Figs. 6** and **7**) and increasing the number of replicates per group reveals the expected, although modest, increase in detection performance (**Supplementary Fig. 8**).

To investigate overall method concordance, we intersected the top ranked DS detections (FDR < 0.05) returned by each method across 5 simulation replicates per DS category (**Fig. 3**). We observed overall high concordance between methods, with the majority of common hits being truly differential. In contrast, most isolated intersections, i.e., hits unique to a certain method, were genes that had been simulated to be EE and thus false discoveries. Methods with vstresiduals as input yielded a noticeably high proportion of false discoveries.

Using a different anchor dataset as input to our simulation framework yielded highly consistent results (**Supplementary Figs. 1b, 2b, 3b, 5b, 6b, 9, 10** and **Supplementary File 2**). Method runtimes varied across several orders of magnitude (**Supplementary Fig. 11**). Mixed models were by far the slowest, followed by AD tests, MAST, and then scDD. Aggregation-based DS methods were the fastest. MAST, scDD, and mixed models provide arguments for parallelization, and all methods could be implemented to parallelize computations across subpopulations. For comparability, all methods were run here on a single core.

Differential state analysis of mouse cortex exposed to LPS treatment. One of the motivating examples for the DS methodological work was a scRNA-seq dataset collected to understand how peripheral lipopolysaccharide (LPS) induces its effects on brain cortex. LPS given peripherally is capable of inducing a neuroinflammatory response. Even if the mechanisms at the base of this response are still not clear, it is known that LPS can penetrate the blood-brain barrier (BBB) or alternatively, can act outside the BBB by stimulating afferent nerves, acting at circumventricular organs, and altering BBB permeabilities and functions[45,46,47,48].

We sought to investigate the effects of peripheral LPS administration on all major cell types

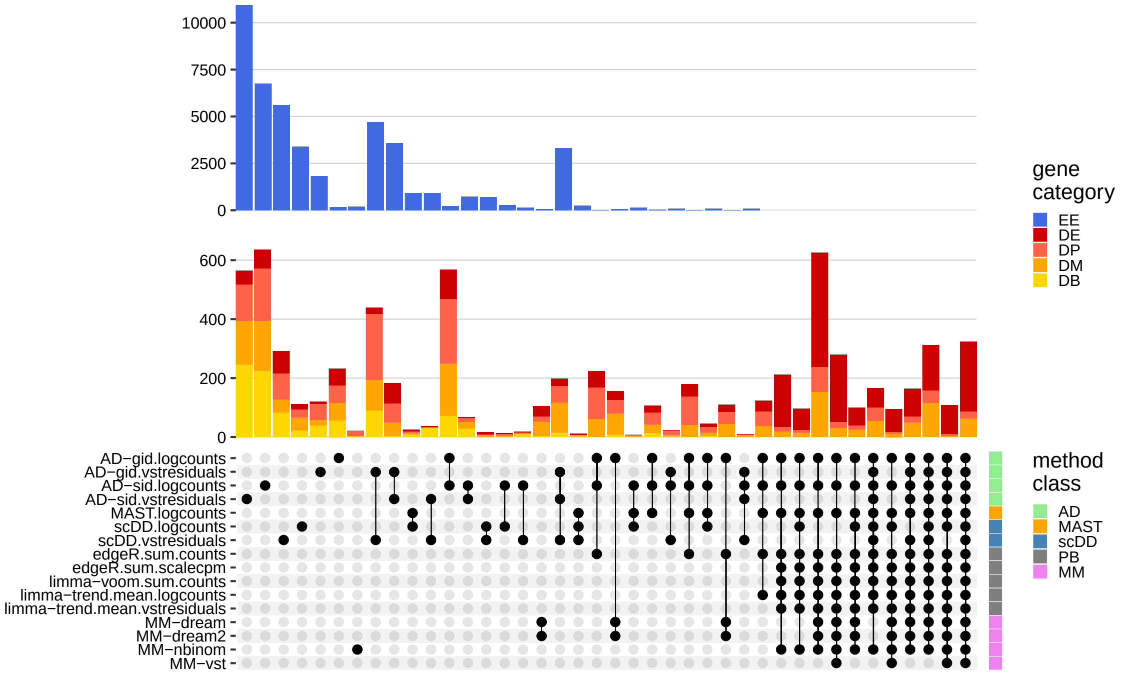

Figure 3: Between-method concordance. Upset plot obtained from intersecting the top-n ranked gene-subpopulation combinations (lowest p-value) across methods and simulation replicates. Here, $n = min(n_1, n_2)$, where n_1 = number of genes simulated to be differential, and n_2 = number of genes called differential at FDR < 0.05. Shown are the 40 most frequent interactions; coloring corresponds to (true) simulated gene categories. Bottom right annotation indicates method types (PB = pseudobulk (aggregation-based) methods, MM = mixed models, AD = Anderson-Darling tests).

in mouse frontal cortex using single-nuclei RNA-seq (snRNA-seq). The goal was to identify genes and pathways affected by LPS in neuronal and non-neuronal cells. To this end, we applied our DS analysis framework to snRNA-seq data of 4 control (vehicle) and 4 LPS-treated mice using pseudobulk (sum of counts) and edgeR. We obtained 12,317 vehicle and 12,907 treated cells that passed filtering and received a subpopulation assignment. Using graph-based clustering (Louvain algorithm[49]), we identified 22 cell clusters and annotated them into 8 subpopulations (using both canonical and computationally-identified marker genes): astrocytes, endothelial cells, microglia, oligodendrocyte progenitor cells (OPC), choroid plexus ependymal (CPE) cells, oligodendrocytes, excitatory neurons, and inhibitory neurons (see Methods and **Supplementary File 3**). Low dimensional projections of cells and pseudobulks (by subtype and condition) are shown in **Figures 4a** through **c**; sample sizes and relative subpopulation abundances are shown in **Supplementary Figure 12**.

We identified 915 genes with differential states (FDR < 0.05, |logFC| > 1) in at least one subpopulation, 751 of which were detected in only a single subpopulation (**Supplementary Fig.**

13). Since relying on thresholds alone is prone to bias, we next clustered the (per-subpopulation) fold-changes across the union of all differentially expressed genes (**Fig. 4d**). We observed a distinct set of genes (consensus clustering ID 3) that were up-regulated across all subpopulations, and enriched for genes associated with response to (external) biotic stimulus, defense and immune response (**Supplementary File 4**). Endothelial cells appeared to be most strongly affected, followed by glial cells (astrocytes, microglia and oligodendrocytes). While the responses for consensus cluster 3 were largely consistent across all subpopulations, some genes' responses departed from the trend (e.g., are specific to a single subpopulation or subset of subpopulations (**Supplementary Fig. 14**).

We next sought to estimate how homogeneous the effects observed at the pseudobulk-level are across cells. To this end, we calculated effect coefficients summarizing the extent to which each cell reflects the population-level fold-changes (**Fig. 4d**, bottom). For endothelial and glial cells, the effect coefficient distributions were well separated between vehicle and LPS samples, indicating that the majority of cells are affected. In contrast, the large overlap of the distributions in neurons suggests that only a minority of cells react. Taken together, these analyses clearly demonstrate the ability of our DS analysis framework to identify and characterize subpopulation-specific as well as global state transitions across experimental conditions.

In order to investigate the concordance of the 16 surveyed DS methods on a real dataset, we applied all methods to the LPS dataset. Intersecting genes reported as differential (at FDR < 0.05) yielded results similar to the simulation study (**Supplementary Fig. 15**); for example, AD, MAST and scDD methods report large numbers of isolated hits, whereas overall high agreement between aggregation- and mixed model-based methods is observed. While formal evaluation of method performance is not possible in the absence of ground truth, these results reveal nonetheless similar trends to the simulation results.

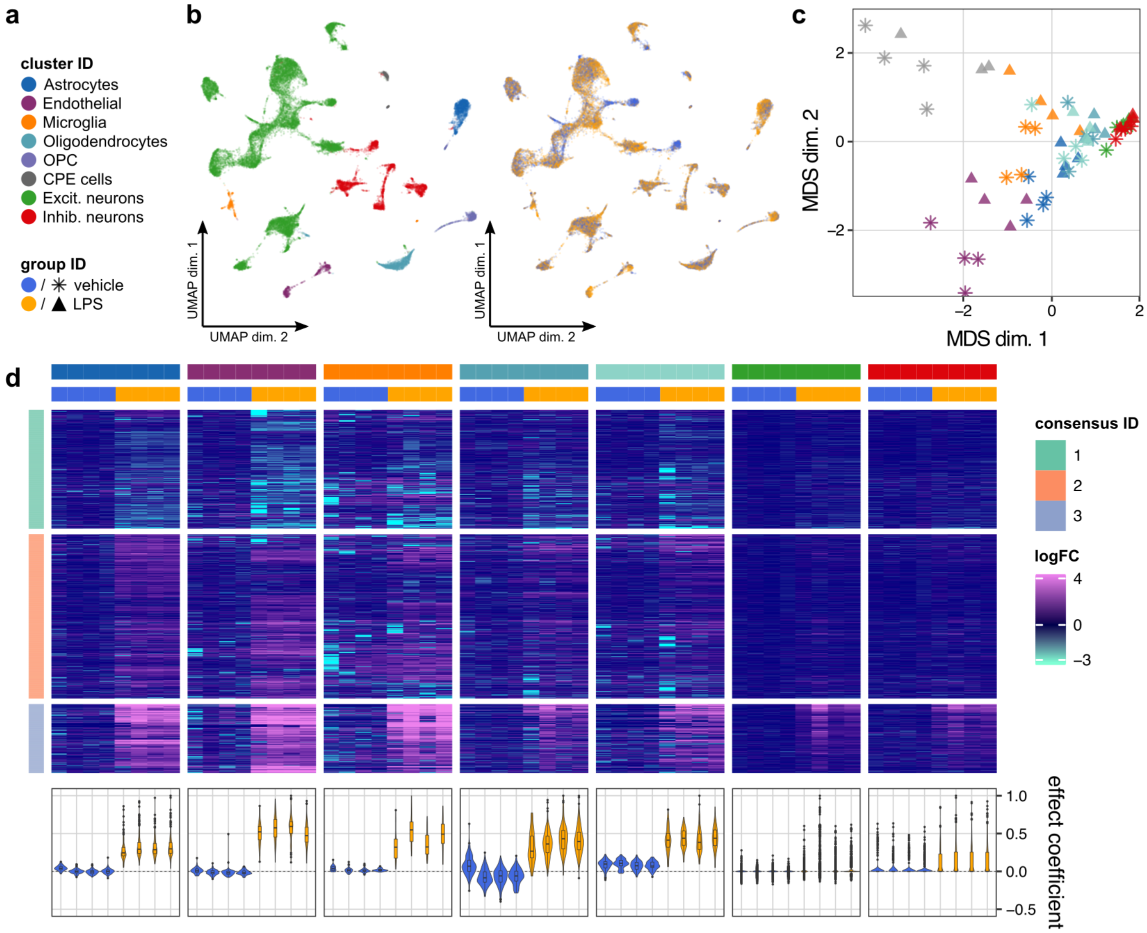

Figure 4: DS analysis of cortex tissue from vehicle- and LPS-treated mice. (**a**) Shared color and shape legend of subpopulation and group IDs. (**b**) UMAP visualization colored by subpopulation (left) and group ID (right). (**c**) Pseudobulk-level Multidimensional Scaling (MDS) plot. Each point represents one subpopulation-sample instance; points are colored by subpopulation and shaped by group ID. (**d**) Heatmap of pseudobulk-level log-expression values normalized to the mean of vehicle samples; rows correspond to genes, columns to subpopulation-sample combinations. Included is the union of DS detections (FDR < 0.05, |logFC| > 1) across all subpopulations. Data is split horizontally by subpopulation and vertically by consensus clustering ID (of genes); top and bottom 1% logFC quantiles were truncated for visualization. Bottom-row violin plots represent cell-level effect coefficients computed across all differential genes, and scaled to a maximum absolute value of 1 (each violin is a sample; coloring corresponds to group ID); effect coefficients summarize the extent to which each cell reflects the population-level fold-changes (see Methods).

Discussion

We have compared what can be considered as *in silico sorting* approaches for multi-subpopulation multi-sample multi-condition scRNA-seq datasets, where the interest is to follow each cell subpopulation along the axis of samples and conditions; we refer to these generally as differential state analyses and have largely leveraged existing tools for running such analyses. A summary of the tested DS methods across several criteria (e.g., sensitivity and runtimes) is given in **Figure 5**; methods were scored quantitatively and partially on visual inspection of the simulation results (see Methods). Furthermore, we have applied DS analysis to a new dataset to uncover subpopulation-specific changes in brain tissue from mice exposed to peripheral LPS treatment.

Aggregating data from a subpopulation to a single observation (per sample) is a natural approach to the DS problem[20,21], but it still remained to be demonstrated how effective it

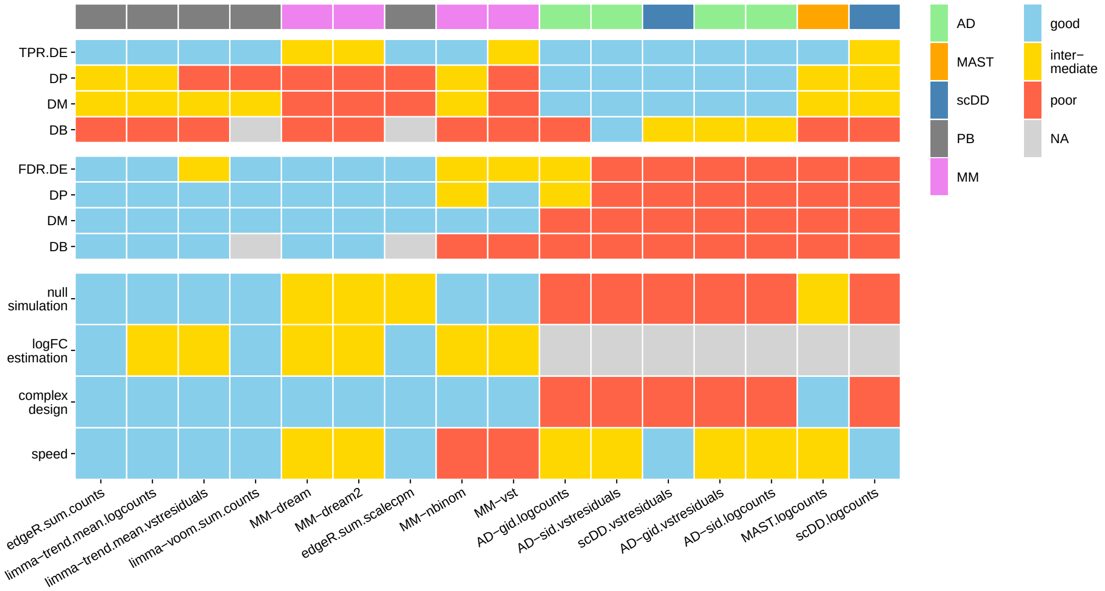

Figure 5: Summary of DS method performance across a set of evaluation criteria. Methods are ranked from left to right by their weighted average score across criteria, with the numerical encoding good = 2, intermediate = 1, and poor/NA = 0. Evaluation criteria (y-axis) comprise: DS detection sensitivity (TPR) and specificity (FDR) for each type of differential distribution, uniformity of p-value distributions under the null (null simulation), concordance between simulated and estimated logFCs (logFC estimation), ability to accommodate complex experimental designs (complex design), and runtimes (speed). Top annotation indicates method types (PB = pseudobulk (aggregation-based) methods, MM = mixed models, AD = Anderson-Darling tests). Null simulation, logFC estimation, complex design and runtimes received equal weights of 0.5; TPR and FDR were weighted according to the frequencies of modalities in scRNA-seq data reported by Korthauer *et al.*[33]: $\sim 75\%$ unimodal, $\sim 5\%$ trimodal and $\sim 25\%$ bimodal, giving weights of 0.75 for DE, 0.125 for DP and DM, and 0.05 for DB.

is. Based on our simulation results, the tested aggregation-based DS methods were extremely fast and showed overall a stable high performance, although depending on the scale of the data analyzed, logFCs were attenuated for some combinations. While mixed model methods performed similarly well, their computational cost may not be worth the flexibility they provide (**Fig. 5** and **Supplementary Fig. 11**). Methods developed specifically for scRNA-seq differential analysis were outperformed by aggregation and mixed models, but it should be mentioned that these methods focus on comparing sets of cells and were not specifically designed for the multi-group multi-sample problem. Furthermore, methods that compared full distributions did not perform well overall (**Fig. 5**). This latter class of methods was used here as a reference point, but could also be improved to be more targeted to the DS inference problem. For example, Anderson-Darling tests were run in two ways, group-wise or sample-wise, where under the null hypothesis, all distributions are equal. In the sample-wise case, departures from the null could happen between replicates of the same experimental condition and in the group-wise case, it is perhaps not ideal to mix distributions from different samples. Thus, while our results suggest that aggregation methods are fast and perform amongst the best, there may still be value in considering full distributions, if bespoke methods were developed. Furthermore, methods that integrate both changes in the mean and changes in variability may be worth exploring.

The starting point of a DS analysis is a count table across genes and cells, where each cell has an appropriate subpopulation and sample label, and metadata (e.g., patient, experimental condition information) accompanies the list of samples. This starting point, organization of cells into subpopulations ("types"), is itself an active and debated area of research[2,3] and one that already applies a computational analysis on a given dataset, whether that be clustering or manual or computational assignment; in fact, combining computational and manual assignment was recently listed as best practice[44].

Although not discussed here, researchers would generally first apply a differential abundance (DA) analysis of subpopulations, which naturally leads to discussions about the ambiguity of cell type definitions. DA analysis will highlight subpopulations whose relative abundance changes according to the treatment; in contrast, DS analysis will identify changes within the defined subpopulations that are associated with the treatment. Thus, the combination of DA and DS should always be capable of detecting interesting differences, with results dependent on how cell types are defined.

Another aspect of subpopulation-level analyses is that there are clear connections to existing tools and practices in the analysis of gene expression. For example, one can visualize data at the aggregate level (e.g., MDS plot for each subpopulation; **Fig. 4c**) and apply standard tools (e.g., geneset analysis, gene network analysis) for discovery and interpretation on each subpopulation, thus leveraging existing methods.

By default, we have focused on subpopulation-specific DS analysis; in particular, the methods fit a separate model (i.e., separate dispersion) for each subpopulation, which explicitly allows them to have different levels of variability. However, some of the models could be reshaped, e.g., to fit a single model over all subpopulations and test parameters within this model. This strategy may allow better separation of features that respond globally versus specific to a given subpopulation, which may be important to separate in the downstream interpretation analyses.

The number of cells required to detect DS changes depends on many factors, including the effect's strength, the number of replicates, the number of cells per sample in each subpopulation, and the sensitivity of the scRNA-seq assay, which itself is a moving target. In general, there is a clear gain in power for larger subpopulations, while FDR control can vary greatly with the number of cells (**Fig. 2**). Going forward, it would be of interest to further explore the origins of this instability, in order to better maximize sensitivity while still controlling for errors.

Another aspect to consider in this context is the resolution of subpopulations subjected to differential testing; for example, there is an analogous tradeoff between sensitivity (e.g., larger subpopulations) and specificity (e.g., effects that target particular subpopulations). Here, methods that integrate the relationship between subpopulations (e.g., treeclimbR[50]) could be applied as an additional layer to improve signal detection.

In the process of this study, we created a flexible simulation framework to facilitate method comparisons as well as data handling tools and pipelines for such experiments, implemented in the muscat R package. By using sample-specific estimates, inter-sample variability present in the reference dataset will be represented in the simulated data. Even though we tested here a broad set of scenarios, there may be other scenarios of interest (e.g., different percentages of the DM mixtures); the simulation framework provided in the muscat package could readily be used to expand the set of simulation scenarios. Furthermore, the simulation framework could be extended to induce batch effects via, for example, incorporating sample-specific logFCs in the computation of simulation means. For this, more research needs to be done to understand how and at what magnitude batch effects manifest. Furthermore, our simulation framework could be extended: i) to accommodate an arbitrary number of groups for which the magnitude of differential signal, the percentage of differential genes, as well as the set of affected subpopulations could be varied; or, ii) implementing *type* genes such that they are not specific to a single subpopulation, perhaps even in a hierarchical structure to represent markers of both broad and specific cell types. Taken together, we expect our simulation framework to be useful to investigate various scRNA-seq data analyses, such as batch correction frameworks, clustering, reference-based cell-type inference methods, marker gene selection methods as well as further developments in DS analysis.

Although we set out with the goal of discovering subpopulation-specific responses across experimental conditions, one needs to be careful in how strongly these claims are made. Absence of evidence is not evidence of absence. In particular, there is a potentially strong bias in statistical power to detect changes in larger cell populations, with decreased power for rarer populations. Statistical power to detect changes in cell states also relates to the depth of sequencing per cell; for example, it has been speculated that cell states are a *secondary* regulatory module[3] and it is unclear at this stage whether we are sequencing deeply enough to access all of the interesting transcriptional programs that relate to cell state. However, despite the potential loss of single-cell resolution, aggregation approaches should be helpful in this regard, accessing more genes at the subpopulation level.

Online Methods

Preprocessing of simulation reference data. As simulation anchors, we used scRNA-seq datasets obtained from i) PBMCs by Kang *et al.*[20] (8 control vs. 8 IFN-β treated samples); and, ii) mouse brain cortex cells (4 vehicle vs. 4 LPS-treated samples; see below). In order to introduce known changes in expression, we only used samples from the reference (control and vehicle, respectively) condition as input to our simulation framework. These were minimally filtered to remove cells with less than 200 detected genes, and genes detected in less than 100 cells. Available metadata was used to filter for singlet cells as well as cells that have been assigned to a cell population. Finally, for more accurate parameter estimation, only subpopulation-sample instances with at least 100 cells were retained, leaving 4 samples per reference dataset, 4 subpopulations for the Kang *et al.*, and 3 subpopulations for the LPS dataset.

Simulation framework. The simulation framework (**Fig. 1a**) comprises: i) estimation of NB parameters from a reference multi-subpopulation, multi-sample dataset; ii) sampling of gene and cell parameters to use for simulation; and, iii) simulation of gene expression data as negative binomial (NB) distributions or mixtures thereof.

Let $Y = (y_{gc}) \in \mathbb{N}_0^{G\times C}$ denote the count matrix of a multi-sample multi-subpopulation reference dataset with genes $\mathcal{G} = \{g_1, \ldots, g_G\}$ and sets of cells $\mathcal{C}_{sk} = \{c_1^{sk}, ..., c_{C_{sk}}^{sk}\}$ for each sample s and subpopulation k (C_{sk} is the number of cells for sample s, subpopulation k). For each gene g, we fit a model to estimate sample-specific means β_g^s, for each sample s, and dispersion parameters ϕ_g using edgeR's $\mathrm{estimateDisp}$ function with default parameters. Thus, we model the reference count data as NB distributed:

$$Y_{gc} \sim \text{NB}(\mu_{gc}, \phi_g)$$

for gene g and cell c, where the mean $\mu_{gc} = \exp(\beta_g^{s(c)}) \cdot \lambda_c$. Here, $\beta_g^{s(c)}$ is the relative abundance of gene g in sample $s(c)$, λ_c is the library size (total number of counts), and ϕ_g is the dispersion.

In order to introduce a multi-subpopulation, multi-sample data structure, we sample a set of K clusters as reference, as well as S reference samples for each of two groups, resulting in an unpaired design. Alternatively, pairing of samples can be mimicked by fixing the same set of reference samples for both groups. For each subpopulation $k \in \{1,...,K\}$, we sample a set of genes $\mathcal{G}_k^* \subset \mathcal{G}$ used for simulation, such that most genes are common to all subpopulations ($\mathcal{G}_1^* \cap \mathcal{G}_2^* \cap ... \cap \mathcal{G}_K^* \approx (1-p) \cdot G$), while a small set ($p \cdot 100$ percent) of *type*-specific genes are sampled separately for each subpopulation ($\mathcal{G}_{k'} \cap \mathcal{G}_k = \emptyset \; \forall \; k \neq k'$), giving rise to distinct subpopulations. Secondly, for each sample s and subpopulation k, we draw a set of cells $\mathcal{C}^*_{sk} \subset \mathcal{C}_{sk}$ (and their corresponding λ_c, $\beta_g^{s(c)}$ and ϕ_g) to simulate (negative binomial random variables) from, where cells $\mathcal{C}_{sk}$ belong to the corresponding reference cluster-sample drawn previously.

Lastly, differential expression of a variety of types is added for a subset of genes. For each subpopulation, we randomly assign each gene to a given *differential distribution* category according to a probability vector $(p_{EE}, p_{EP}, p_{DE}, p_{DP}, p_{DM}, p_{DB})$; see **Figure 1b**. For each gene and subpopulation, we draw a vector of fold changes from a Gamma distribution with shape 4 and rate $4/\mu_{\text{logFC}}$, where μ_{logFC} is the desired average logFC across all genes and subpopulations. The direction of differential expression is randomized for each gene, with equal probability of up- and down-regulation. We split the cells in a given subpopulation-sample combination into two sets (representing treatment groups), $\mathcal{T}_A$ and $\mathcal{T}_B$, which are in turn split again into two sets each (representing subpopulations within the given treatment group), $\mathcal{T}_{A_1}/\mathcal{T}_{A_2}$ and $\mathcal{T}_{B_1}/\mathcal{T}_{B_2}$.
For EE genes, counts for $\mathcal{T}_A$ and $\mathcal{T}_B$ are drawn using identical means. For EP genes, we multiply the effective means for identical fractions of cells per group by the sampled FCs, i.e., cells are split such that $\dim \mathcal{T}_{A_1} = \dim \mathcal{T}_{B_1}$ and $\dim \mathcal{T}_{A_2} = \dim \mathcal{T}_{B_2}$. For DE genes, the means of one group, A or B, are multiplied with the sampled FCs. DP genes are simulated analogously to EP genes with $\dim \mathcal{T}_{A_1} = a \cdot \dim \mathcal{T}_A$ and $\dim \mathcal{T}_{B_1} = b \cdot \dim \mathcal{T}_B$, where $a + b = 1$ and $a \neq b$ (default $a = 0.3, b = 0.7$). For DM genes, 50% of cells from one group are simulated at $\mu \cdot \text{FC}$. For DB genes, all cells from one group are simulated at $\mu \cdot {}^{\text{FC}}\!/_2$, and the second group is split into equal proportions of cells simulated at μ and $\mu \cdot \text{FC}$, respectively.
Details on all simulation parameters, illustrative examples of their effects, and instructions on how to generate an interactive quality control report and benchmark DS methods through simulated data are provided in the muscat R/Bioconductor package's documentation (see Software specification and code availability).

Aggregation-based methods. We summarize the input measurement values for a given gene over all cells in each subpopulation and by sample. The resulting pseudobulk data matrix has dimensions $G \times S$, where S denotes the number of samples, with one matrix obtained per subpopulation. Depending on the specific method, which includes both a type of data to operate on (e.g., counts, logcounts) and summary function (e.g., mean, sum), the varying number of cells between samples and subpopulations is accounted for prior to or following aggregation. For logcounts methods, we apply a library size normalization to the input raw counts. vstresiduals are computed using R package sctransform's vst function[38]. For scalecpm, we calculate the total library size of each subpopulation k and sample s as

$$\Lambda_{sk} = \sum_{g=1}^{G} \sum_{c=1}^{C_{sk}} y_{gc}$$

where G represents the number of genes, C_{sk} is the total number of cells in sample s that have been assigned to subpopulation k, and y_{gc} denotes the counts observed for gene g in cell c. We then multiply the CPM of a given sample and subpopulation with the respective total library size in millions to scale the CPM values back to the count scale:

$$\text{CPM}^*_{sk} = \text{CPM}_{sk} \cdot \Lambda_{sk} \cdot 1\text{e}^{-6}$$

edgeR-based methods were run using glmQLFit and glmQLFTtest[51]; methods based on limma-voom and limma-trend were run using default parameters.

Mixed models. Mixed model methods were implemented using three main approaches: i) fitting linear mixed models (LMMs) on log-normalized data with observational weights, ii) fitting LMMs on variance-stabilized data, iii) fitting generalized linear mixed models (GLMMs) directly on counts. Subpopulations with less than 10 cells in any sample and genes detected in fewer than 20 cells were excluded from differential testing. In each case, a $\sim 1 + \text{group\_id} + (1|\text{sample\_id})$ model was fit for each gene, optimizing the restricted maximum likelihood (i.e. $\text{REML} = \text{TRUE}$), and p-values were calculated using Satterthwaite estimates of degrees of freedom (the Kenward-Roger approach being longer to compute and having a negligible impact on the final results). Fitting, testing and moderation were applied subpopulation-wise.
For the first approach (MM-dream), we relied on the variancePartition[52] package's implementation for repeated measurement bulk RNA-seq, using voom's[25] precision weights as originally described but without empirical Bayes moderation and the duplicateCorrelation[53] step, as this was computationally intensive and had a negligible impact on the significance (as also observed previously for batch effects[21]). Method MM-dream2 uses an updated alternative to this approach using variancePartition's new weighting scheme[Hoffman2020-variancePartition] instead of voom.
For the second approach (MM-vst), we first applied the variance-stabilizing transformation globally before splitting cells into subpopulations, and then fitted the model using the lme4 package[54] directly on transformed data (and without observational weights). We then applied eBayes moderation as in the first approach. We tested both the variance-stabilizing transformation from the DESeq2 package[23], and that from the sctransform package[38], the latter of which was specifically designed for Unique Molecular Identifier (UMI) based scRNA-seq; since the latter outperformed the former (data not shown), it was retained for the main results shown here.

For the GLMM-based approach (MM-nbinom), we supplemented the model with an offset equal to the library size factors, and fitted it directly on counts using both Poisson and negative binomial distributions (with log-link). The Poisson-distributed model was fit using the bglmer function of the blme package, while the negative binomial model was fit with the glmmTMB framework (family = nbinom1). As eBayes moderation did not improve performance on these results, it was not applied in the final implementation.

All these methods and variations thereof are available through the mmDS function of the muscat package.

Other methods. For Anderson-Darling tests, we used the ad.test function from the kSamples R package[55], which applies a permutation test that uses the Anderson-Darling criterion[32] to test the hypothesis that a set of independent samples arose from a common, unspecified distribution. Method AD-sid uses sample labels as grouping variables, thus testing whether any sample from any group arose from a different distribution than the remaining samples. For method AD-gid, we used group labels as grouping variable, thus testing against the null hypothesis that both groups share a common underlying distribution; with disregard of sample labels. For both methods, we require genes to be expressed in at least 10 cells in a given cluster to be tested for differential states.

scDD[33] was run using default prior parameters and min.nonzero = 20, thus requiring a gene to be detected in at least 20 cells per group to be considered for differential testing in a given subpopulation. For MAST[34], we fit a subpopulation-level zero-inflated regression model for each gene (function zlm) and applied a likelihood-ratio test (function lrTest) to test for between-group differences in each subpopulation. Both steps were run using default parameters. AD methods and scDD were run on both logcounts and vstresiduals; MAST was run on logcounts only.

Animal studies - LPS dataset. Ethical approval for this study was provided by the Federal Food Safety and Veterinary Office of Switzerland. All animal experiments were conducted in strict adherence to the Swiss federal ordinance on animal protection and welfare as well as according to the rules of the Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC).

CD1 male mice (Charles River Laboratories, Germany) age 11 weeks were divided into two groups with 4 animals each: a vehicle and a lipopolysaccharide (LPS) treatment group. The LPS-treated group was given a single intraperitoneal injection of LPS from Escherichia coli O111:B4 (Sigma Aldrich, L2630) at a dose of 5mg/kg, dissolved in 0.9% NaCl. Vehicle mice were injected with a solution of DMSO/Tween80/NaCl (10%/10%/80%). The mice were sacrificed 6 hours later by anesthetizing the animals with isoflurane followed by decapitation. Brains were quickly frozen and stored at -80°C.

Nuclei isolation, mRNA-seq library preparation and sequencing - LPS dataset. Nuclei were prepared using the NUC201 isolation kit from Sigma Aldrich. Briefly, 8×50µm sagittal sections of cortex from each animal were prepared using a microtome and placed in 200µl of cold Nuclei Pure Lysis Buffer (Nuclei Pure Prep Nuclei isolation kit - Sigma Aldrich) with 1M dithiothreitol (DTT) and 0.2U/µl SUPERase inhibitor (Invitrogen) freshly added before use. Nuclei were extracted using a glass dounce homogenizer with Teflon pestle using 10-12 up and down strokes in lysis buffer. 360µl of cold 1.8M Sucrose Cushion solution was added to lysate which was then filtered through a 30µm strainer. 560µl of filtered solution was carefully overlayed on 200µl of Sucrose solution and nuclei were purified by centrifugation for 45min at 16,000g. The nuclei pellet was re-suspended in 50µl cold Nuclei Pure Storage Buffer (Nuclei Pure Prep Nuclei isolation kit Sigma Aldrich) with 0.2U/µl SUPERase inhibitor and centrifuged for 5min at 500g. The supernatant was removed, the pellet washed again with Nuclei Pure Storage Buffer with 0.2U/µl SUPERase inhibitor, and centrifuged for 5min at 500g. Finally, the pellet was re-suspended in 50µl cold Nuclei Pure Storage Buffer with 0.2U/µl SUPERase inhibitor. Nuclei were counted using trypan blue staining on Countess II (Life technology). A total of 12,000 estimated nuclei from each sample was loaded on the 10x Single Cell B Chip.

cDNA libraries from each sample were prepared using the Chromium Single Cell 3' Library and Gel Bead kit v3 (10x Genomics) according to the manufacturers instructions. cDNA libraries were sequenced using Illumina Hiseq 4000 using the HiSeq 3000/4000 SBS kit (Illumina) and HiSeq 3000/4000 PE cluster kit to get a sequencing depth of 30K reads/nuclei.

Single nucleus RNA-seq data processing and quality control. Paired end sequencing reads from the eight samples were preprocessed using 10X Genomics Cell Ranger 3.0 software for sample demultiplexing, barcode processing and single-nucleus 3' gene counting (single nuclei mode; counting performed on unspliced Ensembl transcripts, as described in the 10x Genomics documentation). Mouse reference genome assembly GRCm38/mm10 was used for alignment of sequencing reads. The gene by cell count matrices generated by Cell Ranger pipeline were used for downstream quality control and analyses.

LPS dataset analysis. Filtering for doublet cells was performed on each sample separately using the hybrid method of the scds package[56], removing the expected 1% per thousand cells captured with the highest doublet score. Quality control and filtering were performed using the scater[57]

R package. Upon removal of genes that were undetected across all cells, we removed cells whose feature counts, number of expressed features, and percentage of mitochondrial genes fell beyond 2.5 Median Absolute Deviations (MADs) of the median. Finally, features with a count > 1 in at least 20 cells were retained for downstream analysis.

Next, we used Seurat[43,9] v3.0 for integration, clustering, and dimension reduction. Integration and clustering were performed using the 2000 most highly variable genes (HVGs) identified via Seurat's FindVariableFeatures function with default parameters; integration was run using the first 30 dimensions of the Canonical Correlation Analysis (CCA) cell embeddings. Clusterings as well as dimension reductions (t-SNE[58] and UMAP[59]) were computed using the first 20 principal components. For clustering, we considered a range of resolution parameters (0.1-2); downstream analyses were performed on cluster assignments obtained from resolution 0.2 (22 subpopulations). Cluster merging and cell-type annotation were performed manually on the basis of a set of known marker genes in conjunction with marker genes identified programmatically with scran's findMarkers function[60], and additional exploration with iSEE[61]. We identified 8 subpopulations that included all major cell types, namely, astrocytes, endothelial cells, microglia, oligodendrocyte progenitor cells (OPC), choroid plexus ependymal (CPE) cells, oligodendrocytes, excitatory neurons, and inhibitory neurons.

DS analysis was run using edgeR[22] on pseudobulk (sum of counts), requiring at least 10 cells in at least 2 samples per group for a subpopulation to be considered for differential testing; the CPE cells subpopulation did not pass this filtering criterion and were excluded from differential analysis. Genes with FDR < 0.05 and |logFC| > 1 were retained from the output. To distinguish subpopulation-specific and shared signatures, we assembled a matrix of logFCs (calculated for each cell subpopulation) of the union of all differential genes (FDR < 0.05 and |logFC| > 1), and performed consensus clustering of the genes using the M3C package[62] (penalty term method), choosing the number of clusters with the highest stability.

To estimate per-cell effect coefficients, we calculated dot products of each cell's normalized log-expression and the group-level logFCs using only the DS genes detected for the corresponding subpopulation.

Performance summary criteria. For each of the metrics in **Figure 5**, method performances are considered to be 'good', 'intermediate', 'poor' or 'NA' (not available). Method assessments were made as follows:

- *TPR/FDR:* For each type of DD category, we consider TPRs and FDR at FDR 5% averaged across two references, 5 simulation replicates and 3 clusters (**Fig. 2a**). Methods are scored according to $\overline{TPR} > 2/3$: good, $> 1/3$: intermediate, otherwise: poor; and $\overline{FDR} < 0.05$: good, < 0.1: intermediate, otherwise: poor.

- *null simulation:* We perform a Kolmogorov-Smirnov (KS) test on the uniformity of p-values ($\mathrm{ks.test}$ with CDF $y = "\mathrm{punif}"$) under the null (**Supplementary Fig. 2**) for each of two references, three simulation replicates and three clusters per simulation, resulting in a total of 18 tests per method. KS statistics (largest difference between observed and uniform empirical cumulative distribution functions) are then averaged, and categorized according to $\overline{KS}_{stat.} < 0.1$: good, < 0.25: intermediate, otherwise: poor.

- *logFC estimation:* From visual inspection, methods that gave logFC estimates near the diagonal (against the true simulated logFC) were labeled as good; methods with attenuated logFC estimates were listed as intermediate; methods that did not return logFC estimates were given ‘NA‘.

- The *complex design* criterion is qualitative. Methods are scored 'good' or 'poor' depending on whether or not they are capable of accommodating the experimental design of interest, i.e., multiple replicates across two conditions.

- *speed* summarizes the runtimes recorded for increasing numbers of cells and genes, respectively (**Supplementary Fig. 11**). Scores are given according to the three major groups observed (in terms of runtimes) with scDD and pseudobulk methods running in the order of seconds; AD, MAST and MM-dream methods two orders of magnitude longer; and MM-nbinom and -vst three to four orders of magnitude longer.

Methods were ranked according to the weighted average score across all metrics, with numerical encoding good = 3, intermediate = 2, poor/NA = 0; a weight of 0.5 for error control, logFC estimation, complex design and speed; and weights of 0.75, 0.125, 0.125 and 0.05 for TPR/FDR on DE, DP, DM and DB genes, respectively. This weighting of the different DD categories is in accordance with the frequencies of multi-modalities in scRNA-seq data reported by Korthauer *et al.* (∼75% unimodal, ∼5% trimodal and ∼25% bimodal, which were split equally between DP and DM).

Software specifications and code availability. All analyses were run in R v3.6.2[63], with Bioconductor v3.10[64]. Performance measures were calculated using iCOBRA[65], and results were visualized with ggplot2[66], ComplexHeatmap[67], and UpSetR[68]. All package versions used throughout this study are captured in **Supplementary File 5**. Data preprocessing, simulation and analysis code are accessible at `https://github.com/HelenaLC/muscat-comparison`, including a browseable workflowr[69] website for the LPS dataset analysis (**Supplementary File 3**). All aggregation and DS analysis methods are provided in the muscat R package, which is available at `https://www.bioconductor.org/packages/muscat` through the open-source Bioconductor project.

Data availability. The original droplet scRNA-seq data from Kang *et al.*[20] is deposited under the Gene Expression Omnibus accession GSE96583. The raw LPS dataset is available from ArrayExpress (accession: E-MTAB-8192) and the Cell Ranger-processed files and metadata are available from DOI:10.6084/m9.figshare.8976473. Both datasets are also available in R through the muscData Bioconductor ExperimentHub package. **Supplementary File 6** is a commpressed archive containing R objects of all simulations and results. **Supplementary Files 1-6** are available from DOI:10.6084/m9.figshare.8986193

Acknowledgments

The authors thank members of the Robinson Lab at the University of Zurich for valuable feedback on methodology, benchmarking and exposition. This work was supported by the Swiss National Science Foundation (grant numbers 310030_175841, CRSII5_177208) and the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (grant number 2018-182828). MDR acknowledges support from the University Research Priority Program Evolution in Action at the University of Zurich.

Author contributions

HLC, CS and MDR developed aggregation-based methods; PLG developed MM-based methods. HLC implemented methods, the simulation framework, and the method comparison; CS assisted in several technical and conceptual aspects. DC, LC, CR and DM designed mouse LPS experiments; LC and CR provided mouse cortex tissue sections for snRNA-seq. PLG and HLC performed data processing, analysis, and interpretation; MDR and DM assisted in designing analyses and DM contributed to interpretation. HLC, MDR, and PLG drafted the manuscript, with contributions from all authors. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

1. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16,** 133–145 (2015).
2. Morris, S. A. The evolving concept of cell identity in the single cell era. *Development* **146** (2019).
3. Xia, B. & Yanai, I. A periodic table of cell types. *Development* **146** (2019).
4. Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8,** e43803 (2019).
5. Tiklová, K. *et al.* Single-cell RNA sequencing reveals midbrain dopamine neuron diversity emerging during mouse brain development. *Nature Communications* **10,** 581 (2019).
6. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods* **15,** 255–261 (2018).
7. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology* **34,** 1145–1160 (2016).
8. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Research* **25,** 1491–1498 (2015).
9. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21 (2019).
10. Diaz-Mejia, J. J. *et al.* Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. *F1000Research* **8,** 296 (2019).
11. Zhang, A. W. *et al.* Probabilistic cell type assignment of single-cell transcriptomic data reveals spatiotemporal microenvironment dynamics in human cancers. *bioRxiv* **521914** (2019).
12. Nowicka, M. *et al.* CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research* **6,** 748 (2019).
13. Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J. & Nolan, G. P. Automated identification of stratifying signatures in cellular subpopulations. *PNAS* **111,** E2770–7 (2014).
14. Arvaniti, E. & Claassen, M. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature Communications* **8,** 14825 (2017).
15. Greene, E. *et al.* A new data-driven cell population discovery and annotation method for single-cell data, FAUST, reveals correlates of clinical response to cancer immunotherapy. *bioRxiv* **702118** (2019).
16. Chevrier, S. *et al.* Compensation of Signal Spillover in Suspension and Imaging Mass Cytometry. *Cell Systems* **6,** 612–620.e5 (2018).
17. Weber, L. M., Nowicka, M., Soneson, C. & Robinson, M. D. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications Biology* **2,** 183 (2019).
18. Fonseka, C. Y. *et al.* Mixed-effects association of single cells identifies an expanded effector CD4 T cell subset in rheumatoid arthritis. *Science Translational Medicine* **10** (2018).
19. Krieg, C. *et al.* Author Correction: High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nature Medicine* **24,** 1773–1775 (2018).
20. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology* **36,** 89–94 (2018).
21. 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).
22. 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).
23. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550 (2014).
24. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43,** e47 (2015).
25. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15,** R29 (2014).

26. Tung, P.-Y. *et al.* Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports* **7,** 39921 (2017).

27. Ma, B. X., Korthauer, K., Kendziorski, C. & Newton, M. A. A Compositional Model to Assess Expression Changes from Single-Cell Rna-Seq Data. *bioRxiv* **655795** (2019).

28. Seiler, C. *et al.* Uncertainty Quantification in Multivariate Mixed Models for Mass Cytometry Data. *arXiv* **1903.07976** (2019).

29. Chen, S. *et al.* Dissecting heterogeneous cell-populations across signaling and disease conditions with PopAlign. *bioRxiv* **421354** (2018).

30. Jaakkola, M. K., Seyednasrollah, F., Mehmood, A. & Elo, L. L. Comparison of methods to detect differentially expressed genes between single-cell populations. *Briefings in Bioinformatics* **18,** 735–743 (2017).

31. Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* **20,** 40 (2019).

32. Scholz, F. W. & Stephens, M. A. K-Sample Anderson-Darling Tests. *Journal of the American Statistical Association* **82,** 918–924 (1987).

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

34. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16,** 278 (2015).

35. Köster, J. & Rahmann, S. Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).

36. Svensson, V. Droplet scRNA-seq is not zero-inflated. *bioRxiv* **582064** (2019).

37. Soneson, C. & Robinson, M. D. Towards unified quality verification of synthetic count data with countsimQC. *Bioinformatics* **34,** 691–692 (2018).

38. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *bioRxiv* **576827** (2019).

39. William Townes, F., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature Selection and Dimension Reduction for Single Cell RNA-Seq based on a Multinomial Model. *bioRxiv* **574574** (2019).

40. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* **7,** 1141 (2018).

41. Freytag, S., Tian, L., Lönnstedt, I., Ng, M. & Bahlo, M. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research* **7,** 1297 (2018).

42. Waltman, L. & van Eck, N. J. A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B* **86,** 471 (2013).

43. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36,** 411–420 (2018).

44. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* **15,** e8746 (2019).

45. Romeo, H. E., Tio, D. L., Rahman, S. U., Chiappelli, F. & Taylor, A. N. The glossopharyngeal nerve as a novel pathway in immune-to-brain communication: relevance to neuroimmune surveillance of the oral cavity. *Journal of Neuroimmunology* **115,** 91–100 (2001).

46. Ulmer, A. J., Th. Rietschel, E., Zähringer, U. & Heine, H. Lipopolysaccharide: Structure, Bioactivity, Receptors, and Signal Transduction. *Trends in Glycoscience and Glycotechnology* **14,** 53–68 (2002).

47. Xaio, H., Banks, W. A., Niehoff, M. L. & Morley, J. E. Effect of LPS on the permeability of the blood–brain barrier to insulin. *Brain Research* **896,** 36–42 (2001).

48. Banks, W. A. & Robinson, S. M. Minimal penetration of lipopolysaccharide across the murine blood–brain barrier. *Brain, Behavior, and Immunity* **24,** 102–109 (2010).

49. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008 (2008).

50. Huang, R. *et al.* treeclimbR pinpoints the data-dependent resolution of hierarchical hypotheses. *bioRxiv* **2020.06.08.140608** (2020).

51. 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. *Methods in Molecular Biology* **1418,** 391–416 (2016).

52. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17,** 483 (2016).

53. Smyth, G. K., Michaud, J. & Scott, H. S. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21,** 2067–2075 (2005).

54. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67,** 1–48 (2015).

55. Scholz, F. & Zhu, A. kSamples: K-Sample Rank Tests and their Combinations. *R package* (2019).

56. Bais, A. S. & Kostka, D. scds: Computational Annotation of Doublets in Single Cell RNA Sequencing Data. *bioRxiv* **564021** (2019).

57. 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* **33,** 1179–1186 (2017).

58. Maaten, L. v. d. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).

59. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **1802.03426** (2018).

60. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5,** 2122 (2016).

61. Rue-Albrecht, K., Marini, F., Soneson, C. & Lun, A. T. L. iSEE: Interactive SummarizedExperiment Explorer. *F1000Research* **7,** 741 (2018).

62. John, C. & Watson, D. M3C: Monte Carlo Reference-based Consensus Clustering. *R package* (2019).

63. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2019).

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

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

66. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).

67. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32,** 2847–2849 (2016).

68. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33,** 2938–2940 (2017).

69. Blischak, J. D., Carbonetto, P. & Stephens, M. Creating and sharing reproducible research code the workflowr way. *F1000Research* **8,** 1749 (2019).