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

High-throughput single-cell data reveal an unprecedented view of cell identity and allow showing complex variations between conditions; nonetheless, most methods for differential expression target differences in the mean, and struggle to identify changes where the mean is not or marginally affected. Here, 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 (scRNA-seq) and high-dimensional flow or mass cytometry (HDCyto) data. *distinct* is based on a hierarchical non-parametric permutation approach and, by comparing empirical cumulative distribution functions (ECDFs), 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 data, where *distinct* shows very 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 (HDCyto) 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) [23], where genes or proteins vary in specific sub-populations of cells (e.g., a cytokine response in myeloid cells but not in other leukocytes [9]). Classical bulk differential expression methods have been shown to perform well when used on single-cell measurements [18, 22] and on aggregated data (i.e., averages or sums across cells), also referred to as pseudo-bulk (PB) [4, 23]. 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 [11, 22]. In addition to bulk and PB tools, other methods were specifically proposed to perform differential analyses on single-cell data (notably, *scDD* [11], *SCDE* [10], *MAST* [7], *BASiCS* [21] and mixed models [19]). Nevertheless, they all present significant limitations: BASiCS does not perform cell-type specific differential testing between conditions, scDD does not handle covariates and biological replicates, while SCDE, MAST and mixed models performed poorly when detecting differential patterns that do not involve the mean [4, 11].

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) of the individual (e.g., single-cell) measurements of each sample, and compares the ECDFs to identify changes between conditions, even when the mean is unchanged or marginally involved (Figure 1 and Supplementary Figure 1). The differences between the ECDFs of the groups, computed at several cutoffs, are used to calculate a test statistic, whose null distribution is estimated via a hierarchical non-parametric permutation approach (see Methods). A major disadvantage of permutation tests, which often restricts its usage on bio-

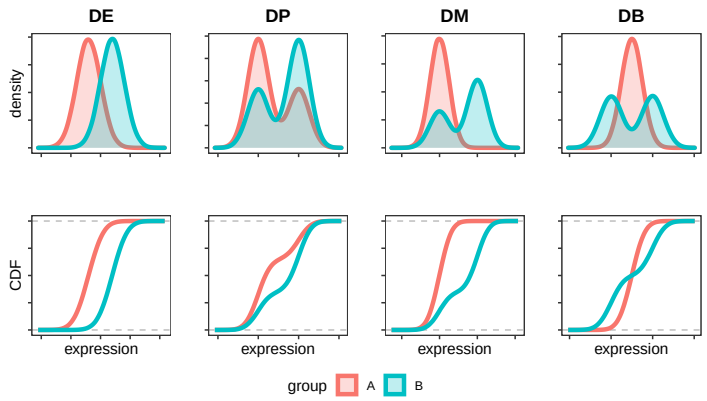

Figure 1: Cumulative distribution functions (CDFs) unravel differences between distributions. Density (top panels) and CDF (bottom panels) of four differential patterns, as proposed by Korthauer et. al. [11]: traditional differential expression (DE), differential proportion (DP), differential modality (DM), and both differential modality and different component means (DB).

logical data, is that usually there are not enough permutations in 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-type specific covariates, such as batch effects (see Methods).

Simulation studies

We conducted an extensive benchmark, based on scRNA-seq and HDCyto simulated and experimental datasets to investigate *distinct*'s ability to identify differential patterns in sub-populations of cells. First, we considered the semi-simulated HDCyto data from Weber *et al.* [23], where spike-in signals were computationally introduced in experimental data, hence maintaining the properties of real biological data while also embedding a known ground truth signal. In particular, we considered three datasets from Weber *et al.* [23]: the main DS dataset and two more where differential effects were diluted by 50 and 75%. Each dataset consists of 14 protein markers, 88,435 cells separated into 8 cell types (identified via unsupervised clustering [14]), and two groups (with and without spike-in signal) of 8 samples each. We evaluated *distinct* and two methods from *diffcyt*, based on *limma* [16] and linear mixed models (LMM), which outperformed competitors on these same data [23]. All methods exhibit similar performance in the main simulation; however, as the magnitude of the differential effect

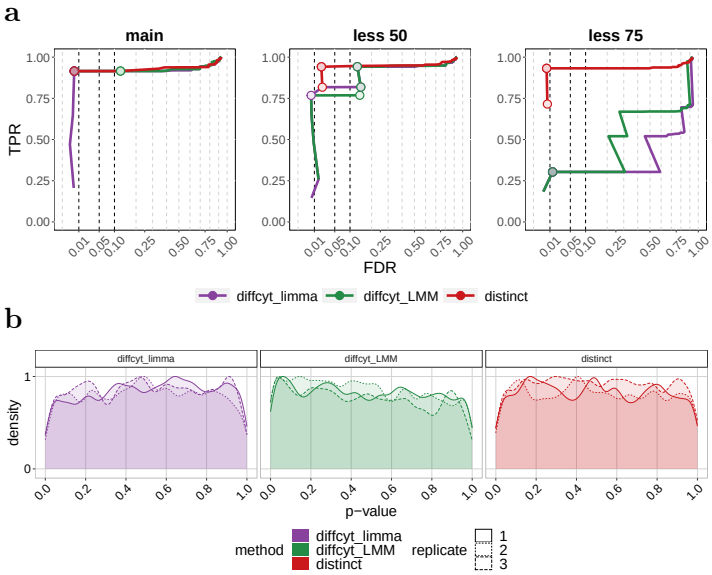

Figure 2: *distinct* shows high power while controlling for FPR and FDR. (a) TPR vs. FDR in *diffcyt* semi-simulated data. 'main' (first row), 'less 50' (second row) and 'less 75' (third row) indicate the main simulation, and those where differential effects are diluted by 50 and 75%, respectively. Each simulation consists of 88,435 cells and two groups of 8 samples each. Circles indicate observed FDR for 0.01, 0.05, and 0.1 significance thresholds. **(b)** 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.* [23], cells were clustered in an unsupervised manner.

decreases, *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) (Figure 2a and Supplementary Figure 2). We also considered the three replicate null datasets from Weber *et al.* [23] (i.e., with no differential effect), containing 14 protein markers and 88,438 cells across 8 cell types, and found that all methods display approximately uniform p-values (Figure 2b).

We further simulated droplet scRNA-seq data via *muscat* [4] (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 (traditional 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, and DB (both differential modality and different component means) compares a unimodal and a bimodal distribution with the same overall mean (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*

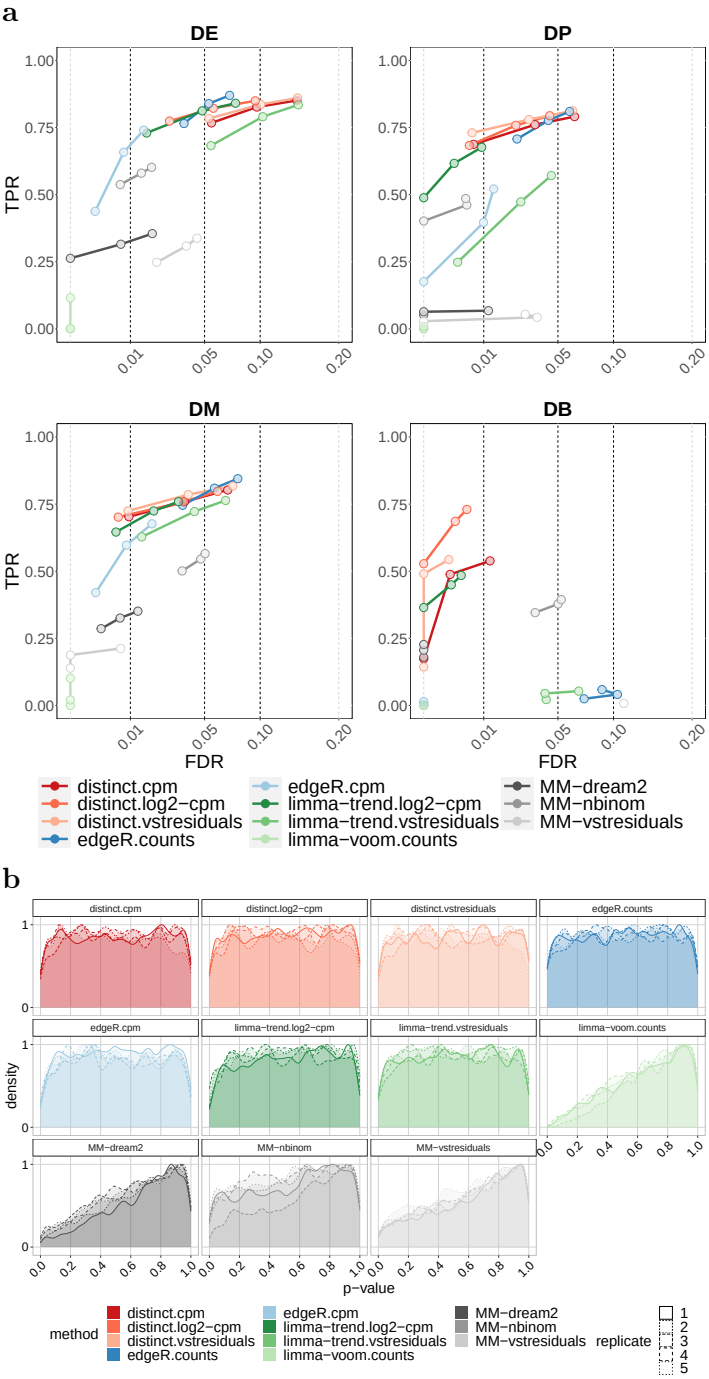

Figure 3: *distinct* identifies various differential patterns and controls for the FPR and FDR. (a) TPR vs. FDR in *muscat* simulated data; DE, DP, DM and DB refer to the differential profiles illustrated in Figure 1. Results are averages across the five simulation replicates. Circles indicate observed FDR for 0.01, 0.05, and 0.1 significance thresholds. **(b)** Density of raw p-values in *muscat* null simulated data; each replicate represents a different null simulation. In both **(a)** and **(b)**, two groups of 3 samples are compared and, on average, 200 cells are available for every sample in each of three clusters.

(vstresiduals) [8]. We compared *distinct* to several PB approaches from *muscat*, based on *edgeR* [17], *limma-voom* and *limma-trend* [16], which emerged among the best performing methods in Crowell *et al.* comprehensive benchmark [4]. We further considered three methods from *muscat* based on mixed models (MM), namely *MM-dream2*, *MM-vstresiduals* and *MM-nbinom* (see Methods). In all scenarios and on all three input data, *distinct* shows very favourable performance; in particular, for DB, it achieves significantly higher TPR, especially when using *log2-CPMs* (Figure 3a). Except for DB, the best performing PB methods, particularly *edgeR.counts* and *limma-trend.log2-CPMs*, display similar performance to *distinct*, while MM methods have significantly lower statistical power in all differential scenarios. We further considered 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 (FPs), with *distinct*, *edgeR* and *limma-trend* showing approximately uniform p-values for all types of input data (Figure 3b). From a computational perspective, *distinct*, having to perform a permutation test for every gene-cluster combination, required an average time of 3.7 to 4.8 minutes per simulation, which is much higher than PB methods (0.1 to 0.2 minutes) but significantly lower than MM approaches (29.6 to 314.5 minutes) (Figure 4 and Supplementary Table 1). We also extended the previous simulations to add a sample-level cell-type specific batch effect: differential results are substantially unchanged (Supplementary Figure 3), 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 and MM methods, *distinct* achieves higher overall TPR, while controlling for the FDR, regardless of the number of available cells (Figure

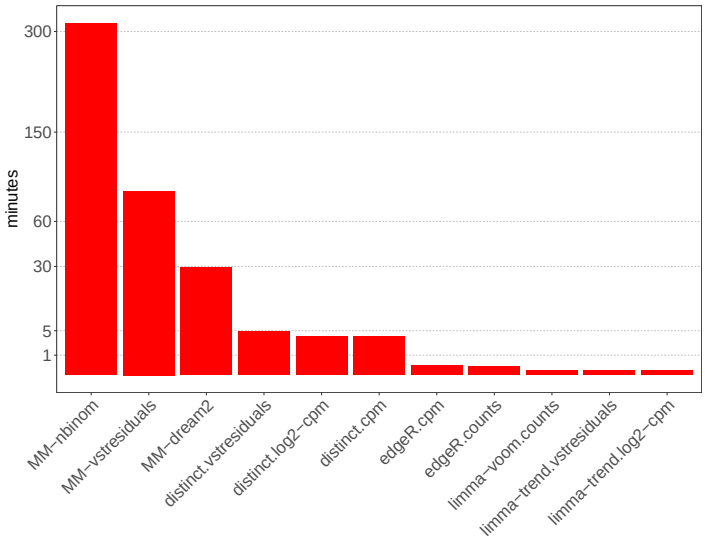

Figure 4: *distinct* requires more computational resources than PB methods, but significantly less than MM models. Average computing time, expressed in minutes, in *muscat* main simulations (Figure 3). For each method, times are averaged across simulation type (DE, DP, DM, DB and null) and, for each type, across the five replicate simulations; in each replicate 3,600 cells are available (200, on average, per cluster-sample combination). *distinct* and MM 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.

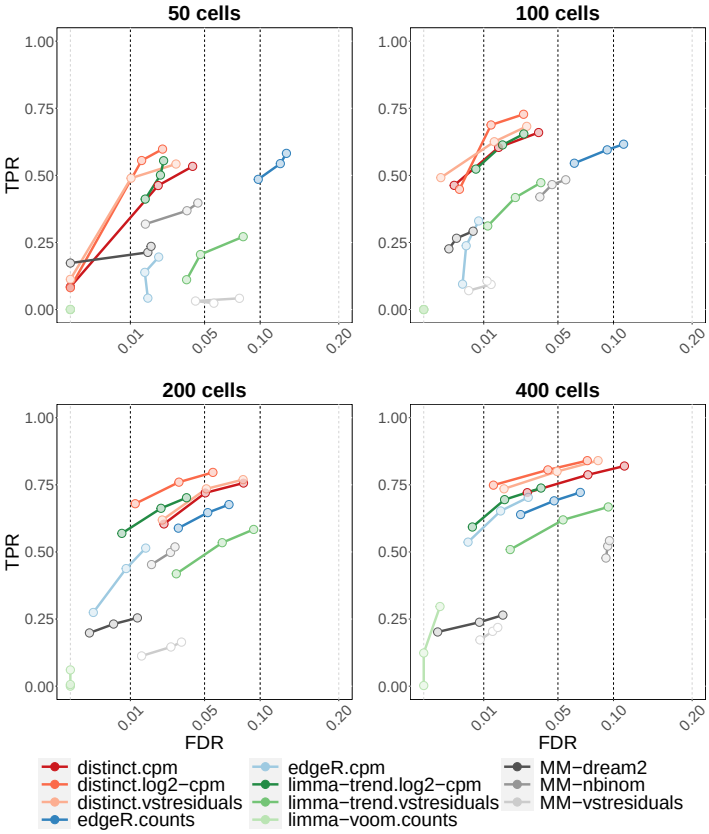

Figure 5: *distinct* achieves good performance when varying the number of available cells. TPR vs. FDR in *muscat* simulated data; with 50, 100, 200 and 400 cells per cluster-sample combination, corresponding to a total of 900, 1,800, 3,600 and 7,200 cells, respectively. Results are aggregated over the five replicate simulations of the four differential types, DE, DP, DM and DB, 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, and 0.1 significance thresholds.

5 and Supplementary Figure 4).

Experimental data analyses

In order to investigate false positive rates 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 models, for the real data analyses, we only considered *distinct* and PB methods. We considered gene-cluster combinations with at least 20 non-zero cells across all samples. The first dataset (labeled *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 (CRC) patients [25]. We automatically separated cells in 11 clusters (via *igraph* [5]), and generated replicate datasets, by randomly separating, three times, the 12 patients to two groups of size 6. The second dataset (labeled *Kang*) is a 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 [9].

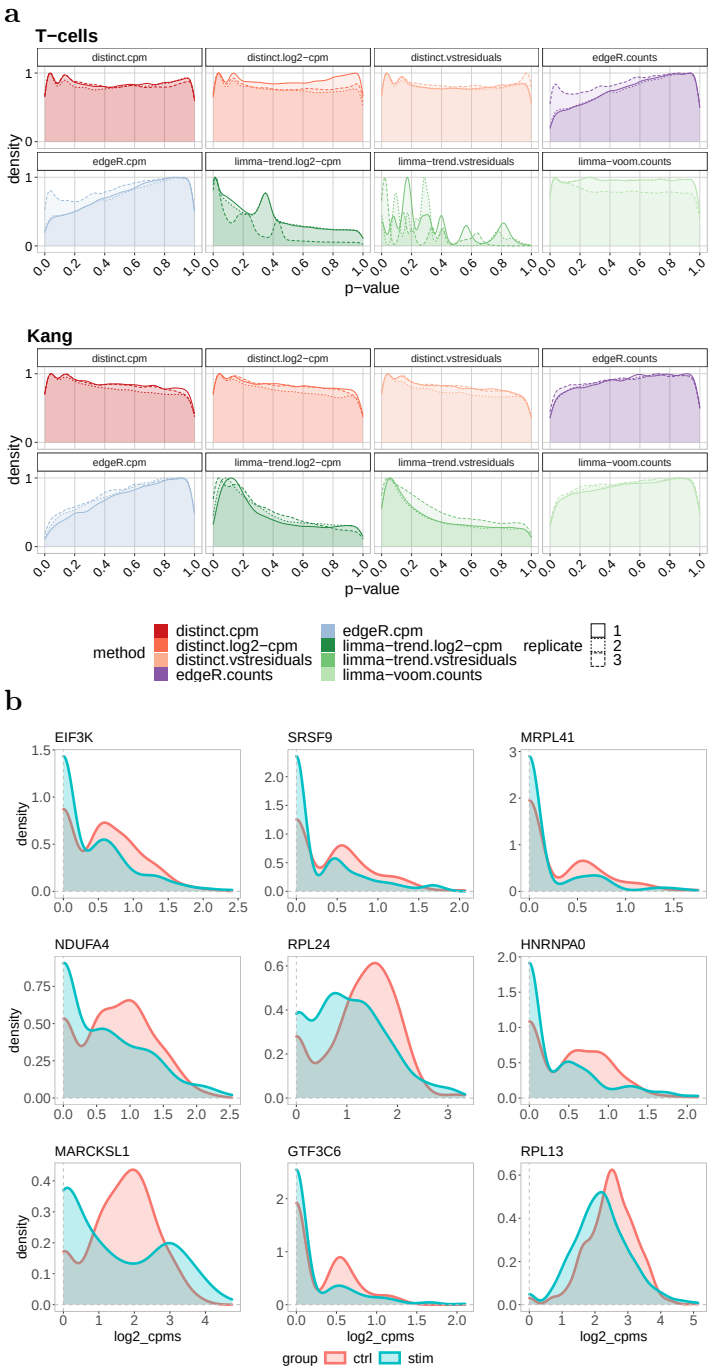

Figure 6: On real data, *distinct* discovers non-canonical differential patterns, and has almost-uniform null p-values. (a) Density of raw p-values in the null *T-cells* (top) and *Kang* (bottom) experimental data. Each replicate represents a random partition of samples in two groups. The *T-cells* data 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 (adjusted p-values < 0.05) by *distinct* on all input data, and not by any PB tool, on the *Kang* dataset when comparing controls and stimulated samples. Gene RPL13 was identified in FCGR3A+ Monocytes cells, while all other genes in Dendritic cells.

The full dataset contains 35,635 genes and 29,065 cells, which are separated 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 6a and Supplementary Tables 3-4).

Finally, we 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, 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 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 6b 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 [20]. This is an interesting association, considering that INF-β stimulation is known to interfere and inhibit tumor progression [6,15]. Finally, Supplementary Figures 9-17 show how *distinct* can identify differences between groups of distributions even when only a portion of the ECDF varies.

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 or marginally affected. To overcome present limitations, we have introduced *distinct*, a general method to identify differential patterns between groups of distributions. Our method exhibits favourable performance in both simulated and experimental datasets, 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 Zenodo with DOI: 10.5281/zenodo.4314375. The *diffcyt* simulated data is available via FlowRepository (accession ID FR-FCM-ZYL8 [23]) and *HDCytoData* R Bioconductor package [24]; the *Kang* dataset can be accessed via *muscData* R Bioconductor package [3]; the *T-cells* dataset is deposited on the European Genome-phenome (accession id EGAD00001003910 [25]).

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

Comparing ECDFs

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. For each sample, we compute the empirical cumulative distribution function of the observations (e.g., single-cell measurements); the ECDF for the i-th sample in the j-th group, is denoted by $ecdf_i^{(j)}(.)$, for $j \in \{A, B\}$ and $i = 1, \ldots, N_j$. We then define K

(25 by default) equally spaced cut-offs between the minimum, min, and maximum, max, values observed across all samples: $b_1, \ldots, b_K$, where $b_k = min + k \times l$, 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 DS analyses, these operations are repeated for every gene-cluster combination.

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:

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

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

If nuisance covariates (e.g., batch effects) are available, *distinct* fits a linear model with the input data (e.g., CPMs or log2-CPMs) as response variable, and the covariates as predictors. 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 [13], while vstresiduals were calculated via *sctransform* R package [8] (except for the *T-cells* data, where, due to a failure of *sctransform*'s variance stabilizing normalization, we used *DESeq2*'s vst transformation [12]).

diffcyt simulation study

The *diffcyt* dataset originates from experimental mass cytometry data of healthy peripheral blood mononuclear cells from two paired groups of 8 samples each [2]; one group contains unstimulated cells, while the other was stimulated with B cell receptor/Fc receptor cross-linker. In Weber *et al.* [23], semi-simulated data were generated by splitting 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 normalized and cells clustered via *diffcyt*'s *transformData* and *generateClusters* functions, respectively. As in Weber *et al.* [23], 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 [14, 23]. *diffcyt*'s *limma* and LMM methods were applied via *diffcyt*'s *testDS_limma* and *testDS_LMM* functions, respectively [23].

muscat simulation study

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 main and batch effect 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. When varying the number of cells, we adjusted these numbers to guarantee a comparable filter: when simulating 50, 100, 200 and 400 cells per sample-cluster combination, we considered a simulated expression mean threshold of 0.8, 0.4, 0.2 and 0.1, respectively; for DB profiles we used 5 times bigger thresholds. For every simulations, five replicates were simulated, and results were averaged across replicates. In the main simulation (Figure 3) 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 Figures 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 cluster. For the differential simulations, we used log2-FC values of 1 for DE, 1.5 for DP and DM, and 3 for DB. 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. We simulated 2 batches, b_1 and b_2, with group A having two samples associated to b_1 and one to b_2, and group B with two samples from batch b_2 and one from b_1. 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.* [4]. PB methods were applied on aggregated data by summing cell-level measurements; for differential testing, we used *muscat*'s *pbDS* function [4]. Mixed model methods were implemented, via *muscat*'s *mmDS* function, using the same approaches as in Crowell *et al.* [4]: 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 in all methods (*distinct*, PB and MM) by modelling the patient id as a covariate.

P-values adjustment

All p-values were adjusted via Benjamini-Hochberg correction [1]. 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 [4], 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] 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.

[2] 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.

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

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

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

[25] 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.