

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

Helena L. Crowell^{1,2}, Charlotte Sonesson^{1,2,3,*}, Pierre-Luc Germain^{1,4,*}, Daniela Calini⁵, Ludovic Collin⁵, Catarina Raposo⁵, Dheeraj Malhotra⁵ & 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

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

Introduction

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

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

35 We intentionally use the term *subpopulation* to be more generic than cell *type*^[7,8], which
36 itself is meant to represent a discrete and stable molecular signature; however, the precise defini-
37 tion of cell type is widely debated^[2,3]. In our framework, a subpopulation is simply a set of cells
38 deemed to be similar enough to be considered as a group and where it is of interest to interrogate
39 such sets of similarly-defined cells across multiple samples and conditions. Therefore, cells from
40 a scRNA-seq experiment are first organized into subpopulations, e.g., by integrating the multiple
41 samples together^[9] and clustering or applying a subpopulation-level assignment algorithm^[10] or
42 cell-level prediction^[11]; clustering and manual annotation is also an option. Regardless of the mode
43 or the uncertainty in subpopulation assignment, the discovery framework we describe provides a
44 basis for biological interpretation and a path to discovering interesting expression patterns *within*
45 subpopulations across samples. Even different subpopulation assignments of the same data could
46 be readily interpretable. For example, T cells could be defined as a single (albeit diverse) cell
47 subpopulation or could be divided into discrete subpopulations, if sufficient information to cate-

48 gorize the cells at this level of resolution is available. In either case, the framework presented here
49 would focus on the subpopulation of interest and look for expression changes *across* conditions.
50 This naturally introduces an interplay with the definition of cell types and states themselves (e.g.,
51 discrete states could be considered as types) and thus with the methods used to computationally
52 or manually classify cells. Overall, our goal here is to explore the space of scRNA-seq datasets with
53 several subpopulations and samples, in order to understand the fidelity of methods to discover cell
54 state changes.

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

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

78 In existing comparison studies of scRNA-seq differential detection methods^[30,6,31], analyses
79 were limited to comparing groups of cells and had not explicitly considered sample-level inferences
80 or aggregation approaches. The rapid uptake of new single-cell technologies has driven the col-
81 lection of scRNA-seq datasets across multiple samples. Thus, it remains to be tested whether

82 existing methods designed for comparing expression in scRNA-seq data are adequate for such
83 cross-sample comparisons, and in particular, how sensitive aggregation methods are to detect
84 subpopulation-level responses.

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

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

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

Results

107 **Simulation framework.** To explore the various aspects of DS analysis, we developed a straight-
108 forward but effective simulation framework that is anchored to a labeled multi-sample multi-
109 subpopulation scRNA-seq reference dataset, and exposes parameters to modulate: the number of
110 subpopulations and samples simulated, the number of cells per subpopulation (and sample), and
111 the type and magnitude of a wide range of patterns of differential expression. Using (non-zero-
112 inflated) negative binomial (NB) as the canonical distribution for droplet scRNA-seq datasets^[6,36],

113 we first estimate subpopulation- and sample-specific means, dispersion and library size parameters
114 from the reference data set (see **Figure 1a**). Baseline multi-sample simulated scRNA-seq data
115 can then be simulated also from a NB distribution, by sampling from the subpopulation/sample-
116 specific empirical distributions of the mean, dispersion and library size. To this baseline, genes
117 can be selected as subpopulation-specific (i.e., mean different in one subpopulation versus the
118 others), or as a state gene (differential expression introduced in the samples from one condition),
119 or neither (equal relative expression across all samples and subpopulations). To introduce changes
120 in expression that represent a change in cell state, we follow the differential distribution approach
121 of Korthauer *et al.*^[33], adding changes in the mean expression (DE), changes in the proportions
122 of low and high expression-state components (DP), differential modality (DM) or changes in both
123 proportions and modality (DB). Genes that are not subject to state changes are either equivalently
124 expressed (EE), or expressed at low and high expression-states by an equal proportion (EP) of cells
125 in both conditions; see **Figure 1b**. Here, the changes are added to samples in a condition-specific
126 manner, thus mimicking a subpopulation-specific state change amongst replicates of one condition.

127 As reference datasets, we used i) scRNA-seq data of Peripheral Blood Mononuclear Cells
128 (PBMCs) from 8 lupus patients measured before and after 6h-treatment with IFN- β (16 samples
129 in total)^[20], where cells were already annotated into various immune subpopulations; and, ii) single-
130 nuclei RNA-seq data of brain cortex tissue from 8 mice split into a vehicle and lipopolysaccharide
131 treatment group. In order to introduce known state changes, simulations were based only on con-
132 trol and vehicle samples, respectively. Importantly, our simulation framework is able to reproduce
133 important characteristics of individual scRNA-seq datasets (e.g., mean-dropout and mean-variance
134 relationships) from a *countsim.QC*^[37] analysis (see **Supplementary File 1**) as well as sample-to-
135 sample variability, as illustrated by pseudobulk-level dispersion-mean trends (**Supplementary Fig.**
136 **1a**). By varying the proportion of subpopulation-specific and DS genes, we are able to generate
137 multiple subpopulations that are distinct but proximal, and clearly separated from one another in
138 lower-dimensional space (**Fig. 1c**); in particular, parameters control the distinctness of each sub-
139 population and of the group-wise state changes. Subpopulation-specific log-fold-changes (logFCs)
140 further allow modulating differential expression to be of equal magnitude across all subpopulations,
141 or such that a given subpopulation exhibits a weakened (logFC < 2), amplified (logFC > 2), or
142 null (logFC = 0) differential signal relative to the default (logFC = 2; see **Figure 1c**). Taken
143 together, we constructed a simulation that replicates aspects of individual scRNA-seq datasets,
144 mimics sample-to-sample variability and offers a high level of flexibility to introduce subpopulation-
145 specific identities (e.g., via marker genes) as well as condition-specific state changes.

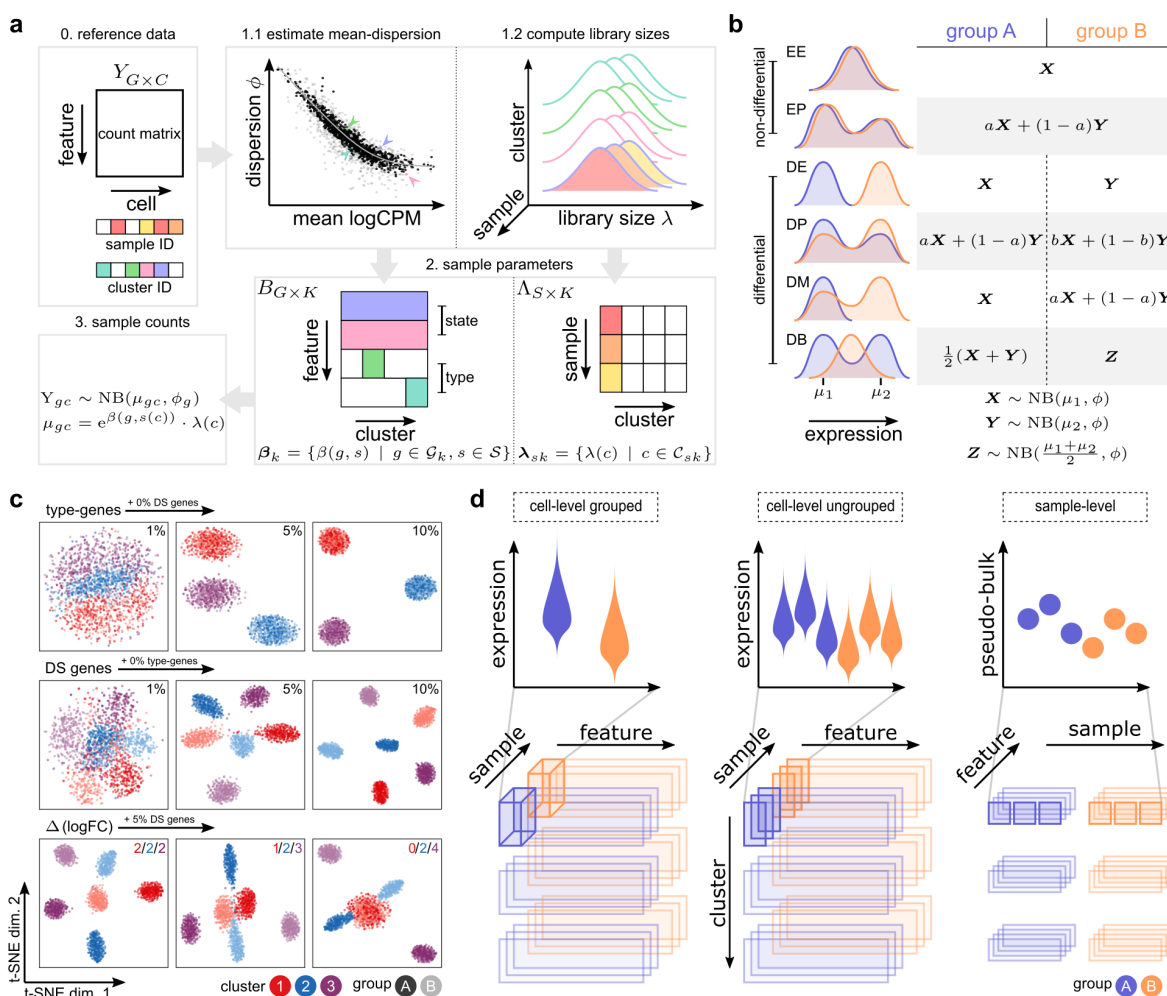


Figure 1: Schematic overview of *muscat*'s simulation framework. (a) Given a count matrix of features by cells and, for each cell, pre-determined cluster (subpopulation) identifiers as well as sample labels (0), dispersion and sample-wise means are estimated from a negative binomial distribution for 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 X , Y and 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.

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

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

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

174 To compare the ability of methods to detect DS genes, we simulated $S_1 = S_2 = 3$ samples
175 across 2 conditions. To retain the empirical distribution of library sizes, we simulated the same
176 number of genes as in the reference dataset, and selected a random subset of $G = 4,000$ genes
177 for further analysis to reduce runtimes. We simulated $K = 3$ subpopulations and introduced
178 10% of genes with DS, with equal magnitude of differential expression across subpopulations
179 ($\mathbb{E}[\log FC] = 2$) and randomly assigned to genes across the range of expression strength.

180 To ensure that method performances are comparable and do not suffer from low cell numbers,
181 we simulated an average of 200 cells per subpopulation-sample instance, amounting to a total of
182 $\sim 200 \times (S_1 + S_2) \times K \approx 3,600$ cells per simulation. Each simulation and method was repeated
183 5 times per scenario, and performances were averaged across replicates.

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

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

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

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

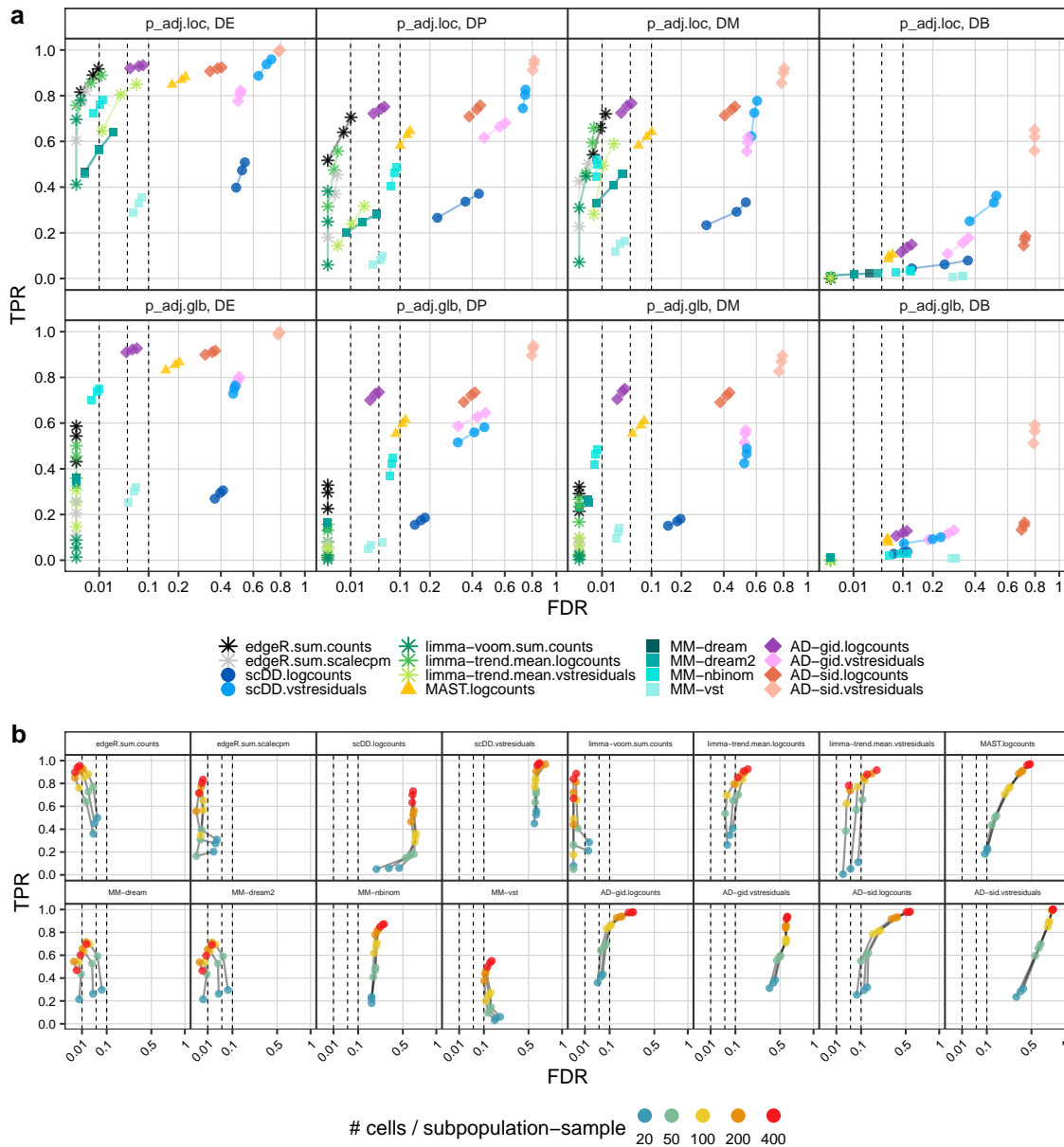


Figure 2: DS method performance across p-value adjustment types, differential distribution categories, and subpopulation-sample cell counts. All panels show observed overall true positive rate (TPR) and false discovery rate (FDR) values at FDR cutoffs of 1%, 5%, and 10%; dashed lines indicate desired 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.

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

216 To investigate the effect of subpopulation size on DS detection, we ran methods on simulations
217 containing 10% of DE genes using subsets of 20 to 400 cells per subpopulation-sample (**Fig. 2b**).
218 For most methods, FDR control varies drastically with the number of cells, while TPRs improve
219 for more cells across all methods. For aggregation-based methods, ~ 100 cells were sufficient
220 to reach decent performance; in particular, there is a sizable gain in performance in going from
221 20 to 100 cells (per subpopulation per sample), but only a moderate gain in deeper sampling
222 of subpopulations (e.g., 200 or 400 cells per subpopulation per sample). Except for *edgeR* on
223 pseudobulk summed scaled CPM, unbalanced sample and group sizes had no effect on method
224 performances (**Supplementary Figs. 6 and 7**) and increasing the number of replicates per group
225 reveals the expected, although modest, increase in detection performance (**Supplementary Fig.**
226 **8**).

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

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

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

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

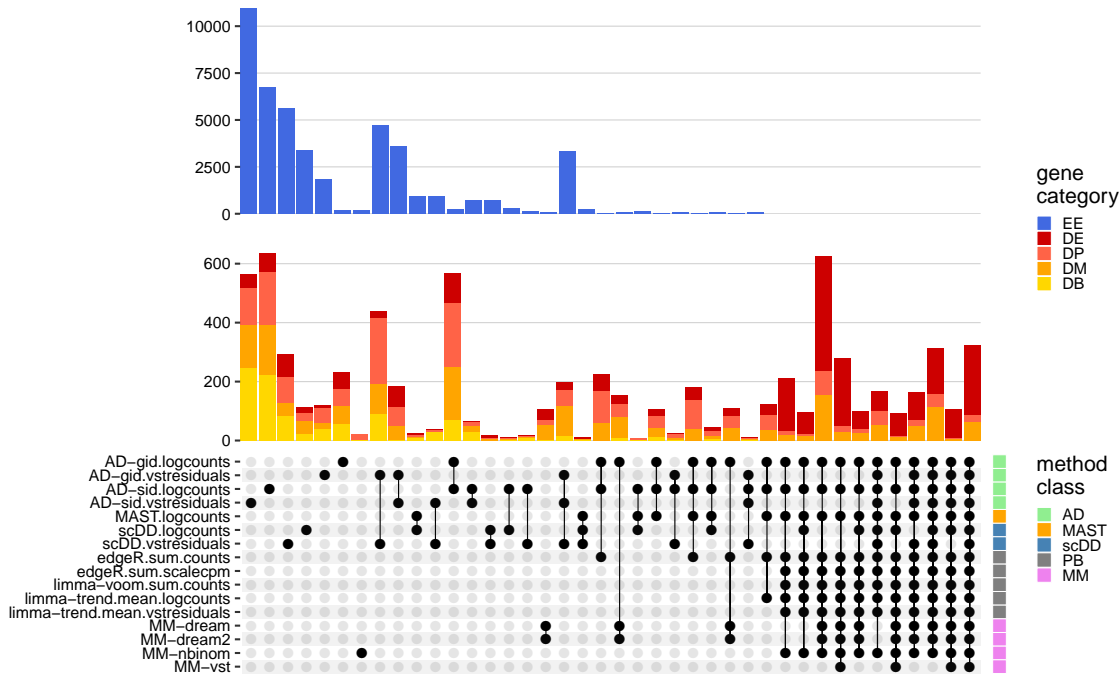


Figure 3: Between-method concordance. Upset plot obtained from intersecting the top- n ranked gene-subpopulation combinations (lowest p-value) across methods and simulation replicates. Here, $n = \min(n_1, n_2)$, where $n_1 =$ number of genes simulated to be differential, and $n_2 =$ number of genes 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).

248 in mouse frontal cortex using single-nuclei RNA-seq (snRNA-seq). The goal was to identify genes
 249 and pathways affected by LPS in neuronal and non-neuronal cells. To this end, we applied our
 250 DS analysis framework to snRNA-seq data of 4 control (vehicle) and 4 LPS-treated mice using
 251 pseudobulk (sum of counts) and *edgeR*. We obtained 12,317 vehicle and 12,907 treated cells that
 252 passed filtering and received a subpopulation assignment. Using graph-based clustering (Louvain
 253 algorithm^[49]), we identified 22 cell clusters and annotated them into 8 subpopulations (using both
 254 canonical and computationally-identified marker genes): astrocytes, endothelial cells, microglia,
 255 oligodendrocyte progenitor cells (OPC), choroid plexus ependymal (CPE) cells, oligodendrocytes,
 256 excitatory neurons, and inhibitory neurons (see Methods and **Supplementary File 3**). Low di-
 257 mensional projections of cells and pseudobulks (by subtype and condition) are shown in **Figures**
 258 **4a** through **c**; sample sizes and relative subpopulation abundances are shown in **Supplementary**
 259 **Figure 12**.

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

262 **13**). Since relying on thresholds alone is prone to bias, we next clustered the (per-subpopulation)
263 fold-changes across the union of all differentially expressed genes (**Fig. 4d**). We observed a dis-
264 tinct set of genes (consensus clustering ID 3) that were up-regulated across all subpopulations,
265 and enriched for genes associated with response to (external) biotic stimulus, defense and immune
266 response (**Supplementary File 4**). Endothelial cells appeared to be most strongly affected, fol-
267 lowed by glial cells (astrocytes, microglia and oligodendrocytes). While the responses for consensus
268 cluster 3 were largely consistent across all subpopulations, some genes' responses departed from
269 the trend (e.g., are specific to a single subpopulation or subset of subpopulations (**Supplementary**
270 **Fig. 14**).

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

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

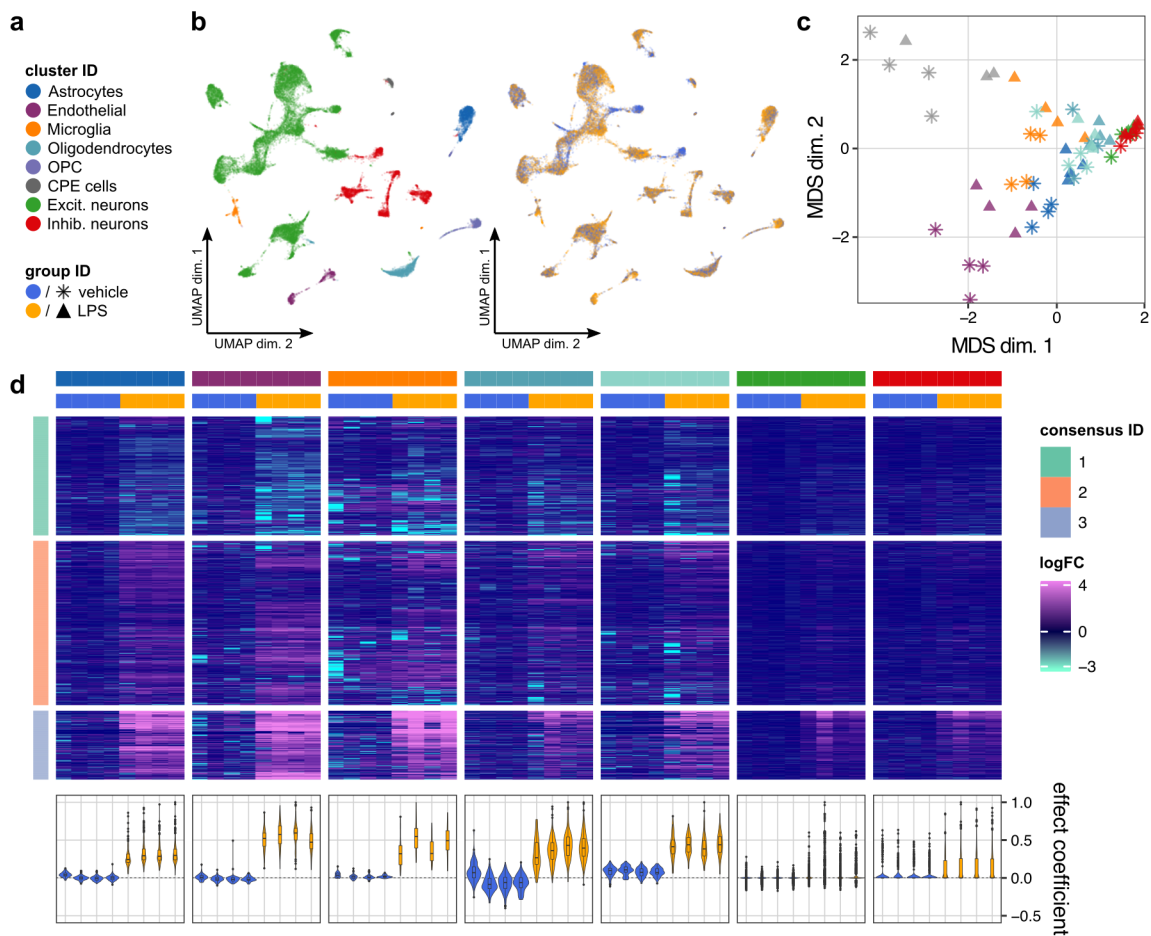


Figure 4: DS analysis of cortex tissue from vehicle- and LPS-treated mice. (a) Shared color and shape legend of subpopulation and group IDs. (b) UMAP visualization colored by subpopulation (left) and group ID (right). (c) Pseudobulk-level Multidimensional Scaling (MDS) plot. Each point represents one 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

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

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

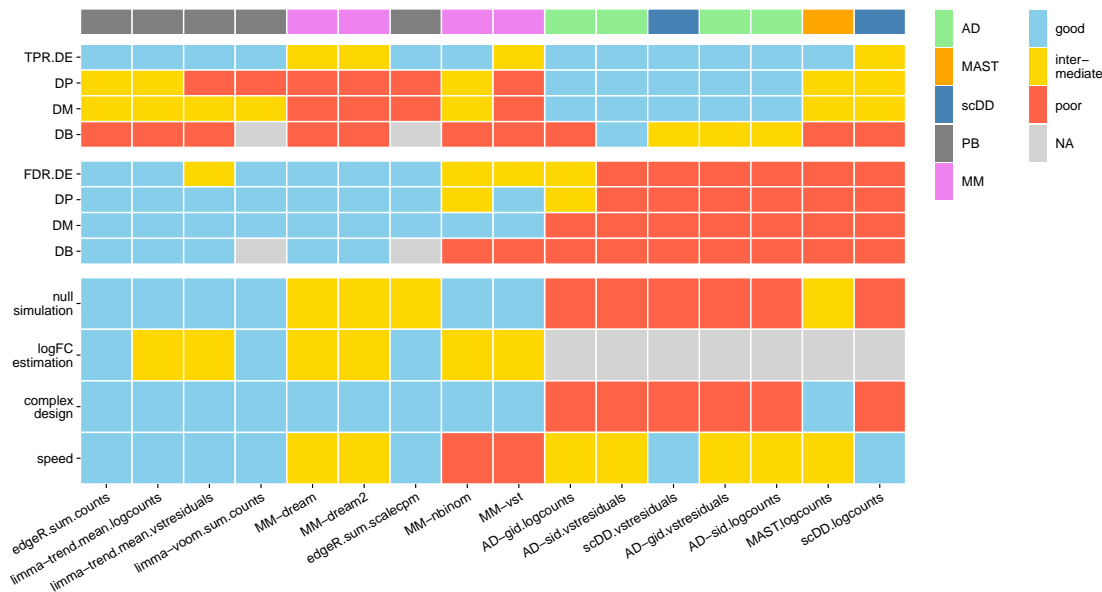


Figure 5: Summary of DS method performance across a set of evaluation criteria. Methods are ranked from left to right by their weighted average score across criteria, with the numerical encoding good = 2, intermediate = 1, and poor/NA = 0. Evaluation criteria (y-axis) comprise: DS detection sensitivity (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]: ~ 75% unimodal, ~ 5% trimodal and ~ 25% bimodal, giving weights of 0.75 for DE, 0.125 for DP and DM, and 0.05 for DB.

296 is. Based on our simulation results, the tested aggregation-based DS methods were extremely
297 fast and showed overall a stable high performance, although depending on the scale of the data
298 analyzed, logFCs were attenuated for some combinations. While mixed model methods performed
299 similarly well, their computational cost may not be worth the flexibility they provide (**Fig. 5** and
300 **Supplementary Fig. 11**). Methods developed specifically for scRNA-seq differential analysis were
301 outperformed by aggregation and mixed models, but it should be mentioned that these methods
302 focus on comparing sets of cells and were not specifically designed for the multi-group multi-sample
303 problem. Furthermore, methods that compared full distributions did not perform well overall (**Fig.**
304 **5**). This latter class of methods was used here as a reference point, but could also be improved to
305 be more targeted to the DS inference problem. For example, Anderson-Darling tests were run in
306 two ways, group-wise or sample-wise, where under the null hypothesis, all distributions are equal.
307 In the sample-wise case, departures from the null could happen between replicates of the same
308 experimental condition and in the group-wise case, it is perhaps not ideal to mix distributions from
309 different samples. Thus, while our results suggest that aggregation methods are fast and perform
310 amongst the best, there may still be value in considering full distributions, if bespoke methods
311 were developed. Furthermore, methods that integrate both changes in the mean and changes in
312 variability may be worth exploring.

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

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

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

330 geneset analysis, gene network analysis) for discovery and interpretation on each subpopulation,
331 thus leveraging existing methods.

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

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

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

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

364 data analyses, such as batch correction frameworks, clustering, reference-based cell-type inference
365 methods, marker gene selection methods as well as further developments in DS analysis.

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

Online Methods

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

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

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

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

396 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
397 of gene g in sample $s(c)$, λ_c is the library size (total number of counts), and ϕ_g is the dispersion.

398 In order to introduce a multi-subpopulation, multi-sample data structure, we sample a set
399 of K clusters as reference, as well as S reference samples for each of two groups, resulting in
400 an unpaired design. Alternatively, pairing of samples can be mimicked by fixing the same set
401 of reference samples for both groups. For each subpopulation $k \in \{1, \dots, K\}$, we sample a set
402 of genes $\mathcal{G}_k^* \subset \mathcal{G}$ used for simulation, such that most genes are common to all subpopulations
403 ($\mathcal{G}_1^* \cap \mathcal{G}_2^* \cap \dots \cap \mathcal{G}_K^* \approx (1 - p) \cdot G$), while a small set ($p \cdot 100$ percent) of *type*-specific genes
404 are sampled separately for each subpopulation ($\mathcal{G}_{k'} \cap \mathcal{G}_k = \emptyset \forall k \neq k'$), giving rise to distinct

405 subpopulations. Secondly, for each sample s and subpopulation k , we draw a set of cells $C_{sk}^* \subset C_{sk}$
406 (and their corresponding λ_c , $\beta_g^{s(c)}$ and ϕ_g) to simulate (negative binomial random variables) from,
407 where cells C_{sk} belong to the corresponding reference cluster-sample drawn previously.

408 Lastly, differential expression of a variety of types is added for a subset of genes. For each
409 subpopulation, we randomly assign each gene to a given *differential distribution* category accord-
410 ing to a probability vector $(p_{EE}, p_{EP}, p_{DE}, p_{DP}, p_{DM}, p_{DB})$; see **Figure 1b**. For each gene and
411 subpopulation, we draw a vector of fold changes from a Gamma distribution with shape 4 and
412 rate $4/\mu_{\log FC}$, where $\mu_{\log FC}$ is the desired average logFC across all genes and subpopulations. The
413 direction of differential expression is randomized for each gene, with equal probability of up- and
414 down-regulation. We split the cells in a given subpopulation-sample combination into two sets
415 (representing treatment groups), \mathcal{T}_A and \mathcal{T}_B , which are in turn split again into two sets each
416 (representing subpopulations within the given treatment group), $\mathcal{T}_{A_1}/\mathcal{T}_{A_2}$ and $\mathcal{T}_{B_1}/\mathcal{T}_{B_2}$.

417 For EE genes, counts for \mathcal{T}_A and \mathcal{T}_B are drawn using identical means. For EP genes, we multiply
418 the effective means for identical fractions of cells per group by the sampled FCs, i.e., cells are
419 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
420 group, A or B , are multiplied with the sampled FCs. DP genes are simulated analogously to EP
421 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
422 $a = 0.3, b = 0.7$). For DM genes, 50% of cells from one group are simulated at $\mu \cdot FC$. For DB
423 genes, all cells from one group are simulated at $\mu \cdot FC/2$, and the second group is split into equal
424 proportions of cells simulated at μ and $\mu \cdot FC$, respectively.

425 Details on all simulation parameters, illustrative examples of their effects, and instructions on
426 how to generate an interactive quality control report and benchmark DS methods through sim-
427 ulated data are provided in the `muscat` R/Bioconductor package's documentation (see Software
428 specification and code availability).

429 **Aggregation-based methods.** We summarize the input measurement values for a given gene over
430 all cells in each subpopulation and by sample. The resulting pseudobulk data matrix has dimensions
431 $G \times S$, where S denotes the number of samples, with one matrix obtained per subpopulation.
432 Depending on the specific method, which includes both a type of data to operate on (e.g., counts,
433 logcounts) and summary function (e.g., mean, sum), the varying number of cells between samples
434 and subpopulations is accounted for prior to or following aggregation. For logcounts methods,
435 we apply a library size normalization to the input raw counts. `vstresiduals` are computed using R
436 package `sctransform`'s `vst` function^[38]. For `scapecpm`, we calculate the total library size of each
437 subpopulation k and sample s as

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

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

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

442 edgeR-based methods were run using `glmQLFit` and `glmQLFTest`^[51]; methods based on
443 `limma-voom` and `limma-trend` were run using default parameters.

444 **Mixed models.** Mixed model methods were implemented using three main approaches: i) fitting
445 linear mixed models (LMMs) on log-normalized data with observational weights, ii) fitting
446 LMMs on variance-stabilized data, iii) fitting generalized linear mixed models (GLMMs) directly on
447 counts. Subpopulations with less than 10 cells in any sample and genes detected in fewer than 20
448 cells were excluded from differential testing. In each case, a $\sim 1 + \text{group_id} + (1|\text{sample_id})$ model
449 was fit for each gene, optimizing the restricted maximum likelihood (i.e. `REML = TRUE`), and
450 p-values were calculated using Satterthwaite estimates of degrees of freedom (the Kenward-Roger
451 approach being longer to compute and having a negligible impact on the final results). Fitting,
452 testing and moderation were applied subpopulation-wise.

453 For the first approach (`MM-dream`), we relied on the `variancePartition`^[52] package's implemen-
454 tation for repeated measurement bulk RNA-seq, using `voom`'s^[25] precision weights as originally
455 described but without empirical Bayes moderation and the `duplicateCorrelation`^[53] step, as this
456 was computationally intensive and had a negligible impact on the significance (as also observed pre-
457 viously for batch effects^[21]). Method `MM-dream2` uses an updated alternative to this approach
458 using `variancePartition`'s new weighting scheme^[Hoffman2020-variancePartition] instead of `voom`.

459 For the second approach (`MM-vst`), we first applied the variance-stabilizing transformation glob-
460 ally before splitting cells into subpopulations, and then fitted the model using the `lme4` package^[54]
461 directly on transformed data (and without observational weights). We then applied `eBayes` mod-
462 eration as in the first approach. We tested both the variance-stabilizing transformation from
463 the `DESeq2` package^[23], and that from the `scTransform` package^[38], the latter of which was
464 specifically designed for Unique Molecular Identifier (UMI) based scRNA-seq; since the latter out-
465 performed the former (data not shown), it was retained for the main results shown here.

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

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

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

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

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

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

499 anesthetizing the animals with isoflurane followed by decapitation. Brains were quickly frozen and
500 stored at -80°C .

501 **Nuclei isolation, mRNA-seq library preparation and sequencing - LPS dataset.** Nuclei were
502 prepared using the NUC201 isolation kit from Sigma Aldrich. Briefly, $8 \times 50\mu\text{m}$ sagittal sections of
503 cortex from each animal were prepared using a microtome and placed in $200\mu\text{l}$ of cold Nuclei Pure
504 Lysis Buffer (Nuclei Pure Prep Nuclei isolation kit - Sigma Aldrich) with 1M dithiothreitol (DTT)
505 and $0.2\text{U}/\mu\text{l}$ SUPERase inhibitor (Invitrogen) freshly added before use. Nuclei were extracted using
506 a glass dounce homogenizer with Teflon pestle using 10-12 up and down strokes in lysis buffer.
507 $360\mu\text{l}$ of cold 1.8M Sucrose Cushion solution was added to lysate which was then filtered through
508 a $30\mu\text{m}$ strainer. $560\mu\text{l}$ of filtered solution was carefully overlaid on $200\mu\text{l}$ of Sucrose solution and
509 nuclei were purified by centrifugation for 45min at $16,000\text{g}$. The nuclei pellet was re-suspended
510 in $50\mu\text{l}$ cold Nuclei Pure Storage Buffer (Nuclei Pure Prep Nuclei isolation kit Sigma Aldrich)
511 with $0.2\text{U}/\mu\text{l}$ SUPERase inhibitor and centrifuged for 5min at 500g . The supernatant was removed,
512 the pellet washed again with Nuclei Pure Storage Buffer with $0.2\text{U}/\mu\text{l}$ SUPERase inhibitor, and
513 centrifuged for 5min at 500g . Finally, the pellet was re-suspended in $50\mu\text{l}$ cold Nuclei Pure
514 Storage Buffer with $0.2\text{U}/\mu\text{l}$ SUPERase inhibitor. Nuclei were counted using trypan blue staining
515 on Countess II (Life technology). A total of 12,000 estimated nuclei from each sample was loaded
516 on the $10\times$ Single Cell B Chip.

517 cDNA libraries from each sample were prepared using the Chromium Single Cell 3' Library and
518 Gel Bead kit v3 ($10\times$ Genomics) according to the manufacturers instructions. cDNA libraries were
519 sequenced using Illumina HiSeq 4000 using the HiSeq 3000/4000 SBS kit (Illumina) and HiSeq
520 3000/4000 PE cluster kit to get a sequencing depth of 30K reads/nuclei.

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

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

531 R package. Upon removal of genes that were undetected across all cells, we removed cells whose
532 feature counts, number of expressed features, and percentage of mitochondrial genes fell beyond
533 2.5 Median Absolute Deviations (MADs) of the median. Finally, features with a count > 1 in at
534 least 20 cells were retained for downstream analysis.

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

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

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

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

- 562 • *TPR/FDR*: For each type of DD category, we consider TPRs and FDR at FDR 5% averaged
563 across two references, 5 simulation replicates and 3 clusters (**Fig. 2a**). Methods are scored

564 according to $\overline{TPR} > 2/3$: good, $> 1/3$: intermediate, otherwise: poor; and $\overline{FDR} < 0.05$:
565 good, < 0.1 : intermediate, otherwise: poor.

566 • *null simulation*: We perform a Kolmogorov-Smirnov (KS) test on the uniformity of p-values
567 (`ks.test` with CDF `y = "punif"`) under the null (**Supplementary Fig. 2**) for each of two
568 references, three simulation replicates and three clusters per simulation, resulting in a total
569 of 18 tests per method. KS statistics (largest difference between observed and uniform
570 empirical cumulative distribution functions) are then averaged, and categorized according to
571 $\overline{KS}_{stat.} < 0.1$: good, < 0.25 : intermediate, otherwise: poor.

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

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

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

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

591 **Software specifications and code availability.** All analyses were run in R v3.6.2^[63], with
592 Bioconductor v3.10^[64]. Performance measures were calculated using `iCOBRA`^[65], and results
593 were visualized with `ggplot2`^[66], `ComplexHeatmap`^[67], and `UpSetR`^[68]. All package versions used
594 throughout this study are captured in **Supplementary File 5**. Data preprocessing, simulation and
595 analysis code are accessible at <https://github.com/HelenaLC/muscat-comparison>, including

596 a browseable workflow^[69] website for the LPS dataset analysis (**Supplementary File 3**). All
597 aggregation and DS analysis methods are provided in the `muscat` R package, which is available
598 at <https://www.bioconductor.org/packages/muscat> through the open-source Bioconductor
599 project.

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

Acknowledgments

607 The authors thank members of the Robinson Lab at the University of Zurich for valuable feedback
608 on methodology, benchmarking and exposition. This work was supported by the Swiss National
609 Science Foundation (grant numbers 310030_175841, CRSII5_177208) and the Chan Zuckerberg
610 Initiative DAF, an advised fund of Silicon Valley Community Foundation (grant number 2018-
611 182828). MDR acknowledges support from the University Research Priority Program Evolution in
612 Action at the University of Zurich.

Author contributions

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

Competing interests

620 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. Sonesson, 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., Sonesson, 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., Kendzioriski, 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. Sonesson, 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. & Sonesson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* **7**, 1141 (2018).
41. Freytag, S., Tian, L., Lönstedt, 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., Sonesson, 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. Sonesson, 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 workflow way. *F1000Research* **8**, 1749 (2019).