distinct: a novel approach to differential distribution analyses

Simone Tiberi[1*], Helena L Crowell[1], Lukas M Weber[2], Pantelis Samartsidis[3] and Mark D Robinson[1]

[1]*Department of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland.*
[2]*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.*
[3]*MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK.*
[*] e-mail: Simone.Tiberi@uzh.ch

Abstract

We present *distinct*, a general method for differential analysis of full distributions that is well suited to applications on single-cell data, such as single-cell RNA sequencing and high-dimensional flow or mass cytometry data. High-throughput single-cell data reveal an unprecedented view of cell identity and allow complex variations between conditions to be discovered; nonetheless, most methods for differential expression target differences in the mean and struggle to identify changes where the mean is only marginally affected. *distinct* is based on a hierarchical non-parametric permutation approach and, by comparing empirical cumulative distribution functions, identifies both differential patterns involving changes in the mean, as well as more subtle variations that do not involve the mean. We performed extensive benchmarks across both simulated and experimental datasets from single-cell RNA sequencing and mass cytometry data, where *distinct* shows favourable performance, identifies more differential patterns than competitors, and displays good control of false positive and false discovery rates. *distinct* is available as a Bioconductor R package.

keywords: Differential distribution; Differential analyses; Differential state; High-throughput single-cell data; Single-cell RNA-seq; Single-cell flow and mass cytometry; Permutation tests.

Background

Technology developments in the last decade have led to an explosion of high-throughput single-cell data, such as single-cell RNA sequencing (scRNA-seq) and high-dimensional flow or mass cytometry data, allowing researchers to investigate biological mechanisms at single-cell resolution. Single-cell data have also extended the canonical definition of differential expression by displaying cell-type specific responses across conditions, known as differential state (DS) [28], where genes or proteins vary in specific sub-populations of cells (e.g., a cytokine response in myeloid cells but not in other leukocytes [10]). Classical bulk differential expression methods have been shown to perform well when used on single-cell measurements [22, 23, 27] and on aggregated data (i.e., averages or sums across cells), also referred to as pseudo-bulk (PB) [5, 28]. However, most bulk and PB tools focus on shifts in the means, and may conceal information about cell-to-cell heterogeneity. Indeed, single-cell data can show more complex variations (Figure 1 and Supplementary Figure 1); such patterns can arise due to increased stochasticity and heterogeneity, for example owing to oscillatory and unsynchronized gene expression between cells, or when some cells respond differently to a treatment than others [12, 27]. In addition to bulk and PB tools, other methods were specifically proposed to perform differential analyses on single-cell data (notably: *scDD* [12], *SCDE* [11], *MAST* [8], *BASiCS* [26] and mixed models [24]). Nevertheless, they all present significant limitations: BASiCS does not perform cell-type specific differential testing between conditions, scDD does not directly handle covariates and biological replicates, while PB, SCDE, MAST and mixed models performed poorly in previous benchmarks when detecting differential patterns that do not involve the mean [5, 12].

Results

distinct's full distribution approach

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

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

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

More formally, assume we are interested in comparing two groups, that we call A and B, for which N_A and N_B samples are available, respectively. The ECDF for the i-th sample in the j-th group, is denoted by $ecdf_i^{(j)}(.)$, for $j \in \{A, B\}$ and $i = 1, \dots, N_j$. We then define K equally spaced cut-offs between the minimum, min, and maximum, max, values observed across all samples: $b_1, \dots, b_K$, where $b_k = min + k \times l$, for $k = 1, \dots, K$, with $l = (max - min)/(K + 1)$ being the distance between two consecutive cut-offs. We exclude min and max from the cut-offs because, trivially, $ecdf_i^{(j)}(min) = 0$ and $ecdf_i^{(j)}(max) = 1$, $\forall j, i$. At every cut-off, we compute the absolute difference between the mean ECDF in the two groups; our test statistic, s^{obs}, is obtained by adding these differences across all cut-offs:

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

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

Intuitively, s^{obs}, which ranges in $[0, \infty)$, approximates the area between the average ECDFs, and represents a measure of distance between two groups of densities: the bigger s^{obs}, the greater the distance between groups. The number of cut-offs K, which can be defined by users, is set to 25 by default, because no detectable difference in performance was observed when further increasing it (data not shown). Note that, although at each cut-off we compute the average across each group's curves, ECDFs are computed separately for each individual sample, therefore our approach still accounts for the within-group variability; indeed, at a given threshold, the average of the sample-specific ECDFs differs from the group-level ECDF (i.e., the curve based on all individual measurements from the group). The null distribution of s^{obs} is then estimated via a hierarchical non-parametric permutation approach (see Methods). A major disadvantage of permutation tests, which often restricts its usage on biological data, is that too few permutations are available from small samples. We overcome this by permuting cells, which is still possible in small samples, because there are many more cells than samples. In principle, this may lead to an inflation of false positives due to lack of exchangability (see Methods); nonetheless, in our analyses, *distinct* provides good control of both false positive and false discovery rates.

Importantly, *distinct* is general and flexible: it targets complex changes between groups, explicitly models biological replicates within a hierarchical framework, does not rely on asymptotic theory, avoids parametric assumptions, and can be applied to arbitrary types of data. Additionally, *distinct* can also adjust for sample-level cell-cluster specific covariates (i.e., whose effect varies across cell clusters), such as batch effects,: *distinct* fits a linear model with the input data (e.g., CPMs or log2-CPMs) as response variable, and the covariates as predictors; the method then removes the estimated effect of covariates, and performs differential testing on these normalized values (see Methods).

Furthermore, to enhance the interpretability of differential results, *distinct* provides functionalities to compute (log) fold changes between conditions, and to plot densities and ECDFs, both for individual samples and at the group-level.

Note that, although *distinct* and the Kolmogorov-Smirnov [15] (KS) test share similarities (they both compare distributions via non-parametric tests), the two approaches present several conceptual differences. Firstly, the KS considers the maximum distance between two ECDFs, while our approach estimates the overall distance between ECDFs, which in our view is a more appropriate way to measure the difference between distributions. Secondly, the KS test only compares two individual densities, while our framework compares groups of distributions. Thirdly, while the KS statistic relies on asymptotic theory, our framework uses a permutation test. Finally, a comparison between *distinct* and *scDD* [12] based on the KS test (labelled *scDD-KS*) shows that our method, compared to the KS test, has greater statistical power to detect differential effects and leads to fewer false discoveries (see Simulation studies).

Simulation studies

We conducted an extensive benchmark, based on scRNA-seq and mass cytometry simulated and experi-

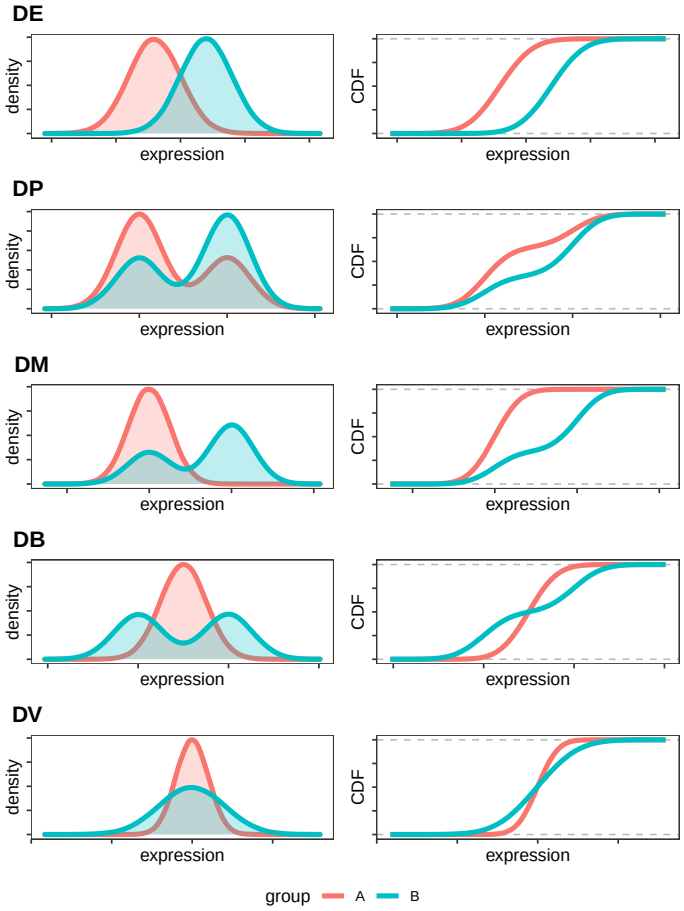

Figure 1: Cumulative distribution functions (CDFs) unravel differences between distributions. Density (left panels) and CDF (right panels) of five differential patterns: differential variability (DV), and the four proposed by Korthauer et. al. [12]: differential expression (DE), differential proportion (DP), differential modality (DM), and both differential modality and different component means (DB).

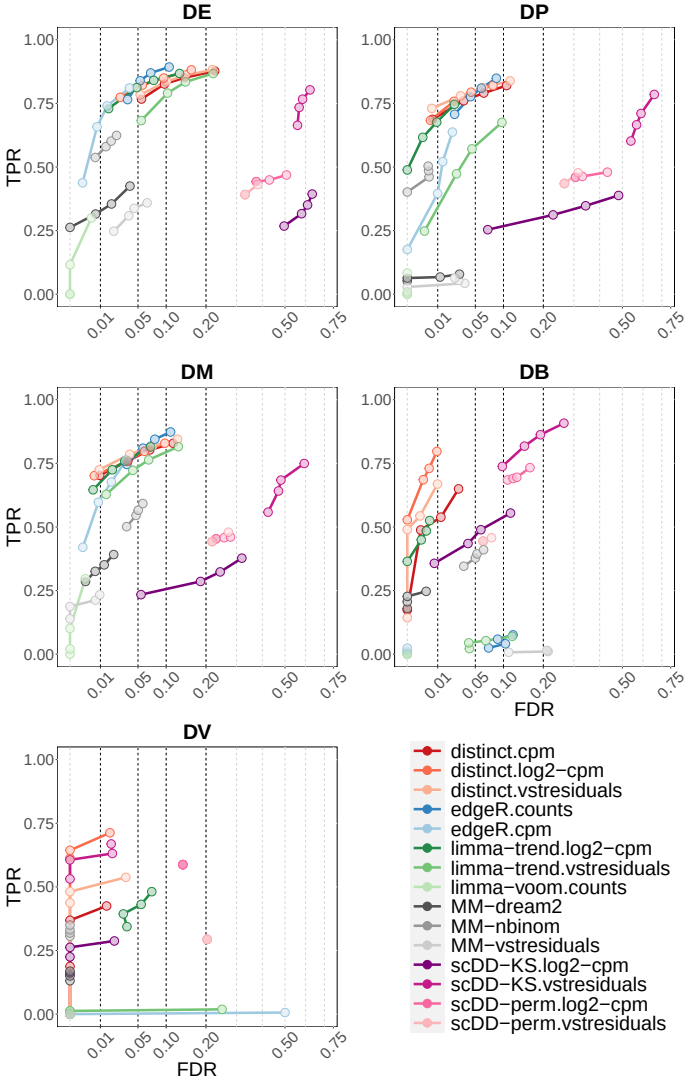

Figure 2: *distinct* identifies various differential patterns and controls for the FDR. TPR *vs.* FDR in *muscat* simulated data; DE, DP, DM, DB and DV refer to the differential profiles illustrated in Figure 1. Results are averages across the five simulation replicates. Circles indicate observed FDR for 0.01, 0.05, 0.1 and 0.2 significance thresholds. Two groups of 3 samples are compared and, on average, 200 cells are available for every sample in each of three clusters.

mental datasets to investigate *distinct*'s ability to identify differential patterns in sub-populations of cells.

First, we simulated droplet scRNA-seq data via *muscat* [5] (see Methods). We ran five simulation replicates for each of the differential profiles in Figure 1, with 10% of the genes being differential in each cluster, where DE (differential expression) indicates a shift in the entire distribution, DP (differential proportion) implies two mixture distributions with different proportions of the two components, DM (differential modality) assumes a unimodal and a bimodal distribution, DB (both differential modality and different component means) compares a unimodal and a bimodal distribution with the same overall mean, and DV (differential variability) refers to two unimodal distributions with the same mean but different variance (Figure 1 and Supplementary Figure 1). Each individual simulation 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.

We considered three different normalizations: counts per million (CPMs), logarithm of CPMs to base 2 (log2-CPMs) and residuals from variance stabilizing normalization from *sctransform* (vstresiduals) [9]. We compared *distinct* to several PB approaches from *muscat*, based on *edgeR* [21], *limma-voom* and *limma-trend* [20], which emerged among the best performing methods for differential analyses from scRNA-seq data [5, 23]. We further considered three methods from *muscat* based on mixed models (MM), namely *MM-dream2*, *MM-vstresiduals* and *MM-nbinom* (see Methods). Finally, we included *scDD* [12], which is conceptually similar to our approach: *scDD* implements a non-parametric method to detect changes between individual distributions from scRNA-seq, based on the Kolmogorov-Smirnov test, *scDD-KS*, and on a permutation ap-

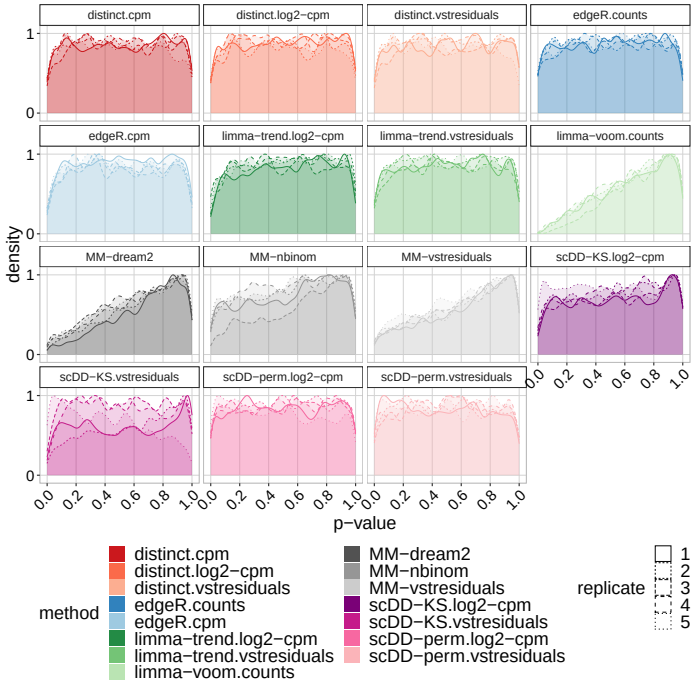

Figure 3: *distinct* has uniform null p-values. Density of raw p-values in *muscat* null simulated data; each replicate represents a different null simulation. Two groups of 3 samples are compared and, on average, 200 cells are available for every sample in each of three clusters.

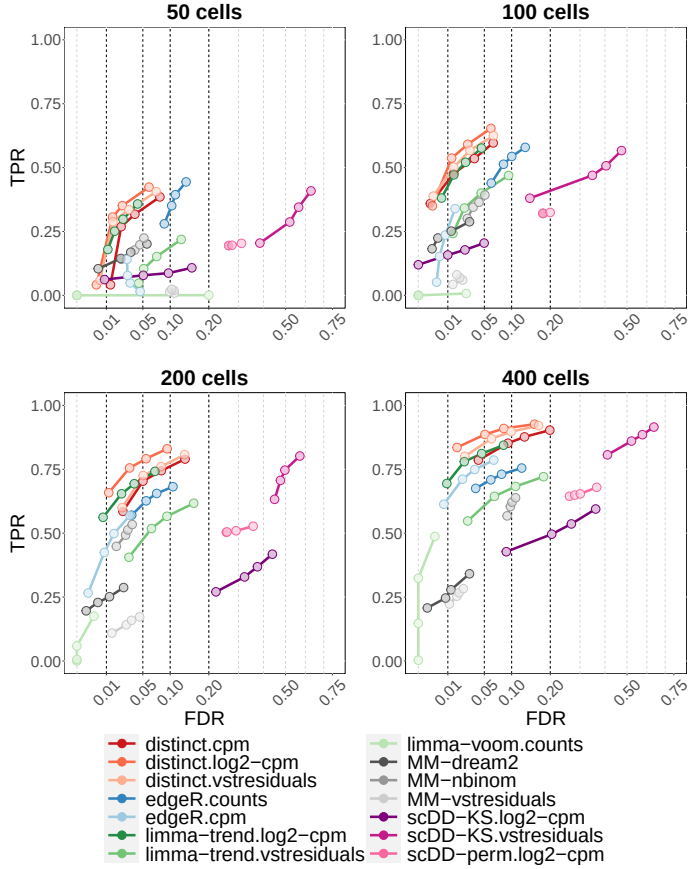

Figure 4: *distinct* achieves good performance when varying the number of available cells. TPR *vs.* FDR in *muscat* simulated data; with 50, 100, 200 and 400 cells per cluster-sample combination, corresponding to a total of 900, 1,800, 3,600 and 7,200 cells, respectively. Results are aggregated over 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.vstresiduals* was excluded from this analysis due to its computational cost.

proach, *scDD-perm*. For *scDD-perm* we used 100 permutations to reduce the computational burden.

In all scenarios and on all three input datasets, *distinct* shows favourable performance: it has good statistical power while controlling for the false discovery rate (FDR) (Figure 2). In particular, for DE, DP and DM, *distinct* has similar performance to the best performing competitors (*edgeR.counts* and *limma-trend.log2-CPMs*), while for DB and DV, it achieves significantly higher true positive rate (TPR), especially when using *log2-CPMs*. PB methods in general perform well for differential patterns involving changes in the mean (DE, DP and DM), but struggle to identify DB and DV patterns. *scDD* provides good TPR across all patterns when using the KS test on vstresiduals (*scDD-KS.vstresiduals*), while the TPR is significantly reduced when using *log2-CPMs* and with the permutation approach(*scDD-perm*); however, *scDD* methods also show a significant inflation of the FDR. In contrast, MM methods provide good control of the FDR but have low statistical power in all differential scenarios.

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, no method presents an inflation of false positives, with *distinct*, *edgeR*, *limma-trend* and *scDD* showing 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) [5,14]. 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 are substantially unchanged (Supplementary Figure 2), which shows *distinct* can effectively remove nuisance confounders. Furthermore, by varying the number of cells in the simulated data, we show that, compared to PB, MM and *scDD* methods, *distinct* achieves higher overall TPR, while controlling for the FDR, regardless of the number of available cells (Figure 5 and Supplementary Figure 3).

From a computational perspective, *distinct* required

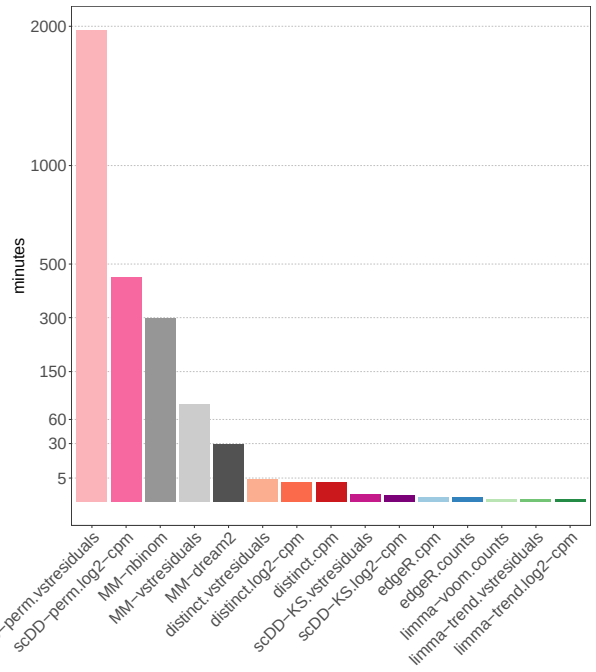

Figure 5: *distinct* requires more computational resources than PB and *scDD-KS* methods, but significantly less than MM and *scDD-perm* models. Average computing time, expressed in minutes, in *muscat* main simulations (Figures 2-3). For each method, times are averaged across simulation types (DE, DP, 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 parellel computing.

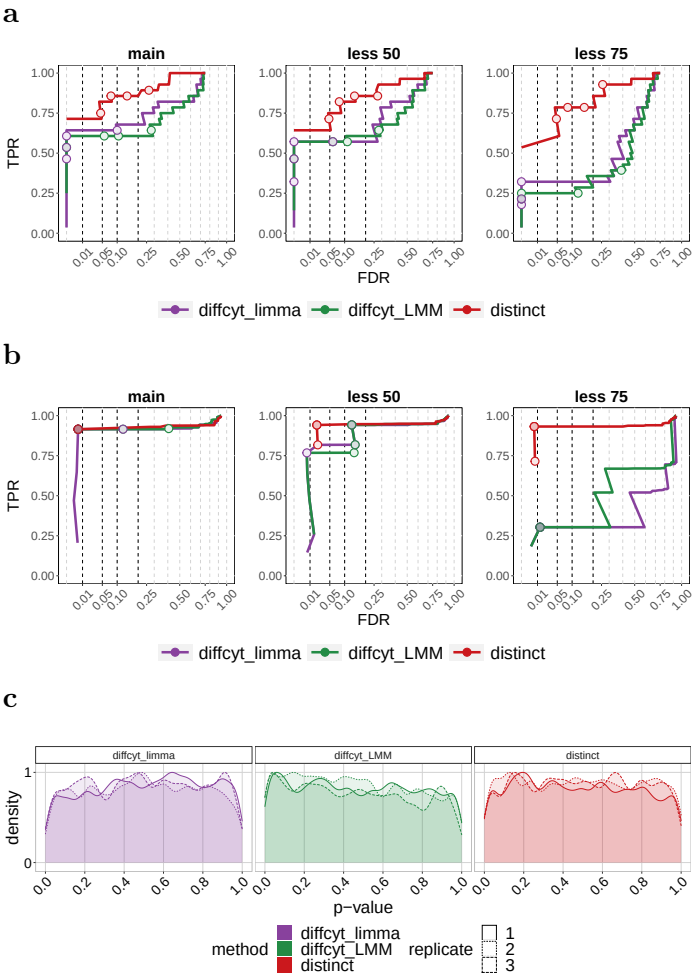

Figure 6: *distinct* shows high power while controlling for false positive and false discovery rates. (a-b) TPR *vs.* FDR in *diffcyt* semi-simulated data. 'main', 'less 50' and 'less 75' indicate the main simulation, and those where differential effects are diluted by 50 and 75%, respectively. Each 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 [28]. **(b)** As in Weber *et al.* [28], 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.* [28], cells were clustered in an unsupervised manner.

an average time of 3.4 to 4.5 minutes per simulation, which is higher than PB methods (0.1 to 0.2 minutes) and *scDD-KS* (0.4 to 0.5 minutes), but significantly lower than MM approaches (29.4 to 297.3 minutes) and *scDD-perm* (447.5 to 1970.1 minutes) (Figure 4 and Supplementary Table 1). All methods were run on 3 cores, except PB approaches, which used a single core, because they do not allow for parellel computing.

We further considered the semi-simulated mass cytometry data from Weber *et al.* [28] (labelled *diffcyt* simulation), where spike-in signals were computationally introduced in experimental data [3], hence maintaining the properties of real biological data while also embedding a known ground truth signal. We evaluated *distinct* and two methods from *diffcyt*, based on *limma* [20] and linear mixed models (LMM), which outperformed competitors on these same data [28]. In particular, we considered three datasets from Weber *et al.* [28]: the main DS dataset and two more where differential effects were diluted by 50 and 75%. Each dataset consists of 24 protein markers, 88,435 cells, and two groups (with and without spike-in signal) of 8 samples each. Measurements were first transformed, and then cells were grouped into sub-populations with two separate approaches (see Methods): i) similarly to the *muscat* simulation study, cell labels were defined based on 8 manually annotated cell types [28] (Figure 6a), and ii) as in the original *diffcyt* study from Weber *et al.* [28], cells were grouped into 100 high-resolution clusters (based on 10 cell-type markers, see Methods) via unsupervised clustering (Figure 6b). In the main simulation, *distinct* achieves higher TPR when considering cell-type labels (Figure 6a, 'main'), while all methods exhibit substantially overlapping performance when using unsupervised clustering (Figure 6b, 'main'). In both clustering approaches, as the magnitude of the differential effect decreases, the distance between methods

increases: *diffcyt* tools show a significant drop in the true positive rate (TPR) whereas *distinct* maintains a higher TPR while effectively controlling for the false discovery rate (FDR) (Figures 6a-b and Supplementary Figure 4). This indicates that *distinct* has good statistical power to detect even small changes between conditions. We also considered the three replicate null datasets from Weber *et al.* [28] (i.e., with no differential effect), containing 24 protein markers and 88,438 cells across 8 cell types, and found that all methods display approximately uniform p-values (Figure 6c).

Experimental data analyses

In order to investigate false positive rates (FPRs) in real data, we considered two experimental scRNA-seq datasets where no differential signals were expected, by comparing samples from the same experimental condition. Given the high computational cost and low power of MM, and the high FDR of *scDD* models, for the real data analyses, we only included *distinct* and PB methods. We considered gene-cluster combinations with at least 20 non-zero cells across all samples. The first dataset (labelled *T-cells*) consists of a Smart-seq2 scRNA-seq dataset of 23,459 genes and 11,138 T cells isolated from peripheral blood from 12 colorectal cancer patients [30]. We automatically separated cells in 11 clusters (via *igraph* [1, 6]), and generated replicate datasets, by randomly separating, three times, the 12 patients to two groups of size 6. The second dataset (labelled *Kang*) contains 10x droplet-based scRNA-seq peripheral blood mononuclear cell data from 8 Lupus patients, before (controls) and after (stimulated) 6h-treatment with interferon-β (INF-β), a cytokine known to alter the transcriptional profile of immune cells [10]. The full dataset contains 35,635 genes and 29,065 cells, which are separated (via manual annotation [10]) into 8 cell types. One of the 8 patients was removed as it appears to be a potential outlier (Supplementary Figures 5-7). Here we only included singlet cells and cells assigned to a cell population, and considered control samples only, resulting in 11,854 cells. Again, we artificially created three replicate datasets by randomly assigning the 7 retained control samples in two groups of size 3 and 4. In both null analyses, we found that *limma-trend* leads to a major increase of FPRs, *distinct*'s p-values are only marginally inflated towards 0, while *edgeR* and *limma-voom* are the most conservative methods and provide the best control of FPRs (Figure 7a and Supplementary Tables 2-3).

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, resulting in 35,635 genes,

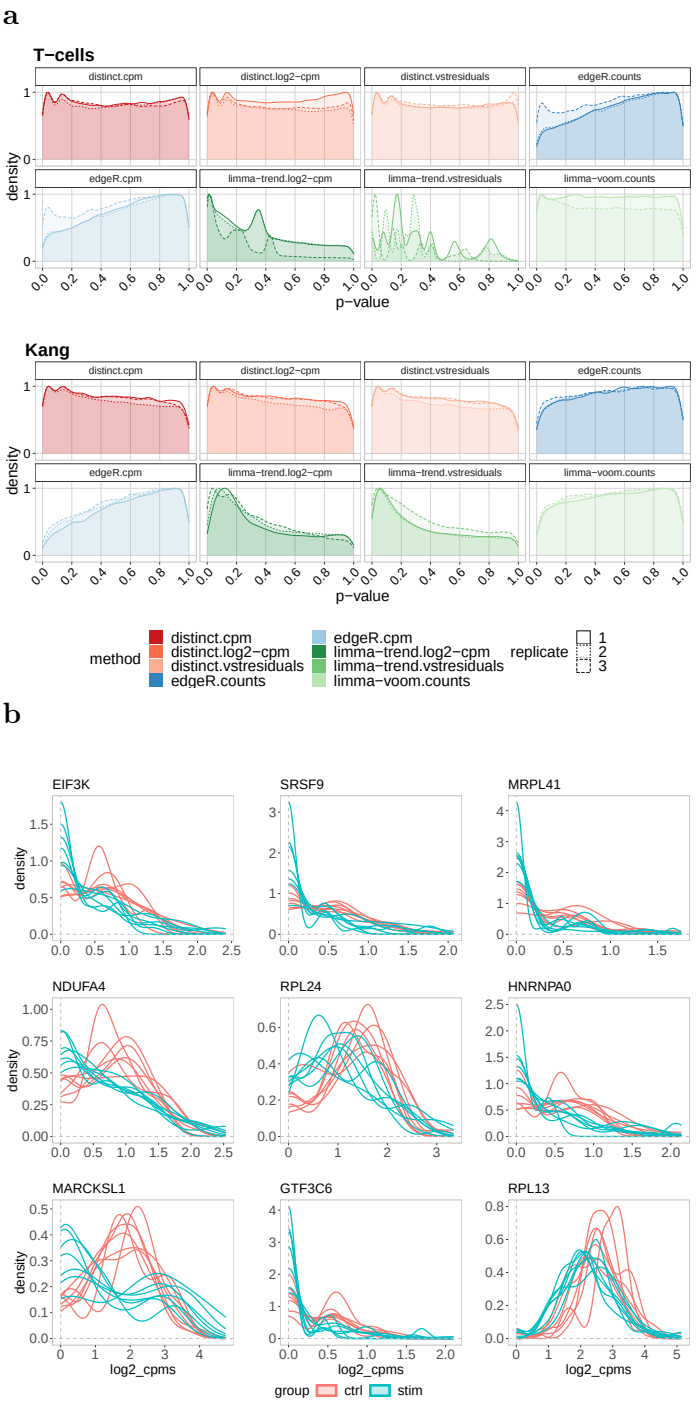

Figure 7: On experimental scRNA-seq data, *distinct* discovers non-canonical differential patterns, and has almost-uniform null p-values. (a) Density of raw p-values in the null *T-cells* (top) and *Kang* (bottom) experimental data. Each replicate represents a random partition of samples in two groups. 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. **(b)** Density of log2-CPMs 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 cells, while all other genes were detected in Dendritic cells. Each line represents a sample.

23,571 cells across 8 cell types and 14 samples; we further filtered gene-cluster combinations with less than 20 non-zero cells across all 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 (Supplementary Figure 8). When visually investigating the gene-cluster combinations detected by *distinct* (adjusted p-value < 0.05), on all input data (CPMs, log2-CPMs and vstresiduals), and not detected by any PB method (adjusted p-value > 0.05), we found several interesting non-canonical differential patterns (Figure 7b and Supplementary Figures 9-17). 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, eight out of nine of these genes are known tumor prognostic markers: EIF3K for cervical and renal cancer, SRSF9 for liver cancer and melanoma, NDUFA4 for renal cancer, RPL24 for renal and thyroid cancer, HNRNPA0 for renal and pancreatic cancer, MARCKSL1 for liver and renal cancer, GTF3C6 for liver cancer and RPL13 for endometrial and renal cancer [25]. This is an interesting association, considering that INF-β stimulation is known to inhibit and interfere with tumor progression [7, 19]. Finally, Supplementary Figures 9-17 show how *distinct* can identify differences between groups of distributions even when only a portion of the ECDF varies between conditions.

Discussion

High-throughput single-cell data can display complex differential patterns; nonetheless, most methods for differential expression fail to identify changes where the mean is not affected. To overcome present limitations, we have introduced *distinct*, a general method to identify differential patterns between groups of distributions, which is particularly well suited to perform differential analyses on high-throughput single-cell data. We ran extensive benchmarks on both simulated and experimental datasets from scRNA-seq and mass cytometry data, where our method exhibits favourable performance, provides good control of the FPR and FDR, and is able to identify more patterns of differential expression compared to canonical tools, even when the overall mean is unchanged. Furthermore, *distinct* allows for biological replicates, can adjust for covariates (e.g., batch effects), and does not rely on asymptotic theory. Finally, note that *distinct* is a very general test that, due to its non-parametric nature, can be applied to various types of data, beyond the single-cell applications shown here.

Availability

distinct is freely available as a Bioconductor R package at: https://bioconductor.org/packages/distinct. The scripts used to run all analyses are available on GitHub (https://github.com/SimoneTiberi/distinct_manuscript, version v2) and Zenodo (DOI: 10.5281/zenodo.4739098). The *diffcyt* simulated data is available via FlowRepository (accession ID FR-FCM-ZYL8 [28]) and *HDCytoData* R Bioconductor package [29]; the *Kang* dataset can be accessed via *muscData* R Bioconductor package [4]; the *T-cells* dataset is deposited on the European Genome-phenome (accession id EGAD00001003910 [30]).

Acknowledgements

We acknowledge Almut Luetge and the entire Robinson lab for precious comments and suggestions. This work was supported by Forschungskredit to ST (grant number FK-19-113) as well as by the Swiss National Science Foundation to MDR (grants 310030_175841, CRSII5_177208). MDR acknowledges support from the University Research Priority Program Evolution in Action at the University of Zurich.

Author contributions

ST conceived the method, implemented it, performed all analyses and wrote the manuscript. ST and MDR designed the study. HLC and LMW contributed to *muscat* and *diffcyt* simulation studies, respectively. PS contributed to the computational development of *distinct*. All authors read, contributed to, and approved the final article.

Competing interests

The authors declare no competing interests.

Methods

Permutation test

In order to test for differences between groups, we employ a hierarchical permutation approach: to estimate the null distribution of s^{obs}, we permute the individual observations (e.g., single-cell measurements) instead of the samples. Note that this violates the exchangeability assumption of permutation tests and, hence, p-values are not guaranteed to be uniformly distributed under

the null hypothesis; nonetheless, in our simulated and experimental analyses, we empirically show that *distinct* provides good control of both false positive and false discovery rates. We randomly permute individual observations P times across all samples and groups, by retaining the original sample sizes. We denote by s_p the test statistic computed at the p-th permutation, $p = 1, \ldots, P$. A p-value, $\tilde{p}$, is obtained as [18]:

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

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

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 measurements). We fit the following linear model:

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

where $y_c^{(i)}$ represents the c-th observation for the i-th sample, β_0 is the intercept of the model, $X_z^{(i)}$ indicates the z-th covariate in the i-th sample, β_z represents the coefficient for the z-th covariate, and $\epsilon_c^{(i)}$ is the residual for the c-th observation in the i-th sample. We infer parameters $\beta_0, \ldots, \beta_Z$ via least squares regression, with the estimated values denoted by $\hat{\beta}_0, \ldots, \hat{\beta}_Z$. We then remove the estimated effect of covariates as $y_c^{(i)} - \sum_{z=1}^{Z} \hat{\beta}_z X_z^{(i)}$; differential testing is performed, as described above, on these normalized values. For DS analyses, model (3) is fit, separately, for every gene-cluster combination, hence accommodating for cell-type specific effects of covariates.

Normalization

In scRNA-seq datasets, CPMs and log2-CPMs were computed via *scater* Bioconductor R package [16], while vstresiduals were calculated via *sctransform* R package [9] (except for the *T-cells* data, where, due to a failure of *sctransform*'s variance stabilizing normalization, we used *DESeq2*'s vst transformation [13]).

In mass cytometry datasets, measurements were transformed via *diffcyt*'s *transformData* function, which applies an *arcsinh* transformation.

diffcyt simulation

The *diffcyt* semi-simulated data originates from a real mass cytometry dataset of healthy peripheral blood mononuclear cells from two paired groups of 8 samples each [3]; one group contains unstimulated cells, while the other was stimulated with B cell receptor/Fc receptor cross-linker. The original dataset contains a total of 172,791 cells and 24 protein markers: 10 of these are cell-type markers used for cell clustering, while 14 are cell state markers used for differential state analyses; the distinction between cell state and cell-type markers is based on prior biological knowledge [28]. In Weber *et al.* [28], semi-simulated data were generated by separating the cells of each unstimulated sample in two artificial samples; a differential signal was then computationally introduced by replacing, in one group, unstimulated B cells with B cells from stimulated samples. Measurements were transformed and cells clustered via *diffcyt*'s *transformData* (which applies an *arcsinh* transformation) and *generateClusters* functions, respectively. For the DS simulation in Figure 6b, as in Weber *et al.* [28], we evaluated methods' performance in terms of detecting DS for phosphorylated ribosomal protein S6 (pS6) in B cells, which is the strongest differential signal across the cell types in this dataset [17, 28]. For the DS simulation in Figure 6a, we considered previously manually annotated cell types [28] and included all 14 cell state markers. *diffcyt*'s *limma* and LMM methods were applied via *diffcyt*'s *testDS_limma* and *testDS_LMM* functions, respectively [28]. We accounted for the paired design by modelling the patient id as a covariate.

muscat simulation and *Kang* data

In all *muscat* simulations, we used the control samples of the *Kang* dataset as a anchor data; as in the real data analyses, we excluded one sample as it emerged as a potential outlier (Supplementary Figures 5-7), and only considered singlet cells and cells assigned to a cell population. In *muscat*'s simulation studies, we considered gene-cluster combinations with simulated expression mean greater than 0.2; for DB patterns, we increased this threshold to 1 because with low expression values differences are not visible by eye. For every

simulations, five replicates were simulated, and results were averaged across replicates. In the main simulation (Figure 2) and the batch effect simulation (Supplementary Figure 3), we simulated from a paired design 2 groups of 3 samples each, with 4,000 genes, and 3,600 cells distributed in 3 clusters (corresponding to an average of 200 cells per sample in each cluster). For the simulation study when varying the number of cells (Figure 5 and Supplementary Figure 3), 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 cluster. 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), which allows simulating cluster-specific batch effects [5,14]. All *muscat* simulation studies, as well as the *Kang* non-null data analysis, were performed by editing the original snakemake workflow from Crowell *et al.* [5]. PB methods were applied on aggregated data by summing cell-level measurements; for differential testing, we used *muscat*'s *pbDS* function [5]. Mixed model methods were implemented, via *muscat*'s *mmDS* function, using the same approaches as in Crowell *et al.* [5]: 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).

P-values adjustment

All p-values were adjusted via Benjamini-Hochberg correction [2]. In *diffcyt* simulations we used globally adjusted p-values for all methods, i.e., p-values from all clusters are jointly adjusted once. However, since PB methods were found to be over-conservative when globally adjusting p-values [5], in *muscat* simulations and *Kang* discovery analyses, we used locally adjusted p-values for all methods.

Software versions

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

References

[1] R. A. Amezquita, A. T. Lun, E. Becht, V. J. Carey, L. N. Carpp, L. Geistlinger, F. Marini, K. Rue-Albrecht, D. Risso, C. Soneson, et al. Orchestrating single-cell analysis with bioconductor. *Nature methods*, 17(2):137–145, 2020.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[3] B. Bodenmiller, E. R. Zunder, R. Finck, T. J. Chen, E. S. Savig, R. V. Bruggner, E. F. Simonds, S. C. Bendall, K. Sachs, P. O. Krutzik, et al. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nature biotechnology*, 30(9):858–867, 2012.

[4] H. L. Crowell. *muscData: Multi-sample multi-group scRNA-seq data*, 2020. R package version 1.1.2.

[5] H. L. Crowell, C. Soneson, P.-L. Germain, D. Calini, L. Collin, C. Raposo, D. Malhotra, and M. D. Robinson. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature Communications*, 11(1):1–12, 2020.

[6] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[7] M. R. Doherty, H. Cheon, D. J. Junk, S. Vinayak, V. Varadan, M. L. Telli, J. M. Ford, G. R. Stark, and M. W. Jackson. Interferon-beta represses cancer stem cell properties in triple-negative breast cancer. *Proceedings of the National Academy of Sciences*, 114(52):13792–13797, 2017.

[8] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome biology*, 16(1):1–13, 2015.

[9] C. Hafemeister and R. Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome biology*, 20(1):1–15, 2019.

[10] H. M. Kang, M. Subramaniam, S. Targ, M. Nguyen, L. Maliskova, E. McCarthy, E. Wan, S. Wong, L. Byrnes, C. M. Lanata, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89, 2018.

[11] P. V. Kharchenko, L. Silberstein, and D. T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

[12] K. D. Korthauer, L.-F. Chu, M. A. Newton, Y. Li, J. Thomson, R. Stewart, and C. Kendziorski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome biology*, 17(1):222, 2016.

[13] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, 2014.

[14] A. Lütge, J. Zyprych-Walczak, U. B. Kunzmann, H. L. Crowell, D. Calini, D. Malhotra, C. Soneson, and M. D. Robinson. Cellmixs: quantifying and visualizing batch effects in single-cell rna-seq data. *Life science alliance*, 4(6), 2021.

[15] F. J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[16] D. J. McCarthy, K. R. Campbell, A. T. Lun, and Q. F. Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186, 2017.

[17] M. Nowicka, C. Krieg, L. M. Weber, F. J. Hartmann, S. Guglietta, B. Becher, M. P. Levesque, and M. D. Robinson. CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research*, 6, 2017.

[18] B. Phipson and G. K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9:Article39, 2010.

[19] X.-Q. Qin, N. Tao, A. Dergay, P. Moy, S. Fawell, A. Davis, J. M. Wilson, and J. Barsoum. Interferon-β gene therapy inhibits tumor formation and causes regression of established tumors in immune-deficient mice. *Proceedings of the National Academy of Sciences*, 95(24):14411–14416, 1998.

[20] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[21] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[22] C. Soneson and M. D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods*, 15(4):255, 2018.

[23] J. W. Squair, M. Gautier, C. Kathe, M. A. Anderson, N. D. James, T. H. Hutson, R. Hudelle, T. Qaiser, K. J. Matson, Q. Barraud, et al. Confronting false discoveries in single-cell differential expression. *bioRxiv*, 2021.

[24] P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, 7:39921, 2017.

[25] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220), 2015.

[26] C. A. Vallejos, J. C. Marioni, and S. Richardson. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol*, 11(6):e1004333, 2015.

[27] T. Wang, B. Li, C. E. Nelson, and S. Nabavi. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC bioinformatics*, 20(1):40, 2019.

[28] L. M. Weber, M. Nowicka, C. Soneson, and M. D. Robinson. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications biology*, 2(1):1–11, 2019.

[29] L. M. Weber and C. Soneson. Hdcytodata: Collection of high-dimensional cytometry benchmark datasets in bioconductor object formats. *F1000Research*, 8, 2019.

[30] Y. Zhang, L. Zheng, L. Zhang, X. Hu, X. Ren, and Z. Zhang. Deep single-cell RNA sequencing data of individual T cells from treatment-naive colorectal cancer patients. *Scientific data*, 6(1):1–15, 2019.