

This is the final peer-reviewed accepted manuscript of:

**Simone Tiberi, Helena L. Crowell, Pantelis Samartsidis, Lukas M. Weber, Mark D. Robinson. (2023). “*distinct: A novel approach to differential distribution analyses*”. *Annals of Applied Statistics*, Vol. 17, Issue 2, June 2023, pp. 1681-1700.**

The final published version is available online at:

<https://doi.org/10.1214/22-AOAS1689>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

*This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)*

***When citing, please refer to the published version.***

# *distinct*: a novel approach to differential distribution analyses

Simone Tiberi<sup>1\*</sup>, Helena L Crowell<sup>1</sup>, Pantelis Samartsidis<sup>2</sup>, Lukas M Weber<sup>3</sup> and Mark D Robinson<sup>1</sup>

<sup>1</sup>*Department of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland.*

<sup>2</sup>*MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK.*

<sup>3</sup>*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.*

\* e-mail: Simone.Tiberi@uzh.ch

## 1 Abstract

2 We present *distinct*, a general method for dif-  
3 ferential analysis of full distributions that is  
4 well suited to applications on single-cell data,  
5 such as single-cell RNA sequencing and high-  
6 dimensional flow or mass cytometry data. High-  
7 throughput single-cell data reveal an unprece-  
8 dented view of cell identity and allow com-  
9 plex variations between conditions to be discov-  
10 ered; nonetheless, most methods for differential  
11 expression target differences in the mean and  
12 struggle to identify changes where the mean is  
13 only marginally affected. *distinct* is based on  
14 a hierarchical non-parametric permutation ap-  
15 proach and, by comparing empirical cumulative  
16 distribution functions, identifies both differen-  
17 tial patterns involving changes in the mean, as  
18 well as more subtle variations that do not in-  
19 volve the mean. We performed extensive bench-  
20 marks across both simulated and experimen-  
21 tal datasets from single-cell RNA sequencing  
22 and mass cytometry data, where *distinct* shows  
23 favourable performance, identifies more differ-  
24 ential patterns than competitors, and displays  
25 good control of false positive and false discovery  
26 rates. *distinct* is available as a Bioconductor R  
27 package.

28 **keywords:** Differential distribution; Differential anal-  
29 yses; Differential state; High-throughput single-cell  
30 data; Single-cell RNA-seq; Single-cell flow and mass cy-  
31 tometry; Permutation tests.

## 32 Background

33 Technology developments in the last decade have led to  
34 an explosion of high-throughput single-cell data, such  
35 as single-cell RNA sequencing (scRNA-seq) and high-  
36 dimensional flow or mass cytometry data, allowing re-

37 searchers to investigate biological mechanisms at single-  
38 cell resolution. Single-cell data have also extended the  
39 canonical definition of differential expression by dis-  
40 playing cell-type specific responses across conditions,  
41 known as differential state (DS) [32], where genes or  
42 proteins vary in specific sub-populations of cells (e.g.,  
43 a cytokine response in myeloid cells but not in other  
44 leukocytes [13]). Classical bulk differential expression  
45 methods have been shown to perform well when used  
46 on single-cell measurements [25, 26, 31] and on aggre-  
47 gated data (i.e., averages or sums across cells), also re-  
48 ferred to as pseudo-bulk (PB) [7, 32]. However, most  
49 bulk and PB tools focus on shifts in the means, and  
50 may conceal information about cell-to-cell heterogene-  
51 ity. Indeed, single-cell data can show more complex  
52 variations (Figure 1 and Supplementary Figure 1); such  
53 patterns can arise due to increased stochasticity and  
54 heterogeneity, for example owing to oscillatory and un-  
55 synchronized gene expression between cells, or when  
56 some cells respond differently to a treatment than oth-  
57 ers [15, 31]. In addition to bulk and PB tools, other  
58 methods were specifically proposed to perform differ-  
59 ential analyses on single-cell data (notably: *scDD* [15],  
60 *SCDE* [14], *MAST* [11], *BASiCS* [10, 29, 30] and mixed  
61 models [27]). Nevertheless, they all present significant  
62 limitations: *BASiCS* does not perform cell-type spe-  
63 cific differential testing between conditions, *scDD* does  
64 not directly handle covariates and biological replicates,  
65 while *PB*, *SCDE*, *MAST* and mixed models performed  
66 poorly in previous benchmarks when detecting differ-  
67 ential patterns that do not involve the mean [7, 15].

## 68 Results

### 69 *distinct*'s full distribution approach

70 To overcome these challenges, we developed *distinct*, a  
71 flexible and general statistical methodology to perform  
72 differential analyses between groups of distributions.

73 *distinct* is particularly suitable to compare groups of  
 74 samples (i.e., biological replicates) on single-cell data.

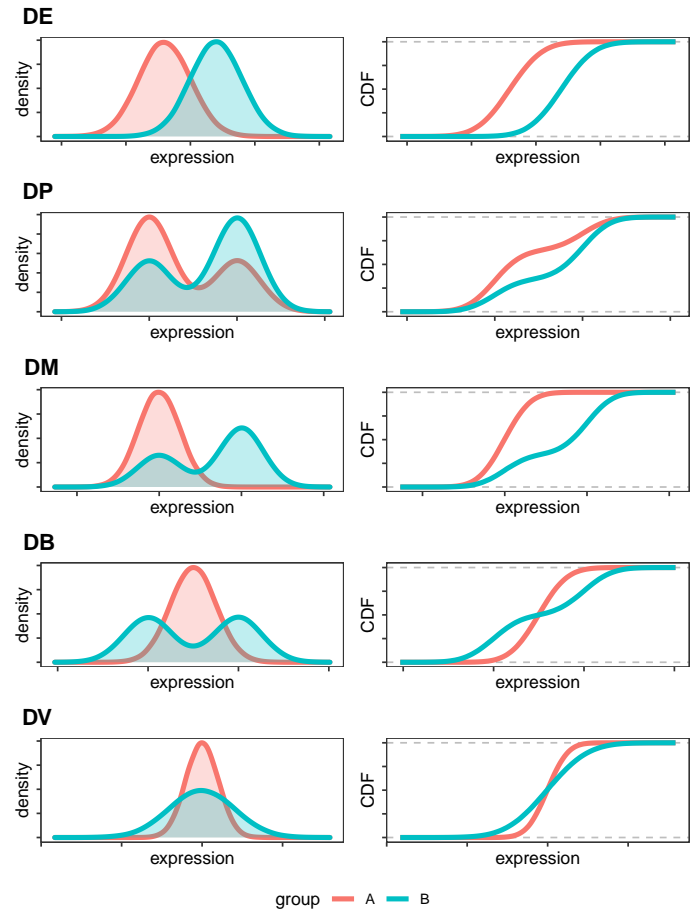
75 Our approach computes the empirical cumulative dis-  
 76 tribution function (ECDF) from the individual (e.g.,  
 77 single-cell) measurements of each sample, and compares  
 78 ECDFs to identify changes between full distributions,  
 79 even when the mean is unchanged or marginally in-  
 80 volved (Figure 1 and Supplementary Figure 1). First,  
 81 we compute the ECDF of each individual sample; then,  
 82 we build a fine grid and, at each cut-off, we average the  
 83 ECDFs within each group, and compute the absolute  
 84 difference between such averages. A test statistic,  $s^{obs}$ ,  
 85 is obtained by adding these absolute differences.

86 More formally, assume we are interested in compar-  
 87 ing two groups, that we call  $A$  and  $B$ , for which  $N_A$   
 88 and  $N_B$  samples are available, respectively. The ECDF  
 89 for the  $i$ -th sample in the  $j$ -th group, is denoted by  
 90  $ecdf_i^{(j)}(\cdot)$ , for  $j \in \{A, B\}$  and  $i = 1, \dots, N_j$ . We  
 91 then define  $K$  equally spaced cut-offs between the mini-  
 92 mum,  $min$ , and maximum,  $max$ , values observed across  
 93 all samples:  $b_1, \dots, b_K$ , where  $b_k = min + k \times l$ , for  
 94  $k = 1, \dots, K$ , with  $l = (max - min)/(K + 1)$  being  
 95 the distance between two consecutive cut-offs. We ex-  
 96 clude  $min$  and  $max$  from the cut-offs because, trivially,  
 97  $ecdf_i^{(j)}(min) = 0$  and  $ecdf_i^{(j)}(max) = 1, \forall j, i$ . At ev-  
 98 ery cut-off, we compute the absolute difference between  
 99 the mean ECDF in the two groups; our test statistic,  
 100  $s^{obs}$ , is obtained by adding these differences across all  
 101 cut-offs:

$$s^{obs} = \sum_{k=1}^K \left| \frac{\sum_{i=1}^{N_A} ecdf_i^{(A)}(b_k)}{N_A} - \frac{\sum_{i=1}^{N_B} ecdf_i^{(B)}(b_k)}{N_B} \right|. \quad (1)$$

102 Note that in differential state analyses, these operations  
 103 are repeated for every gene-cluster combination.

104 Intuitively,  $s^{obs}$ , which ranges in  $[0, \infty)$ , approximates  
 105 the area between the average ECDFs, and represents  
 106 a measure of distance between two groups of densities:  
 107 the bigger  $s^{obs}$ , the greater the distance between groups.  
 108 The number of cut-offs  $K$ , which can be defined by  
 109 users, is set to 25 by default, because no detectable  
 110 difference in performance was observed when further  
 111 increasing it (data not shown). Note that, although at  
 112 each cut-off we compute the average across each group's  
 113 curves, ECDFs are computed separately for each indi-  
 114 vidual sample, therefore our approach still accounts for  
 115 the within-group variability; indeed, at a given thresh-  
 116 old, the average of the sample-specific ECDFs differs  
 117 from the group-level ECDF (i.e., the curve based on  
 118 all individual measurements from the group). The null  
 119 distribution of  $s^{obs}$  is then estimated via a hierarchical



**Figure 1: Cumulative distribution functions (CDFs) unravel differences between distributions.** Density (left panels) and CDF (right panels) of five differential patterns: differential variability (DV), and the four proposed by Korthauer et. al. [15]: differential expression (DE), differential proportion (DP), differential modality (DM), and both differential modality and different component means (DB).

120 non-parametric permutation approach (see Methods).  
 121 A major disadvantage of permutation tests, which of-  
 122 ten restricts its usage on biological data, is that too  
 123 few permutations are available from small samples. We  
 124 overcome this by permuting cells, which is still pos-  
 125 sible in small samples, because there are many more  
 126 cells than samples. In principle, this may lead to an  
 127 inflation of false positives due to lack of exchangeabil-  
 128 ity (see Methods); nonetheless, in our analyses, *distinct*  
 129 provides good control of both false positive and false  
 130 discovery rates.

131 Importantly, *distinct* is general and flexible: it targets  
 132 complex changes between groups, explicitly models bio-  
 133 logical replicates within a hierarchical framework, does  
 134 not rely on asymptotic theory, avoids parametric as-  
 135 sumptions, and can be applied to arbitrary types of  
 136 data. Additionally, *distinct* can also adjust for sample-  
 137 level cell-cluster specific covariates (i.e., whose effect  
 138 varies across cell clusters), such as batch effects. In par-  
 139 ticular, *distinct* fits a linear mixed effects model with

140 the input data (e.g., normalized counts) as response  
141 variable, nuisance covariates as fixed effects, and sam-  
142 ples as random effects. The method then removes the  
143 estimated impact of fixed effect covariates, and per-  
144 forms differential testing on these normalized values  
145 (see Methods).

146 Furthermore, to enhance the interpretability of differen-  
147 tial results, *distinct* provides functionalities to compute  
148 (log) fold changes between conditions, and to plot den-  
149 sities and ECDFs, both for individual samples and at  
150 the group-level.

151 Note that, although *distinct* and the Kolmogorov-  
152 Smirnov [18] (KS) test share similarities (they both  
153 compare distributions via non-parametric tests), the  
154 two approaches present several conceptual differences.  
155 Firstly, the KS considers the maximum distance be-  
156 tween two ECDFs, while our approach estimates the  
157 overall distance between ECDFs, which in our view is  
158 a more appropriate way to measure the difference be-  
159 tween distributions. Secondly, the KS test only com-  
160 pares two individual densities, while our framework  
161 compares groups of distributions. Thirdly, while the  
162 KS statistic relies on asymptotic theory, our framework  
163 uses a permutation test. Finally, a comparison between  
164 *distinct* and *scDD* [15] based on the KS test (labelled  
165 *scDD-KS*) shows that our method, compared to the KS  
166 test, has greater statistical power to detect differential  
167 effects and leads to fewer false discoveries (see Simula-  
168 tion studies).

## 169 Simulation studies

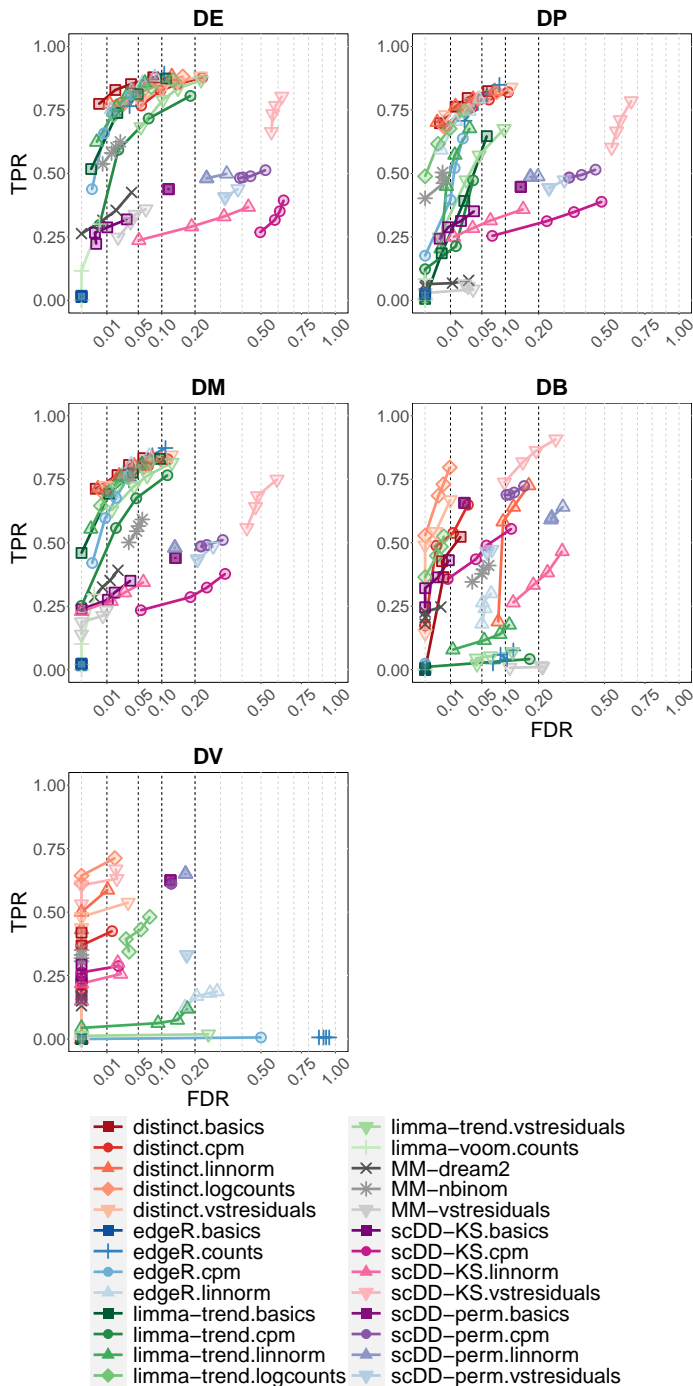
170 We conducted an extensive benchmark, based on  
171 scRNA-seq and mass cytometry simulated and experi-  
172 mental datasets to investigate *distinct*'s ability to iden-  
173 tify differential patterns in sub-populations of cells.

174 First, we simulated droplet scRNA-seq data via *mus-*  
175 *cat* [7] (see Methods). We ran five simulation repli-  
176 cates for each of the differential profiles in Figure 1,  
177 with 10% of the genes being differential in each clus-  
178 ter, where DE (differential expression) indicates a shift  
179 in the entire distribution, DP (differential proportion)  
180 implies two mixture distributions with different propor-  
181 tions of the two components, DM (differential modal-  
182 ity) assumes a unimodal and a bimodal distribution,  
183 DB (both differential modality and different component  
184 means) compares a unimodal and a bimodal distribu-  
185 tion with the same overall mean, and DV (differential  
186 variability) refers to two unimodal distributions with  
187 the same mean but different variance (Figure 1 and  
188 Supplementary Figure 1). Each individual simulation  
189 consists of 4,000 genes, 3,600 cells, separated into 3 clus-

ters, and two groups of 3 samples each, corresponding  
191 to an average of 200 cells per sample in each cluster.

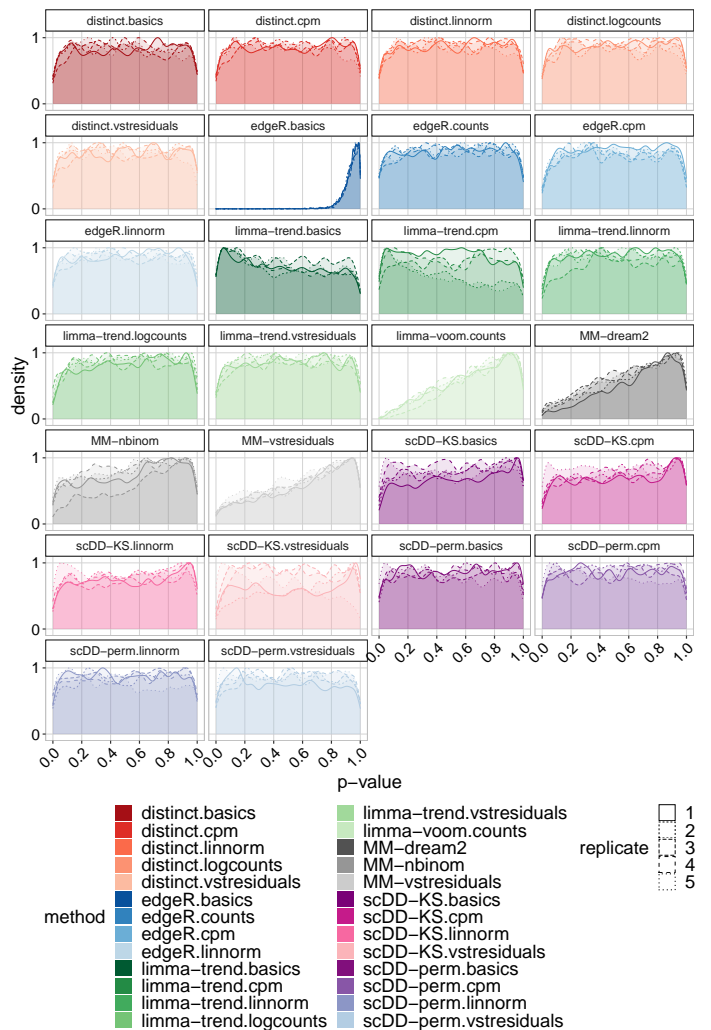
192 We considered six different normalization approaches:  
193 counts per million (CPMs), *scater*'s logcounts [19],  
194 *linnorm* [34], *BASiCS* [10,29,30], *SCnorm* [3] and  
195 residuals from variance stabilizing normalization from  
196 *sctransform* (vstresiduals) [12]. We compared *dis-*  
197 *tinct* to several PB approaches from *muscat*, based on  
198 *edgeR* [24], *limma-voom* and *limma-trend* [23], which  
199 emerged among the best performing methods for differ-  
200 ential analyses from scRNA-seq data [7,26]. We further  
201 considered three methods from *muscat* based on mixed  
202 models (MM), namely *MM-dream2*, *MM-vstresiduals*  
203 and *MM-nbinom* (see Methods). Finally, we included  
204 *scDD* [15], which is conceptually similar to our ap-  
205 proach: *scDD* implements a non-parametric method to  
206 detect changes between individual distributions from  
207 scRNA-seq, based on the Kolmogorov-Smirnov test,  
208 *scDD-KS*, and on a permutation approach, *scDD-perm*.  
209 For *scDD-perm* we used 100 permutations to reduce the  
210 computational burden.

211 In all scenarios and on all six input data, *distinct* shows  
212 favourable performance: it has good statistical power  
213 while controlling for the false discovery rate (FDR)  
214 (Figure 2). In particular, for DE, DP and DM, *distinct*  
215 has similar performance to the best performing com-  
216 petitors (*edgeR.linnorm* and *limma-trend.logcounts*),  
217 while for DB and DV, it achieves significantly higher  
218 true positive rate (TPR), especially when using *log-*  
219 *counts*. PB methods in general perform well for differ-  
220 ential patterns involving changes in the mean (DE, DP  
221 and DM), but struggle to identify DB and DV patterns.  
222 *scDD* provides good TPR across all patterns when us-  
223 ing the KS test on vstresiduals (*scDD-KS.vstresiduals*),  
224 while the TPR is significantly reduced when using  
225 other inputs and with the permutation approach(*scDD-*  
226 *perm*); however, *scDD* methods (in particular, *scDD-*  
227 *KS.vstresiduals*) also show a significant inflation of the  
228 FDR. In contrast, MM methods provide good control of  
229 the FDR but have low statistical power in all differen-  
230 tial scenarios. We also investigated how normalization  
231 influences each method's results (Supplementary Fig-  
232 ure 2): *distinct* appears to be the least affected method  
233 and displays the smallest variation across normaliza-  
234 tion inputs, possibly due to its non-parametric struc-  
235 ture, which can more flexibly accommodate various in-  
236 puts. Given the computational cost of *SCnorm*, which  
237 is significantly higher than the other normalizations,  
238 we only included this approach in the results from the  
239 main simulations. Furthermore, among the 25 replicate  
240 datasets in Figure 2, *SCnorm* ran in a few minutes on  
241 10 simulations, while it failed to run within a week time



**Figure 2: *distinct* identifies various differential patterns and controls for the FDR.** TPR vs. FDR in *muscat* simulated data; DE, DP, DM, DB and DV refer to the differential profiles illustrated in Figure 1. Circles indicate observed FDR for 0.01, 0.05, 0.1 and 0.2 significance thresholds. Results are averages across the five simulation replicates. Each individual replicate consists of 4,000 genes, 3,600 cells, separated into 3 clusters, and two groups of 3 samples each, corresponding to an average of 200 cells per sample in each cluster.

(on 10 cores) on the remaining 15 datasets. Therefore, we excluded *SCnorm* from Figure 2 and, in Supplementary Figures 3 and 4, we report a comparison of *SCnorm* to the remaining normalization methods, on the subset of 10 simulations where all normalizations successfully ran. For *distinct*, *edgeR* and *limma*, no noticeable differences are detected between *SCnorm* and

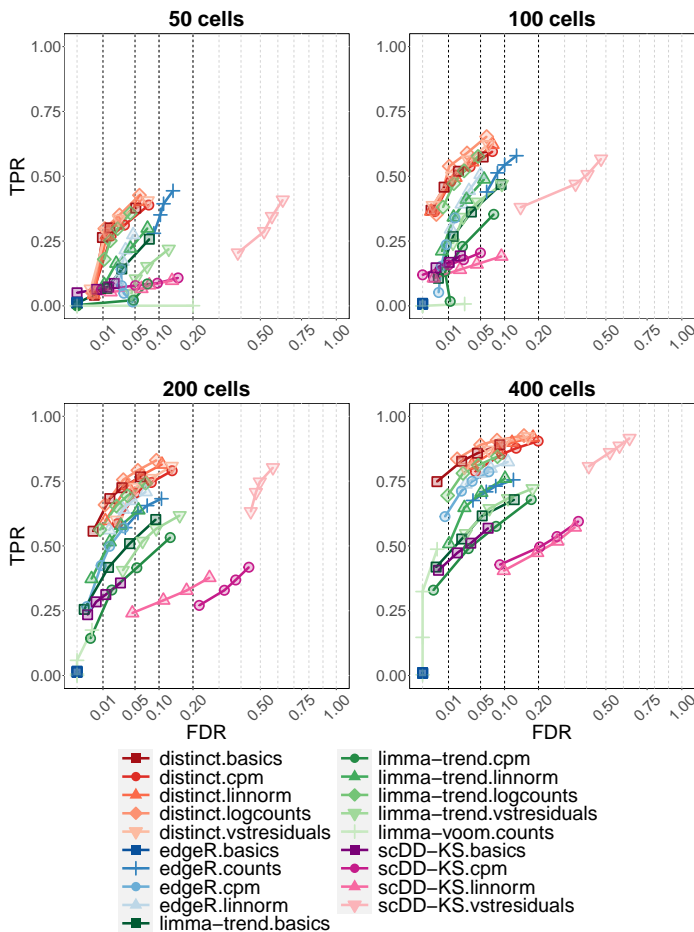


**Figure 3: *distinct* has uniform null p-values.** Density of raw p-values in *muscat* null simulated data; each replicate represents a different null simulation. Each individual replicate consists of 4,000 genes, 3,600 cells, separated into 3 clusters, and two groups of 3 samples each, corresponding to an average of 200 cells per sample in each cluster.

the remaining normalization methods, while for *scDD-KS SCnorm* leads to a higher inflation of the FDR.

We further simulated five null simulation replicates with no differential patterns; again with each simulation having 4,000 genes, 3,600 cells, 3 cell clusters and two groups of 3 samples each. In the null simulated data, only *limma-trend.basics* and *limma-trend.cpm* present a mild inflation of false positives, while MM and, particularly, *edgeR.basics* lead to overly conservative p-values; instead, *distinct* and *scDD* show approximately uniform p-values for all types of input data (Figure 3).

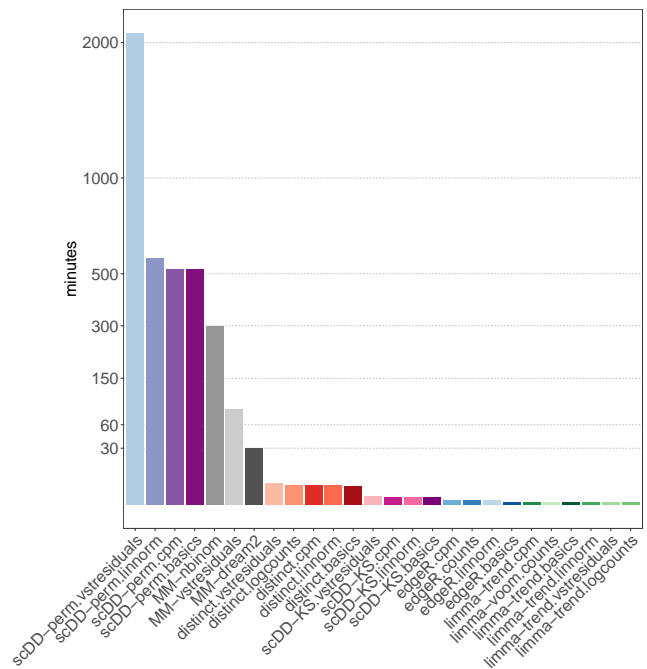
We also extended previous simulations to add a cell-type specific batch effect (i.e., a batch effect that affects differently each cell-type) [7, 17]. In particular, we simulated 2 batches, that we call  $b_1$  and  $b_2$ , with one group of samples having two samples associated to  $b_1$  and one to  $b_2$ , and the other group of samples having two samples from batch  $b_2$  and one from  $b_1$ . Differential results



**Figure 4:** *distinct* achieves good performance when varying the number of available cells. TPR vs. FDR in *muscat* simulated data; with 50, 100, 200 and 400 cells per cluster-sample combination, corresponding to a total of 900, 1,800, 3,600 and 7,200 cells, respectively. Results are aggregated over the five replicate simulations of each differential type (DE, DP, DM, DB and DV), contributing in equal fraction. Each individual simulation replicate consists of 4,000 genes, 3 cell clusters and two groups of 3 samples each. Circles indicate observed FDR for 0.01, 0.05, 0.1 and 0.2 significance thresholds. Note that *scDD-perm* and MM were excluded from this analysis due to their computational cost.

are substantially unchanged (Supplementary Figure 5), which shows *distinct* can effectively remove nuisance confounders.

Furthermore, we performed various sensitivity analyses and investigated how results are affected when varying: i) the number of cells, ii) the library size, iii) the dispersion parameter, iv) the fraction of significant genes, and v) the sample sizes in each group. In particular, we simulated 50, 100, 200 (as in the original simulation) and 400 cells per sample in each cluster. We further modified the library size and dispersion parameters of the negative binomial model used by *muscat* to simulate scRNA-seq data, influencing the mean expression and cell-to-cell variability respectively, by considering values 1/5, 1/2, 2 and 5 times as big as those used in the original simulation. In addition, we varied the per-



**Figure 5:** *distinct* requires more computational resources than PB and *scDD-KS* methods, but significantly less than MM and *scDD-perm* models. Average computing time, expressed in minutes, in *muscat* main simulations (Figures 2-3). For each method, times are averaged across simulation types (DE, DP, DM, DB, DV and null) and, for each type, across the five replicate simulations; in each replicate 3,600 cells are available (200, on average, per cluster-sample combination). *distinct*, MM and *scDD* models were run on 3 cores, while pseudo-bulk methods based on *edgeR* and *limma* used a single core because they do not allow for parallel computing. Note that *scDD-perm* requires much longer on vstresiduals than on the other normalized data, because *scDD* performs differential testing on non-zero values: vstresiduals, (unlike linnorm, cpm and basics normalized data) are not zero-inflated and, therefore, many more cells have to be used for differential testing.

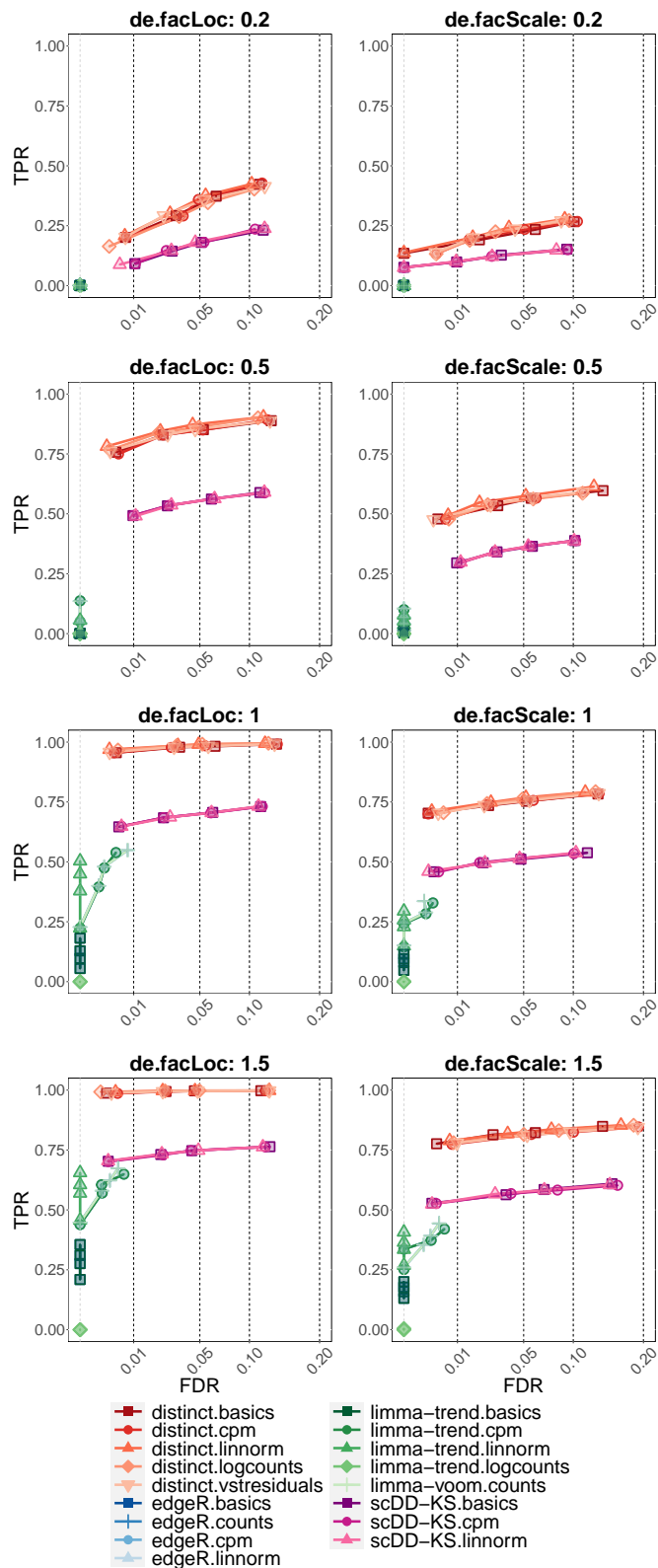
percentage of simulated differential genes as 1, 5, 10 (as in the original simulation) and 20%, and considered various unbalanced designs by comparing two groups of different sample sizes: 3 vs. 2, 4 vs. 3, and 5 vs. 3. Overall, increasing the number of cells or the library size and decreasing the dispersion have a positive impact on the performance of all methods, by improving their ability to detect differential effects (i.e., true positive rate); nonetheless, none of these factors seem to affect the relative ranking of methods, which remains globally stable (Figure 4 and Supplementary Figures 6-7). In addition, changing the fraction of significant genes and considering unbalanced designs does not appear to introduce systematic changes in performance (Supplementary Figures 8-9). Note that, in these sensitivity analyses, we excluded MM models due to the high computational cost and low statistical power displayed in the previous analyses.

From a computational perspective, *distinct* required an average time of 3.2 to 4.5 minutes per simulation, which is higher than PB methods (0.1 to 0.2 minutes)

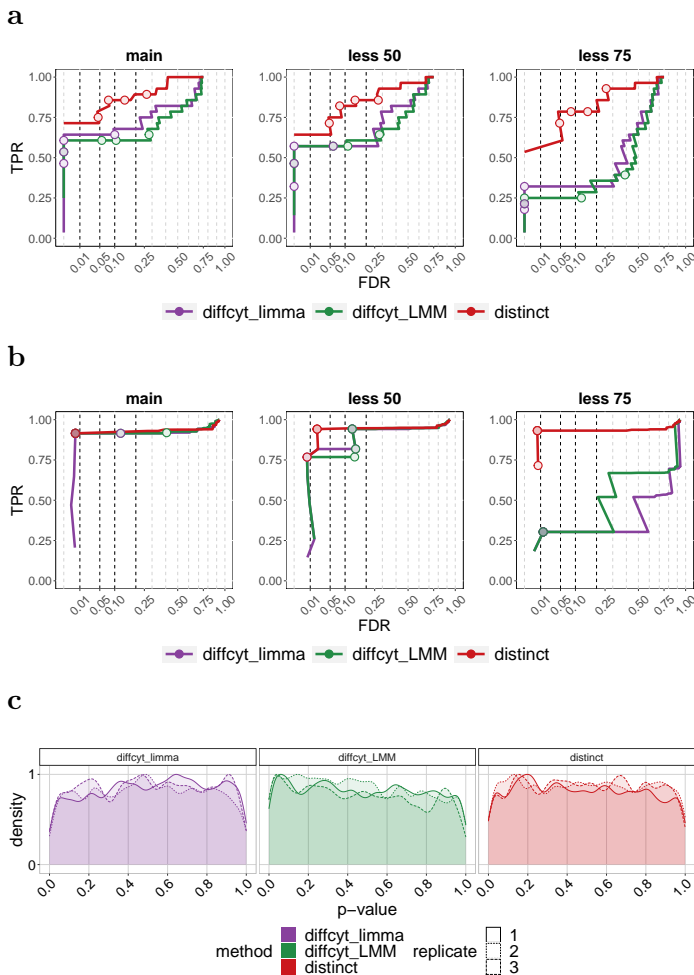
304 and *scDD-KS* (0.5 to 0.7 minutes), but significantly  
 305 lower than MM approaches (29.4 to 297.3 minutes) and  
 306 *scDD-perm* (544.7 to 2085.6 minutes) (Figure 5 and  
 307 Supplementary Table 1). All methods were run on 3  
 308 cores, except PB approaches, which used a single core,  
 309 because they do not allow for parallel computing.

310 We also considered an alternative popular droplet  
 311 scRNA-seq data simulator, *SplatPOP* [2], which rep-  
 312 resents a generalization of *Splatter* [35], that allows  
 313 multi-sample multi-group synthetic data to be gener-  
 314 ated. In particular, we simulated 20,345 genes from  
 315 a human genome with two groups of 4 samples each,  
 316 and 100 cells per sample, belonging to the same cluster  
 317 of cells, for a total of 800 cells across all samples.  
 318 We ran 8 differential simulations, with 10% of genes  
 319 truly differential between groups, by varying the location  
 320 (*de.facLoc*) and scale (*de.facScale*) differential  
 321 parameters, mainly affecting the mean and variance,  
 322 respectively (see Methods). We considered the same  
 323 normalization and differential methods as in the *mus-*  
 324 *cat* simulation (except MM and *scDD-perm*, which were  
 325 not considered due to the high computational cost and  
 326 low statistical power displayed above). As expected, for  
 327 all methods, differential patterns are easier to detect as  
 328 the magnitude of the difference increases, with differ-  
 329 ential location patterns having a higher true positive rate  
 330 than differential scale patterns. While all methods control  
 331 the FDR, in all simulations, *distinct* achieves subst-  
 332 antially higher TPR than competitors (Figure 6). We  
 333 also repeated the same simulations including a batch  
 334 effect, with two batches, with the same scale and lo-  
 335 cation differential parameters for the batch and group  
 336 differences (i.e., increasing together from 0.2 to 1.5).  
 337 Again, we excluded *scDD* from these analyses because  
 338 it cannot handle covariates directly. Results agree with  
 339 those from the *muscat* batch effect simulation study:  
 340 FDR and TPRs are mostly unchanged when introduc-  
 341 ing nuisance covariates, with only a minor decrease in  
 342 the TPR in stronger batch effects, i.e., when *de.facLoc*  
 343 and *de.facScale* are 1 and 1.5 (Supplementary Figure  
 344 10), which again indicates that *distinct* can effectively  
 345 control for nuisance covariates.

346 We further considered the semi-simulated mass cytometry  
 347 data from Weber *et al.* [32] (labelled *diffcyt* sim-  
 348 ulation), where spike-in signals were computationally  
 349 introduced in experimental data [5], hence maintain-  
 350 ing the properties of real biological data while also  
 351 embedding a known ground truth signal. We evalu-  
 352 ated *distinct* and two methods from *diffcyt*, based on  
 353 *limma* [23] and linear mixed models (LMM), which out-  
 354 performed competitors on these same data [32]. In  
 355 particular, we considered three datasets from Weber



**Figure 6: *distinct* displays higher TPR than competitors.** TPR vs. FDR in *SplatPop* simulated data, with various degrees of differential location (left) and scale (right) parameters, primarily affecting the mean and variance, respectively. Circles indicate observed FDR for 0.01, 0.05, 0.1 and 0.2 significance thresholds. Each simulation consists of 20,345 genes, 800 cells (belonging to the same cluster), and two groups of 4 samples each, corresponding to an average of 100 cells per sample.



**Figure 7: *distinct* shows high power while controlling for false positive and false discovery rates.** (a-b) TPR vs. FDR in *diffcyt* semi-simulated data. ‘main’, ‘less 50’ and ‘less 75’ indicate the main simulation, and those where differential effects are diluted by 50 and 75%, respectively. Each simulation consists of 88,435 cells and two groups of 8 samples each. Circles indicate observed FDR for 0.01, 0.05, 0.1 and 0.2 significance thresholds. (a) As in the *muscat* simulation study, cells were clustered into 8 populations based on manually annotated cell types [32]. (b) As in Weber *et al.* [32], cells were grouped in 100 high-resolution clusters via unsupervised clustering. (c) Density of raw p-values in *diffcyt* null semi-simulated data; each replicate represents a different null simulation. Each replicate consists of 88,438 cells and two groups of 8 samples each. As in Weber *et al.* [32], cells were clustered in an unsupervised manner.

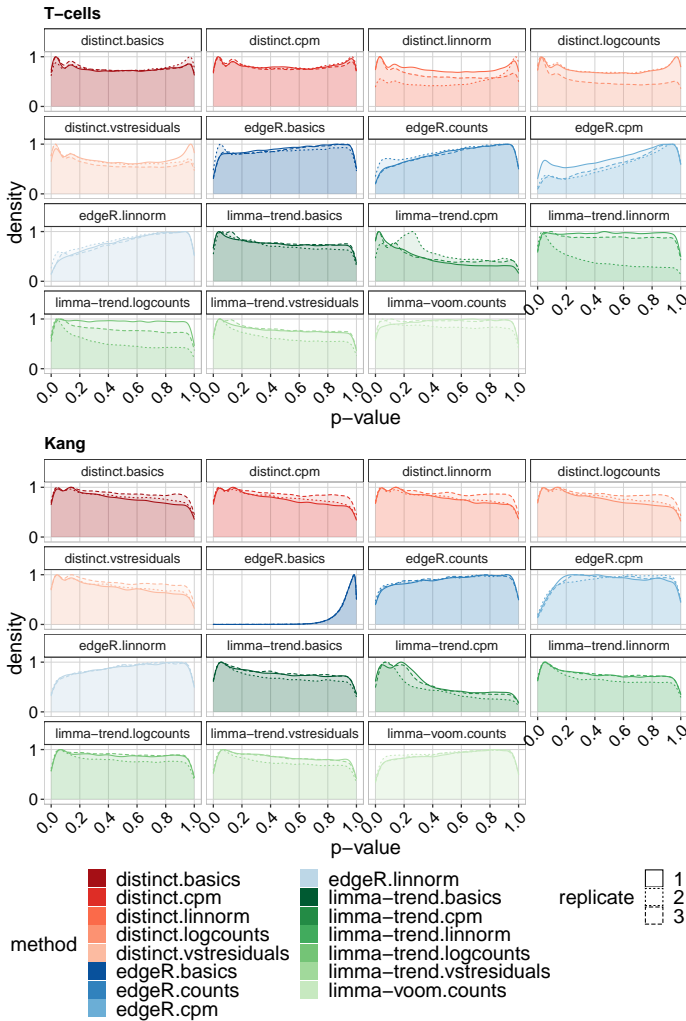
356 *et al.* [32]: the main DS dataset and two more where  
 357 differential effects were diluted by 50 and 75%. Each  
 358 dataset consists of 24 protein markers, 88,435 cells, and  
 359 two groups (with and without spike-in signal) of 8 sam-  
 360 ples each. Measurements were first transformed, and  
 361 then cells were grouped into sub-populations with two  
 362 separate approaches (see Methods): i) similarly to the  
 363 *muscat* simulation study, cell labels were defined based  
 364 on 8 manually annotated cell types [32] (Figure 7a),  
 365 and ii) as in the original *diffcyt* study from Weber  
 366 *et al.* [32], cells were grouped into 100 high-resolution clus-  
 367 ters (based on 10 cell-type markers, see Methods) via  
 368 unsupervised clustering (Figure 7b). In the main simu-

369 lation, *distinct* achieves higher TPR when considering  
 370 cell-type labels (Figure 7a, ‘main’), while all methods  
 371 exhibit substantially overlapping performance when using  
 372 unsupervised clustering (Figure 7b, ‘main’). In both  
 373 clustering approaches, as the magnitude of the differen-  
 374 tial effect decreases, the distance between methods in-  
 375 creases: *diffcyt* tools show a significant drop in the true  
 376 positive rate whereas *distinct* maintains a higher TPR  
 377 while effectively controlling for the false discovery rate  
 378 (FDR) (Figures 7a-b and Supplementary Figure 11).  
 379 This indicates that *distinct* has good statistical power  
 380 to detect even small changes between conditions. We  
 381 also considered the three replicate null datasets from  
 382 Weber *et al.* [32] (i.e., with no differential effect), con-  
 383 taining 24 protein markers and 88,438 cells across 8  
 384 cell types, and found that all methods display approx-  
 385 imately uniform p-values (Figure 7c).

## 386 Experimental data analyses

387 In order to investigate false positive rates (FPRs) in  
 388 real data, we considered two experimental scRNA-seq  
 389 datasets where no differential signals were expected, by  
 390 comparing samples from the same experimental condi-  
 391 tion. Given the high computational cost and low  
 392 power of MM, and the high FDR of *scDD* models, for  
 393 the real data analyses, we only included *distinct* and  
 394 PB methods. We considered gene-cluster combinations  
 395 with at least 20 non-zero cells across all samples. The  
 396 first dataset (labelled *T-cells*) consists of a Smart-seq2  
 397 scRNA-seq dataset of 19,875 genes and 11,138 T cells  
 398 isolated from peripheral blood from 12 colorectal can-  
 399 cer patients [36]. We automatically separated cells in  
 400 11 clusters (via *igraph* [1, 8]), and generated replicate  
 401 datasets, by randomly separating, three times, the 12  
 402 patients to two groups of size 6. The second dataset  
 403 (labelled *Kang*) contains 10x droplet-based scRNA-seq  
 404 peripheral blood mononuclear cell data from 8 Lupus  
 405 patients, before (controls) and after (stimulated) 6h-  
 406 treatment with interferon- $\beta$  (INF- $\beta$ ), a cytokine known  
 407 to alter the transcriptional profile of immune cells [13].  
 408 The full dataset contains 35,635 genes and 29,065 cells,  
 409 which are separated (via manual annotation [13]) into 8  
 410 cell types. One of the 8 patients was removed as it ap-  
 411 pears to be a potential outlier (Supplementary Figures  
 412 12-14). Here we only included singlet cells and cells  
 413 assigned to a cell population, and considered control  
 414 samples only, resulting in 11,854 cells and 10,891 genes.  
 415 Again, we artificially created three replicate datasets  
 416 by randomly assigning the 7 retained control samples  
 417 in two groups of size 3 and 4. In both null analyses, we  
 418 found that *limma-trend*, particularly when using CPMs,  
 419 leads to an increase of FPRs, *distinct*’s p-values are only

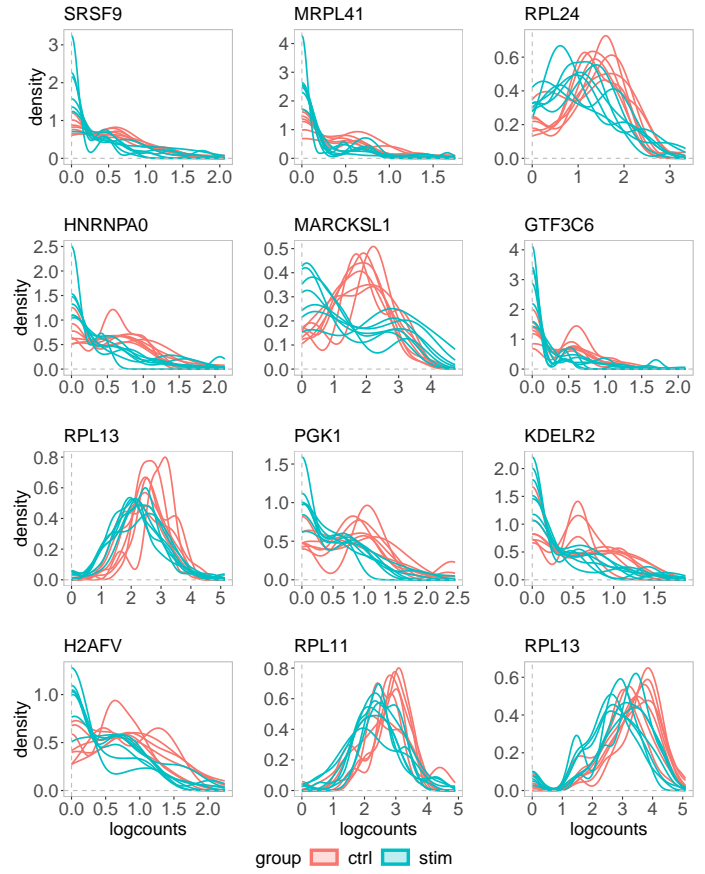




**Figure 8: On experimental scRNA-seq data, *distinct* has almost-uniform null p-values.** Density of raw p-values in the null *T-cells* (top) and *Kang* (bottom) experimental data. Each replicate represents a random partition of samples in two groups. The *T-cells* data consists of 12 samples and 11,138 cells across 11 clusters. For the *Kang* dataset, we retained 7 samples and 11,854 cells across 8 clusters.

marginally inflated towards 0, while *edgeR* and *limma-voom* are the most conservative methods and provide the best control of FPRs (Figure 8 and Supplementary Tables 2-3). Regarding normalization, *linnorm* and *BASiCS* lead to the most conservative p-values and smallest false positive rates.

We then considered again the *Kang* dataset, and performed a DS analysis between controls and stimulated samples. Again, we removed one potential outlier patient, and only considered singlet cells and cells assigned to a cell population; we further filtered gene-cluster combinations with less than 20 non-zero cells across all samples, resulting in 12,045 genes and 23,571 cells across 8 cell types and 14 samples. We found that *distinct* identifies more differential patterns than PB methods, with *edgeR* and *limma-voom* being the most conservative methods, and that its results are very coherent across different input data (Supplemen-



**Figure 9: *distinct* discovers non-canonical differential patterns.** Density of logcounts for nine examples of differential patterns identified by *distinct* on all input data (adjusted p-values < 0.05), and not by any PB tool (adjusted p-values > 0.05), on the *Kang* dataset when comparing controls and stimulated samples. Gene RPL13 was identified in FCGR3A+ Monocytes (third row) and in NK cells (fourth row), while all other genes were detected in Dendritic cells. Each line represents a sample.

tary Figure 15). When visually investigating the gene-cluster combinations detected by *distinct* (adjusted p-value < 0.1), on all five input data (CPMs, logcounts, linnorm, BASiCS and vstresiduals), and not detected by any of the ten PB approaches (adjusted p-value > 0.1), we found several interesting non-canonical differential patterns (Figure 9 and Supplementary Figures 16-27). In particular, gene MARCKSL1 displays a DB pattern, with stimulated samples having higher density on the tails and lower in the centre of the distribution, gene RPL13 mirrors classical DE, while the other genes seem to emulate DP profiles. Interestingly, ten out of eleven of these genes are known tumor prognostic markers: H2AZ2 for cervical and renal cancer, SRSF9 for liver cancer and melanoma, RPL24 for renal and thyroid cancer, HNRNPA0 for renal and pancreatic cancer, MARCKSL1 for liver and renal cancer, GTF3C6 for liver cancer, RPL13 for endometrial and renal cancer, PGK1 for breast, head and neck, cervical, liver, and pancreatic cancer, KDELR2 for renal, head

method	% of unique results
distinct.logcounts	0.3
distinct.basics	0.8
limma-trend.logcounts	0.9
distinct.cpm	1.0
distinct.vstresiduals	1.1
edgeR.linnorm	1.2
limma-trend.vstresiduals	1.5
limma-trend.basics	1.5
edgeR.counts	1.7
edgeR.basics	3.0
distinct.linnorm	3.6
limma-trend.linnorm	3.7
limma-voom.counts	5.6
edgeR.cpm	10.4
limma-trend.cpm	26.8

**Table 1:** Percentage of unique gene/cell-type identifications that are unique to each method. Since methods return significantly different number of significant results, for each method, we selected the most significant 1,000 results. For every method, we then compute the fraction of such results that are unique, i.e., not in common with the top 1,000 results returned by any other method.

and neck and glioma cancer, and RPL11 for renal and breast cancer [28]. This is an interesting association, considering that  $\text{INF-}\beta$  stimulation is known to inhibit and interfere with tumor progression [9, 22]. Additionally, Supplementary Figures 16-27 show how *distinct* can identify differences between groups of distributions even when only a portion of the ECDF varies between conditions. Finally, we computed the fraction of detected genes that are unique by each method. Given that a ground truth is absent, we speculate that gene-cluster combinations detected by multiple methods are more likely to be truly differential, while those detected by a single method are more likely to be false positive detections. Since methods return widely different number of significant genes, for each method, we considered the top (i.e., smallest p-value) 1,000 genes per cell-type. We then computed the percentage of results that are unique to each method (Table 1), i.e., not in common with the top 1,000 results returned by any other method. Overall, *distinct* displays a lower fraction of unique results (1.4% on average across all input data) compared to *edgeR* (4%) and *limma* (6.7%). It is also interesting to note that *scater*'s logcounts normalization lead to the 2 smallest fractions of unique values (i.e., *distinct.logcounts* and *limma-trend.logcounts*).

## Discussion

High-throughput single-cell data can display complex differential patterns; nonetheless, most methods for dif-

ferential expression fail to identify changes where the mean is not affected. To overcome the limitations of present differential tools, we have developed *distinct*, a novel method to identify differential patterns between groups of distributions, which is particularly well suited to perform differential analyses on high-throughput single-cell data. *distinct* is based on a flexible hierarchical multi-sample full-distribution non-parametric approach. In order to compare it to state-of-the-art differential methods, we ran extensive benchmarks on both simulated and experimental datasets from scRNA-seq and mass cytometry data, where our approach exhibits favourable performance, provides good control of the FNR and FDR, and is able to identify more patterns of differential expression compared to canonical tools, even when the overall mean is unchanged. In particular, our approach displays a higher statistical power (i.e., TPR) not only than PB methods, but also compared to other non-parametric frameworks from *scDD*, based on the Kolmogorov-Smirnov test statistic (*scDD-KS*) and on permutation tests (*scDD-perm*). *distinct* also allows for biological replicates, does not rely on asymptotic theory, which could be inaccurate in small sample sizes (typical of biological data), and avoids parametric assumptions, that may be challenging to meet in single-cell data. Additionally, *distinct* can also effectively adjust for sample-level cell-cluster specific covariates (i.e., whose effect varies across cell clusters), such as batch effects (Supplementary Figure 5). Importantly, *distinct* is a very general test that, due to its non-parametric nature, can be applied to various types of data, even beyond the single-cell applications shown here. Furthermore, thanks to its flexible form, we have shown in our simulations that *distinct* has the most consistent performance across normalization approaches (Supplementary Figure 2 and 4).

However, these advantages come at the expense of a higher computational burden, particularly when compared to PB methods or KS approaches (Figure 5). Nonetheless, by employing clever computational techniques (i.e., parallel computing and C++ coding within R), the method runs within minutes on a laptop, even for large datasets. Overall, we believe that *distinct* represents a valid alternative for differential detections from single-cell data, particularly when interest lies beyond canonical differences in means, as it allows to enhance statistical power at the cost of a reasonable increase in the computational time.

Finally, although we have focused here on comparing two groups of samples, several future extensions are possible to allow our framework to be applied to different scenarios. For instance, by suitably modifying

538 the test statistics in (1), one may ideally extend our ap-  
539 proach to perform a joint differential test between three  
540 of more groups of samples. Although, it is worth not-  
541 ing that, in the presence of three or more experimental  
542 conditions, at present, it is still possible to run pairwise  
543 comparisons between pairs of conditions. While a joint  
544 test across all groups may certainly be of interest in  
545 some cases, from our experience, comparisons between  
546 pairs of groups are usually more used among scientists.  
547 In addition, as we were suggested by a user, *distinct*  
548 could be employed to compare cell clusters instead of  
549 experimental conditions, hence discovering differential  
550 genes between cell clusters (e.g., cell types), even from  
551 individual samples.

## 552 Availability

553 *distinct* is freely available as a Bioconductor R pack-  
554 age at: <https://bioconductor.org/packages/distinct>.  
555 The scripts used to run all analyses are avail-  
556 able on GitHub ([https://github.com/SimoneTiberi/](https://github.com/SimoneTiberi/distinct_manuscript)  
557 *distinct\_manuscript*, version v3) and Zenodo (DOI:  
558 10.5281/zenodo.6397114). The *diffcyt* simulated data  
559 is available via FlowRepository (accession ID FR-FCM-  
560 ZYL8 [32]) and *HDCytoData* R Bioconductor pack-  
561 age [33]; the *Kang* dataset can be accessed via *musc-*  
562 *Data* R Bioconductor package [6]; the *T-cells* dataset  
563 is deposited on the European Genome-phenome (acces-  
564 sion id EGAD00001003910 [36]).

## 565 Acknowledgements

566 We acknowledge Almut Luetge, Brian D M Tom,  
567 Christina Azodi, Davis McCarthy, Reinhard Fur-  
568 rer, and the entire Robinson lab for precious com-  
569 ments and suggestions. This work was supported by  
570 Forschungskredit to ST (grant number FK-19-113) as  
571 well as by the Swiss National Science Foundation to  
572 MDR (grants 310030\_175841, CRSII5\_177208). MDR  
573 acknowledges support from the University Research  
574 Priority Program Evolution in Action at the Univer-  
575 sity of Zurich.

## 576 Author contributions

577 ST conceived the method, implemented it, performed  
578 all analyses and wrote the manuscript. ST and MDR  
579 designed the study. HLC and LMW contributed to  
580 *muscat* and *diffcyt* simulation studies, respectively. PS  
581 contributed to the computational development of *dis-*  
582 *tinct* and to the revision process. All authors read,  
583 contributed to, and approved the final article.

## 584 Competing interests

585 The authors declare no competing interests.

## 586 Methods

### 587 Permutation test

588 In order to test for differences between groups, we em-  
589 ploy a hierarchical permutation approach: to estimate  
590 the null distribution of  $s^{obs}$ , we permute the individual  
591 observations (e.g., single-cell measurements) instead of  
592 the samples. Note that this violates the exchangeability  
593 assumption of permutation tests and, hence, p-values  
594 are not guaranteed to be uniformly distributed under  
595 the null hypothesis; nonetheless, in our simulated and  
596 experimental analyses, we empirically show that *dis-*  
597 *tinct* provides good control of both false positive and  
598 false discovery rates. We randomly permute individual  
599 observations  $P$  times across all samples and groups, by  
600 retaining the original sample sizes. We denote by  $s_p$   
601 the test statistic computed at the  $p$ -th permutation,  
602  $p = 1, \dots, P$ . A p-value,  $\tilde{p}$ , is obtained as [21]:

$$\tilde{p} = \frac{\sum_{p=1}^P \mathbf{1}(s_p \geq s^{obs}) + 1}{P + 1}, \quad (2)$$

603 where  $\mathbf{1}(cond)$  is 1 if *cond* is true, and 0 otherwise. In  
604 order to accurately infer small p-values, when  $\tilde{p}$  is below  
605 some pre-defined thresholds, the number of permuta-  
606 tions are automatically increased and  $\tilde{p}$  is re-computed.  
607 By default, *distinct* initially computes 100 permuta-  
608 tions; when  $\tilde{p} \leq 0.1$  these are increased to 500; when  
609 the new  $\tilde{p} \leq 0.01$  we use 2,000 permutations, which  
610 are further increased to 10,000 if  $\tilde{p} \leq 0.001$ . Note that  
611 the number of permutations (i.e., 100, 500, 2,000 and  
612 10,000) can be specified by the user.

## 613 Covariates

Assume we observe  $Z$  nuisance covariates, and that  $N$   
samples are available across all groups, where for the  
 $i$ -th sample we observe  $C_i$  values (e.g., single-cell mea-  
surements). We fit the following linear mixed effects  
model:

$$y_c^{(i)} = \beta_0 + \sum_{z=1}^Z \beta_z X_z^{(i)} + \alpha_i + \epsilon_c^{(i)}, \text{ for } i = 1, \dots, N, \\ \text{and } c = 1, \dots, C_i, \quad (3)$$

614 where  $y_c^{(i)}$  represents the  $c$ -th observation for the  $i$ -th  
615 sample,  $\beta_0$  is the intercept of the model,  $X_z^{(i)}$  indicates  
616 the  $z$ -th covariate in the  $i$ -th sample,  $\beta_z$  denotes the

617 fixed effect coefficient for the  $z$ -th covariate,  $\alpha_i$  rep- 666  
618 resents the random effect term for the  $i$ -th sample, 667  
619 and  $\epsilon_c^{(i)}$  is the (zero-mean) residual for the  $c$ -th obser- 668  
620 vation in the  $i$ -th sample. We assume that random 669  
621 terms are normally distributed as  $\alpha_i \sim \mathcal{N}(0, \sigma_i^2)$ , where 670  
622  $\mathcal{N}(a, b)$  denotes the normal distribution with mean  $a$  671  
623 and variance  $b$ . Note that, due to the random effect 672  
624 terms, observations from the same sample are posi- 673  
625 tively correlated while, observations between different 674  
626 samples are independent. We infer model parameters 675  
627 via maximum likelihood, with the estimated values for 676  
628 the fixed effect terms denoted by  $\hat{\beta}_0, \dots, \hat{\beta}_Z$ . We then 677  
629 remove the estimated effect of nuisance covariates as 678  
630  $y_c^{(i)} - \sum_{z=1}^Z \hat{\beta}_z X_z^{(i)}$ ; differential testing is performed, as 679  
631 described above, on these normalized values. In DS 680  
632 analyses, model (3) is fit, separately, for every gene- 681  
633 cluster combination, hence accommodating for cell-type 682  
634 specific effects of covariates. 683

## 635 Normalization

636 In scRNA-seq datasets, CPMs and logcounts were com- 687  
637 puted via *scater* Bioconductor R package [19], *vstresiduals* 688  
638 were calculated via *sctransform* R package [12] 689  
639 (except for the *T-cells* data, where, due to a fail- 690  
640 ure of *sctransform*'s variance stabilizing normalization, 691  
641 we used *DESeq2*'s *vst* transformation [16]), while *lin-* 692  
642 *norm*, *BASiCS* and *SCnorm*, normalized data were 693  
643 calculated with the respective Bioconductor R pack- 694  
644 ages [3, 10, 29, 30, 34]. For *SCnorm*, following the 695  
645 author's suggestions, we normalized each cell cluster (3 in 696  
646 total) separately, using samples as *Conditions* param- 697  
647 eter. 698

648 In mass cytometry datasets, measurements were trans-  
649 formed via *diffcyt*'s *transformData* function, which ap-  
650 plies an *arcsinh* transformation.

## 651 *muscat* simulation and *Kang* data

652 In all *muscat* simulations, we used the control samples  
653 of the *Kang* dataset as a anchor data; as in the real  
654 data analyses, we excluded one sample as it emerged as  
655 a potential outlier (Supplementary Figures 12-14), and  
656 only considered singlet cells and cells assigned to a cell  
657 population. In *muscat*'s simulation studies, we con-  
658 sidered gene-cluster combinations with simulated ex-  
659 pression mean greater than 0.2; for DB patterns, we  
660 increased this threshold to 1 because with low expres-  
661 sion values differences are not visible by eye. In the  
662 simulation when varying the library size (Supplemen-  
663 tary Figure 6), we filtered gene-clusters combinations  
664 with at least 50 non-zero cells. For every simulations,  
665 five replicates were simulated, and results were aver-

aged across replicates. In the main simulation (Figure  
2) and the batch effect simulation (Supplementary Fig-  
ure 5), we simulated from a paired design 2 groups of  
3 samples each, with 4,000 genes, and 3,600 cells dis-  
tributed in 3 clusters (corresponding to an average of  
200 cells per sample in each cluster). For the simu-  
lation study when varying the number of cells (Fig-  
ure 4), the total numbers of available cells were 900,  
1,800, 3,600 and 7,200, corresponding to an average of  
50, 100, 200 and 400 cells per sample in every clus-  
ter. For the differential simulations, we used log2-FC  
values of 1 for DE, 1.5 for DP and DM, and 3 for DB  
and DV. For the batch effect simulation study we used a  
modified version of *muscat*, developed by Almut Luetge  
at the Robinson lab (available at: [https://github.com/  
SimoneTiberi/distinct\\_manuscript](https://github.com/SimoneTiberi/distinct_manuscript)), which allows sim-  
ulating cluster-specific batch effects [7, 17]. All *mus-*  
*cat* simulation studies, as well as the *Kang* non-null  
data analysis, were performed by editing the original  
snakemake workflow from Crowell *et al.* [7]. PB meth-  
ods were applied on aggregated data by summing cell-  
level measurements; for differential testing, we used  
*muscat*'s *pbDS* function [7]. Mixed model methods  
were implemented, via *muscat*'s *mmDS* function, us-  
ing the same approaches as in Crowell *et al.* [7]: in  
*MM-dream2* and *MM-vstresiduals* linear mixed models  
were applied to log-normalized data with observational  
weights and variance-stabilized data, respectively, while  
in *MM-nbinom* generalized linear mixed models were  
fitted directly to raw counts. In the *muscat* simulations  
and in the *Kang* non-null data analysis, we accounted  
for the paired design by modelling the patient id as a  
covariate in all methods that allow for covariates (i.e.,  
*distinct*, PB and MM).

## 700 *splatPop* simulation

701 In *SplatPOP* simulated data, we used a hu-  
702 man genome, version 19, downloaded from  
703 [https://www.encodegenes.org/human/release\\_19.html](https://www.encodegenes.org/human/release_19.html).  
704 We ran a total of 16 simulations: 8 with and 8 without  
705 batch effects as nuisance covariate. In each case, we  
706 ran 4 differential location ("de.facLoc" parameter)  
707 and 4 differential scale ("de.facScale" parameter)  
708 simulations, with differential parameters equals to  
709 0.2, 0.5, 1 and 1.5. In every simulation, 10% of  
710 genes were differential between groups, and a total  
711 of 20,345 genes and 800 cells were simulated (100  
712 per sample). In the simulation with batch effects,  
713 the 8 samples were randomly assigned to 2 batches,  
714 and the differential location and scale parameters  
715 between batches ("batch.facLoc" and "batch.facScale",  
716 respectively) matched those between groups of samples

717 (“de.facLoc” and “de.facScale”). For more details on  
 718 how *SplatPOP*’s data is simulated, please refer to the  
 719 original manuscript [2] and vignettes.

## 720 *diffcyt* simulation

721 The *diffcyt* semi-simulated data originates from a real  
 722 mass cytometry dataset of healthy peripheral blood  
 723 mononuclear cells from two paired groups of 8 samples  
 724 each [5]; one group contains unstimulated cells, while  
 725 the other was stimulated with B cell receptor/Fc recep-  
 726 tor cross-linker. The original dataset contains a total  
 727 of 172,791 cells and 24 protein markers: 10 of these  
 728 are cell-type markers used for cell clustering, while 14  
 729 are cell state markers used for differential state anal-  
 730 yses; the distinction between cell state and cell-type  
 731 markers is based on prior biological knowledge [32].  
 732 In Weber *et al.* [32], semi-simulated data were gener-  
 733 ated by separating the cells of each unstimulated sam-  
 734 ple in two artificial samples; a differential signal was  
 735 then computationally introduced by replacing, in one  
 736 group, unstimulated B cells with B cells from stimu-  
 737 lated samples. Measurements were transformed and  
 738 cells clustered via *diffcyt*’s *transformData* (which ap-  
 739 plies an *arcsinh* transformation) and *generateClusters*  
 740 functions, respectively. For the DS simulation in Fig-  
 741 ure 7b, as in Weber *et al.* [32], we evaluated methods’  
 742 performance in terms of detecting DS for phosphory-  
 743 lated ribosomal protein S6 (pS6) in B cells, which is  
 744 the strongest differential signal across the cell types in  
 745 this dataset [20, 32]. For the DS simulation in Figure  
 746 7a, we considered previously manually annotated cell  
 747 types [32] and included all 14 cell state markers. *dif-*  
 748 *fcyt*’s *limma* and LMM methods were applied via *dif-*  
 749 *fcyt*’s *testDS\_limma* and *testDS\_LMM* functions, re-  
 750 spectively [32]. We accounted for the paired design by  
 751 modelling the patient id as a covariate.

## 752 P-values adjustment

753 All p-values were adjusted via Benjamini-Hochberg cor-  
 754 rection [4]. In *diffcyt* simulations we used globally ad-  
 755 justed p-values for all methods, i.e., p-values from all  
 756 clusters are jointly adjusted once. However, since PB  
 757 methods were found to be over-conservative when glob-  
 758 ally adjusting p-values [7], in *muscat* simulations and  
 759 *Kang* discovery analyses, we used locally adjusted p-  
 760 values for all methods.

## 761 Software versions

762 All analyses were performed via R software version  
 763 4.0.0, with Bioconductor packages from release 3.11.

## References

- 765 [1] R. A. Amezquita, A. T. Lun, E. Becht, V. J. Carey, L. N. Carpp,  
 766 L. Geistlinger, F. Marini, K. Rue-Albrecht, D. Risso, C. Sonesson, et al.  
 767 Orchestrating single-cell analysis with bioconductor. *Nature methods*,  
 768 17(2):137–145, 2020.
- 769 [2] C. B. Azodi, L. Zappia, A. Oshlack, and D. J. McCarthy. *splatPop*: simu-  
 770 lating population scale single-cell RNA sequencing data. *Genome biology*,  
 771 22(1):1–16, 2021.
- 772 [3] R. Bacher, L.-F. Chu, N. Leng, A. P. Gasch, J. A. Thomson, R. M. Stewart,  
 773 M. Newton, and C. Kendziorski. Scnorm: robust normalization of single-cell  
 774 rna-seq data. *Nature methods*, 14(6):584–586, 2017.
- 775 [4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a prac-  
 776 tical and powerful approach to multiple testing. *Journal of the Royal sta-  
 777 tistical society: series B (Methodological)*, 57(1):289–300, 1995.
- 778 [5] B. Bodenmiller, E. R. Zunder, R. Finck, T. J. Chen, E. S. Savig, R. V.  
 779 Bruggner, E. F. Simonds, S. C. Bendall, K. Sachs, P. O. Krutzik, et al.  
 780 Multiplexed mass cytometry profiling of cellular states perturbed by small-  
 781 molecule regulators. *Nature biotechnology*, 30(9):858–867, 2012.
- 782 [6] H. L. Crowell. *muscData: Multi-sample multi-group scRNA-seq data*,  
 783 2020. R package version 1.1.2.
- 784 [7] H. L. Crowell, C. Sonesson, P.-L. Germain, D. Calini, L. Collin, C. Rap-  
 785 poso, D. Malhotra, and M. D. Robinson. muscat detects subpopulation-  
 786 specific state transitions from multi-sample multi-condition single-cell tran-  
 787 scriptomics data. *Nature Communications*, 11(1):1–12, 2020.
- 788 [8] G. Csardi and T. Nepusz. The igraph software package for complex network  
 789 research. *InterJournal, Complex Systems*:1695, 2006.
- 790 [9] M. R. Doherty, H. Cheon, D. J. Junk, S. Vinayak, V. Varadan, M. L. Telli,  
 791 J. M. Ford, G. R. Stark, and M. W. Jackson. Interferon-beta represses  
 792 cancer stem cell properties in triple-negative breast cancer. *Proceedings of  
 793 the National Academy of Sciences*, 114(52):13792–13797, 2017.
- 794 [10] N. Eling, A. C. Richard, S. Richardson, J. C. Marioni, and C. A. Vallejos.  
 795 Correcting the mean-variance dependency for differential variability testing  
 796 using single-cell RNA sequencing data. *Cell systems*, 7(3):284–294, 2018.
- 797 [11] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K.  
 798 Slichter, H. W. Miller, M. J. McElrath, M. P. Plic, et al. MAST: a flexible  
 799 statistical framework for assessing transcriptional changes and characterizing  
 800 heterogeneity in single-cell RNA sequencing data. *Genome biology*, 16(1):1–  
 801 13, 2015.
- 802 [12] C. Hafemeister and R. Satija. Normalization and variance stabilization  
 803 of single-cell RNA-seq data using regularized negative binomial regression.  
 804 *Genome biology*, 20(1):1–15, 2019.
- 805 [13] H. M. Kang, M. Subramaniam, S. Targ, M. Nguyen, L. Maliskova, E. Mc-  
 806 Carthy, E. Wan, S. Wong, L. Byrnes, C. M. Lanata, et al. Multiplexed  
 807 droplet single-cell RNA-sequencing using natural genetic variation. *Nature  
 808 biotechnology*, 36(1):89, 2018.
- 809 [14] P. V. Kharchenko, L. Silberstein, and D. T. Scadden. Bayesian approach to  
 810 single-cell differential expression analysis. *Nature methods*, 11(7):740–742,  
 811 2014.
- 812 [15] K. D. Korthauer, L.-F. Chu, M. A. Newton, Y. Li, J. Thomson, R. Stewart,  
 813 and C. Kendziorski. A statistical approach for identifying differential dis-  
 814 tributions in single-cell RNA-seq experiments. *Genome biology*, 17(1):222,  
 815 2016.
- 816 [16] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change  
 817 and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550,  
 818 2014.
- 819 [17] A. Lütge, J. Zypych-Walczak, U. B. Kunzmann, H. L. Crowell, D. Calini,  
 820 D. Malhotra, C. Sonesson, and M. D. Robinson. Cellmixs: quantifying and  
 821 visualizing batch effects in single-cell rna-seq data. *Life science alliance*,  
 822 4(6), 2021.
- 823 [18] F. J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal  
 824 of the American statistical Association*, 46(253):68–78, 1951.
- 825 [19] D. J. McCarthy, K. R. Campbell, A. T. Lun, and Q. F. Wills. Scater:  
 826 pre-processing, quality control, normalization and visualization of single-cell  
 827 RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186, 2017.
- 828 [20] M. Nowicka, C. Krieg, L. M. Weber, F. J. Hartmann, S. Guglietta,  
 829 B. Becher, M. P. Levesque, and M. D. Robinson. CyTOF workflow: dif-  
 830 ferential discovery in high-throughput high-dimensional cytometry datasets.  
 831 *F1000Research*, 6, 2017.
- 832 [21] B. Phipson and G. K. Smyth. Permutation p-values should never be zero:  
 833 Calculating exact p-values when permutations are randomly drawn. *Statisti-  
 834 cal applications in genetics and molecular biology*, 9:Article39, 2010.
- 835 [22] X.-Q. Qin, N. Tao, A. Dergay, P. Moy, S. Fawell, A. Davis, J. M. Wilson, and  
 836 J. Barsoum. Interferon- $\beta$  gene therapy inhibits tumor formation and causes  
 837 regression of established tumors in immune-deficient mice. *Proceedings of  
 838 the National Academy of Sciences*, 95(24):14411–14416, 1998.
- 839 [23] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K.  
 840 Smyth. limma powers differential expression analyses for RNA-sequencing  
 841 and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

- 842 [24] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor  
843 package for differential expression analysis of digital gene expression data.  
844 *Bioinformatics*, 26(1):139–140, 2010.
- 845 [25] C. Sonesson and M. D. Robinson. Bias, robustness and scalability in single-  
846 cell differential expression analysis. *Nature methods*, 15(4):255, 2018.
- 847 [26] J. W. Squair, M. Gautier, C. Kathe, M. A. Anderson, N. D. James, T. H.  
848 Hutson, R. Hudelle, T. Qaiser, K. J. Matson, Q. Barraud, et al. Confronting  
849 false discoveries in single-cell differential expression. *bioRxiv*, 2021.
- 850 [27] P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K.  
851 Pritchard, and Y. Gilad. Batch effects and the effective design of single-cell  
852 gene expression studies. *Scientific reports*, 7:39921, 2017.
- 853 [28] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold,  
854 A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, et al.  
855 Tissue-based map of the human proteome. *Science*, 347(6220), 2015.
- 856 [29] C. A. Vallejos, J. C. Marioni, and S. Richardson. BASiCS: Bayesian  
857 analysis of single-cell sequencing data. *PLoS computational biology*,  
858 11(6):e1004333, 2015.
- 859 [30] C. A. Vallejos, S. Richardson, and J. C. Marioni. Beyond comparisons of  
860 means: understanding changes in gene expression at the single-cell level.  
861 *Genome biology*, 17(1):1–14, 2016.
- 862 [31] T. Wang, B. Li, C. E. Nelson, and S. Nabavi. Comparative analysis of dif-  
863 ferential gene expression analysis tools for single-cell RNA sequencing data.  
864 *BMC bioinformatics*, 20(1):40, 2019.
- 865 [32] L. M. Weber, M. Nowicka, C. Sonesson, and M. D. Robinson. diffcyt: Differ-  
866 ential discovery in high-dimensional cytometry via high-resolution cluster-  
867 ing. *Communications biology*, 2(1):1–11, 2019.
- 868 [33] L. M. Weber and C. Sonesson. Hdcytodata: Collection of high-  
869 dimensional cytometry benchmark datasets in bioconductor object formats.  
870 *F1000Research*, 8, 2019.
- 871 [34] S. H. Yip, P. Wang, J.-P. A. Kocher, P. C. Sham, and J. Wang. Linnorm: im-  
872 proved statistical analysis for single cell RNA-seq expression data. *Nucleic  
873 acids research*, 45(22):e179–e179, 2017.
- 874 [35] L. Zappia, B. Phipson, and A. Oshlack. Splatter: simulation of single-cell  
875 RNA sequencing data. *Genome biology*, 18(1):1–15, 2017.
- 876 [36] Y. Zhang, L. Zheng, L. Zhang, X. Hu, X. Ren, and Z. Zhang. Deep single-cell  
877 RNA sequencing data of individual T cells from treatment-naive colorectal  
878 cancer patients. *Scientific data*, 6(1):1–15, 2019.