diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering

Lukas M. Weber[1,2], Malgorzata Nowicka[1,2,†], Charlotte Soneson[1,2,‡], Mark D. Robinson[1,2,*]

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
[2]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland

†Current address: F. Hoffmann-La Roche AG, Basel, Switzerland

‡Current address: Friedrich Miescher Institute for Biomedical Research and SIB Swiss Institute of Bioinformatics, Basel, Switzerland

*Corresponding author

November 22, 2018

1 Abstract

High-dimensional flow and mass cytometry allow cell types and states to be characterized in great detail by measuring expression levels of more than 40 targeted protein markers per cell. Here, we present `diffcyt`, a new computational framework for differential discovery analyses in these datasets, based on (i) high-resolution clustering and (ii) empirical Bayes moderated tests adapted from transcriptomics. Our approach provides improved statistical performance, including for rare cell populations, along with flexible experimental designs and fast runtimes in an open-source framework.

2 Introduction

High-dimensional flow cytometry and mass cytometry (or CyTOF, for 'cytometry by time-of-flight mass spectrometry') characterize cell types and states by measuring expression levels of pre-defined sets of surface and intracellular proteins in individual cells, using antibodies tagged with either fluorochromes (flow cytometry) or heavy metal isotopes (mass cytometry). Modern flow cytometry systems allow simultaneous detection of more than 20 proteins per cell, in thousands of cells per second [1]. In mass cytometry, the use of metal tags significantly reduces signal interference due to spectral overlap and autofluorescence, enabling detection of more than 40 proteins per cell in hundreds of cells per second [1, 2]. Recently, further increases in the number of detected proteins have been demonstrated using oligonucleotide-tagged antibodies and single-cell sequencing [3]; this has also been combined with single-cell RNA sequencing on the same cells [4, 5].

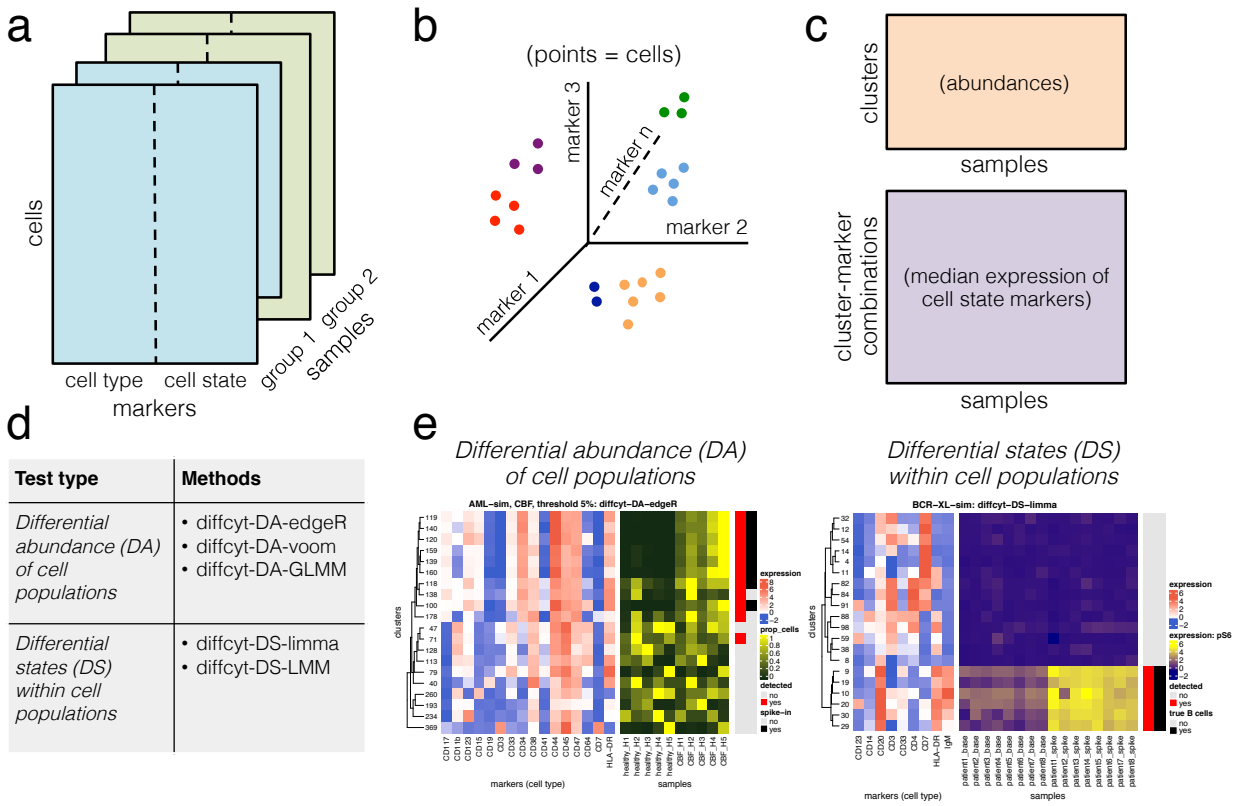
The rapid increase in dimensionality has led to serious bottlenecks in data analysis. Traditional analysis by visual inspection of scatterplots ('manual gating') is unreliable and inefficient in high-dimensional data, does not scale readily, and cannot easily reveal unknown cell populations [1]. Significant efforts have been made to develop computationally guided or automated methods that do not suffer from these limitations. For example, unsupervised clustering algorithms are commonly used to define cell populations in one or more biological samples. Recent benchmarking studies have demonstrated that several clustering methods can accurately detect known cell populations in low-dimensional flow cytometry data [6],

and both major and rare known cell populations in high-dimensional data [7]. A further benchmarking study comparing supervised methods for inferring cell populations associated with a censored continuous clinical variable demonstrated good performance for two methods using data of moderate dimensionality [8].

Several new methods have recently been developed for performing (partially) supervised analyses with the aim of inferring cell populations or states associated with an outcome variable in high-dimensional cytometry data, including `Citrus` [9], `CellCnn` [10], `cydar` [11], and a `classic` regression-based approach [12] (a similar regression-based approach was also recently described by [13]). However, these existing methods have a number of limitations (summarized in Supplementary Table 1). In particular: (i) detected features from `Citrus` cannot be ranked by importance, and the ranking of detected cells from `CellCnn` cannot be interpreted in terms of statistical significance; (ii) rare cell populations are difficult to detect with `Citrus` and `cydar` (by contrast, `CellCnn` is optimized for analysis of rare populations); (iii) the response variable in the models for `Citrus` and `CellCnn` is the outcome variable, which makes it difficult to account for complex experimental designs; and (iv) `CellCnn` and `cydar` do not distinguish between 'cell type' and 'cell state' (or functional) markers, which can make interpretation difficult.

Here, we present `diffcyt`, a new computational framework using high-resolution unsupervised clustering together with supervised statistical analyses to detect cell populations or states associated with an outcome variable in high-dimensional cytometry data. Figure 1 provides an overview of the `diffcyt` methodology (a detailed description is included in Supplementary Note 1). Clustering is used to define cell populations, and empirical Bayes moderated tests adapted from the transcriptomics field are used for differential analysis. By default, our implementation uses the `FlowSOM` clustering algorithm [14], given its strong performance and fast runtimes [7]. For the differential analyses, we use methods from `edgeR` [15, 16], `limma` [17], and `voom` [18], which are widely used in transcriptomics; in addition, we include alternative methods adapted from the `classic` regression-based framework [12]. In principle, other high-resolution clustering algorithms or differential testing methods could also be substituted.

Our methods consolidate several aspects of functionality from the existing methods described above. Similar to `cydar` and the `classic` regression framework, our model specification uses the cytometry-measured features (cell population abundances or median expression of cell state markers within populations) as response variables, which enables analysis of complex experimental designs, including batch effects, paired designs, and continuous covariates. Linear contrasts enable testing of a wide range of hypotheses. Rare cell populations can easily be investigated, since the use of high-resolution clustering ensures that rare populations are unlikely to be merged into larger ones. In addition, as in `Citrus` and the `classic` regression framework, we optionally allow the user to split the set of protein markers into 'cell type' and 'cell state' markers. In this setup, cell type markers are used to define clusters representing cell populations, which are tested for differential abundance (DA); and median cell state marker signals per cluster are used to test for differential states (DS) within populations. We note that the underlying definitions of cell type and cell state can be challenging to apply to observed data, and may partially overlap. In general, 'cell type' refers to relatively stable or permanent features of a cell's identity, while 'cell state' refers to transient features such as signaling or other functional states or the cell cycle [19–21]. In our view, providing the ability to maintain this distinction within the methodology greatly improves biological interpretability, since the results can be directly linked back to known cell types or populations of interest [12]. Finally, our methods have fast runtimes, enabling exploratory and interactive analyses.



Test type	Methods
Differential abundance (DA) of cell populations	• diffcyt-DA-edgeR • diffcyt-DA-voom • diffcyt-DA-GLMM
Differential states (DS) within cell populations	• diffcyt-DS-limma • diffcyt-DS-LMM

Figure 1. Overview of diffcyt methodology. The `diffcyt` framework applies high-resolution clustering and empirical Bayes moderated tests for differential discovery analyses in high-dimensional cytometry data. (a) Input data are provided as tables of protein marker expression values per cell, one table per sample. Markers may be split into 'cell type' and 'cell state' categories; in the standard setup, cell type markers are then used for clustering. (b) High-resolution clustering summarizes the data into a large number (e.g. 100–400) of clusters representing cell subsets. (c) Features are calculated at the cluster level, including cluster cell counts (abundances), and median expression of cell state markers within clusters. (d) Differential testing methods can be grouped into two types: differential abundance (DA) of cell populations, and differential states (DS) within cell populations. Test results are returned as sets of significant detected clusters (DA tests) or cluster-marker combinations (DS tests). (e) Results are interpreted with the aid of visualizations, such as heatmaps. Example heatmaps show cluster phenotypes (expression profiles) and differential signal of interest (relative cluster abundances or expression of signaling marker pS6, by sample), with annotation for detected significant clusters (red) and true differential clusters (black). A detailed description of the `diffcyt` methodology is provided in Supplementary Note 1; existing methods are described in Supplementary Note 2.

3 Results

We demonstrate the performance of our methods on two semi-simulated datasets ('AML-sim' and 'BCR-XL-sim') and two published experimental datasets ('Anti-PD-1' and 'BCR-XL'). The semi-simulated datasets have been constructed by computationally introducing an artificial signal of interest (an *in silico* spike-in signal) into experimental data, thus reflecting the properties of real experimental data while also including a known 'ground truth' that can be used to calculate performance metrics. The experimental datasets, which do not contain a ground truth, are evaluated in qualitative terms. A complete description of all datasets is provided in Supplementary Note 3 (including Supplementary Tables 2–5 and Supplementary Figures 2–3).

The AML-sim dataset evaluates performance for detecting differential abundance of rare cell populations. The dataset contains a spiked-in population of acute myeloid leukemia (AML) blast cells, in a comparison of 5 vs. 5 paired samples of otherwise healthy bone marrow mononuclear cells, which simulates the phenotype of minimal residual disease in AML patients (the data generation strategy is adapted from [10], and uses original data from [22]). The simulation was repeated for two subtypes of AML (cytogenetically normal, CN; and core binding factor translocation, CBF), and three thresholds of abundance for the spiked-in population (5%, 1%, and 0.1%). Figure 2(a) displays representative results for one subtype (CN) and one threshold (1%), for all `diffcyt` DA methods as well as `Citrus`, `CellCnn`, and `cydar` (complete results are included in Supplementary Figure 3). Methods `diffcyt-DA-edgeR`, `diffcyt-DA-voom`, and `CellCnn` give the best performance; the `diffcyt` results can also be interpreted as adjusted p-values, enabling a standard statistical framework where a list of significant 'detected' clusters is determined by specifying a cutoff for the false discovery rate (FDR). `diffcyt-DA-GLMM` has inferior error control at the given FDR cutoffs, and reduced sensitivity at the highest spike-in threshold (5%). `Citrus` detects only a subset of the spiked-in cells, and `cydar` cannot reliably distinguish these rare populations. Figure 2(b) displays p-value distributions from an accompanying 'null simulation', where no true spike-in signal was included; the p-value distributions for the `diffcyt` methods are approximately uniform, indicating good error control and model fit (additional replicates

are included in Supplementary Figure 4). Figure 2(c) illustrates the expression profiles (phenotypes) and relative abundances by sample for the detected and true differential clusters (additional heatmaps are included in Supplementary Figure 5). Figure 2(d) demonstrates the effect of varying the number of clusters across a broad range (between 9 and 1,600). Performance is reduced when there are too few clusters (due to merging of populations) or too many clusters (due to low power). The number of clusters is the main parameter choice in the `diffcyt` methods; an optimum is achieved around 400 clusters for this dataset (the remaining thresholds and condition are shown in Supplementary Figure 6).

Additional results provide further details on overall performance and robustness of the `diffcyt` methods. The top detected clusters represent high-precision subsets of the spiked-in population, confirming that the high-resolution clustering strategy has worked as intended (Supplementary Figure 7). Filtering clusters with low cell counts (using default parameters) did not remove any clusters from this dataset. An alternative implementation of the `diffcyt-DA-voom` method (using random effects for paired data) gives similar overall performance (Supplementary Figure 8). Using `FlowSOM` 'meta-clustering' to generate 40 merged clusters instead of testing at high resolution worsens both error control and sensitivity (Supplementary Figure 9). The influence of random seeds used for the clustering and data generation procedures is greatest at the 0.1% threshold, as expected (Supplementary Figures 10–11). Similarly, additional simulations containing 'less distinct' populations of interest (see Supplementary Note 3) show that reducing signal strength has a strong negative influence on performance at the 0.1% threshold (Supplementary Figure 12). Using smaller sample sizes (2 vs. 2) affects performance noticeably at the lower thresholds (Supplementary Figure 13). Finally, runtimes are fastest for methods `diffcyt-DA-edgeR` and `diffcyt-DA-voom` (Supplementary Figure 14).

The second dataset, BCR-XL-sim, evaluates performance for detecting differential states within cell populations. This dataset contains a spiked-in population of B cells stimulated with B cell receptor / Fc receptor cross-linker (BCR-XL), in a comparison of 8 vs. 8 paired samples of healthy peripheral blood mononuclear cells (original data sourced from [23]). The stimulated B cells have elevated expression of several signaling state markers, in particular phosphorylated ribosomal protein S6 (pS6); methods are evaluated by their ability to detect

differential expression of pS6 within the population of B cells. Figure 2(e) summarizes performance for the `diffcyt` DS methods and the existing methods. The `diffcyt` methods give the best performance, with `diffcyt-DS-limma` having better error control. `Citrus` and `CellCnn` detect differential expression of pS6 for only a subset of the spiked-in cells, and `cydar` gives poor performance (likely due to ambiguity in assigning cells to overlapping 'hyperspheres' in the high-dimensional space in order to calculate performance metrics). Figure 2(f) displays p-value distributions from a null simulation; p-values are approximately uniform across replicates, as previously (additional replicates are included in Supplementary Figure 15). Figure 2(g) displays expression profiles of detected and true differential clusters, along with expression by sample of the signaling marker pS6 (additional heatmaps are included in Supplementary Figure 16). Figure 2(h) demonstrates the effect of varying the number of clusters. Performance is reduced when there are too few or too many clusters; for this dataset, an optimum is observed across a broad range, including 100 clusters.

As previously, the top detected clusters represent high-precision subsets of the population of interest (Supplementary Figure 17). Filtering with default parameters did not remove any clusters. To judge the benefit of splitting markers into 'cell type' and 'cell state' categories, we re-ran the analyses treating all markers as 'cell type' (i.e. used for clustering), and using methods to test for DA instead of DS. This gave similar performance, but makes interpretation more difficult, since clusters may mix elements from canonical cell type and cell state phenotypes (Supplementary Figure 18). Alternative implementations of `diffcyt-DS-limma` (using random effects for paired data) and `diffcyt-DS-LMM` (using fixed effects for paired data) give similar performance overall (Supplementary Figure 19). For this dataset, using `FlowSOM` meta-clustering to merge clusters does not reduce performance (Supplementary Figure 20). Varying random seeds for the clustering and data generation procedures does not significantly affect performance (Supplementary Figures 21–22). Additional simulations containing 'less distinct' populations of interest (see Supplementary Note 2) show deteriorating performance when the signal is reduced by 75% (Supplementary Figure 23). Using smaller sample sizes (4 vs. 4 and 2 vs. 2) worsens error control, especially for `diffcyt-DS-LMM` (Supplementary Figure 24). Runtimes are fastest for `diffcyt-DS-limma` (Supplementary Figure 25).

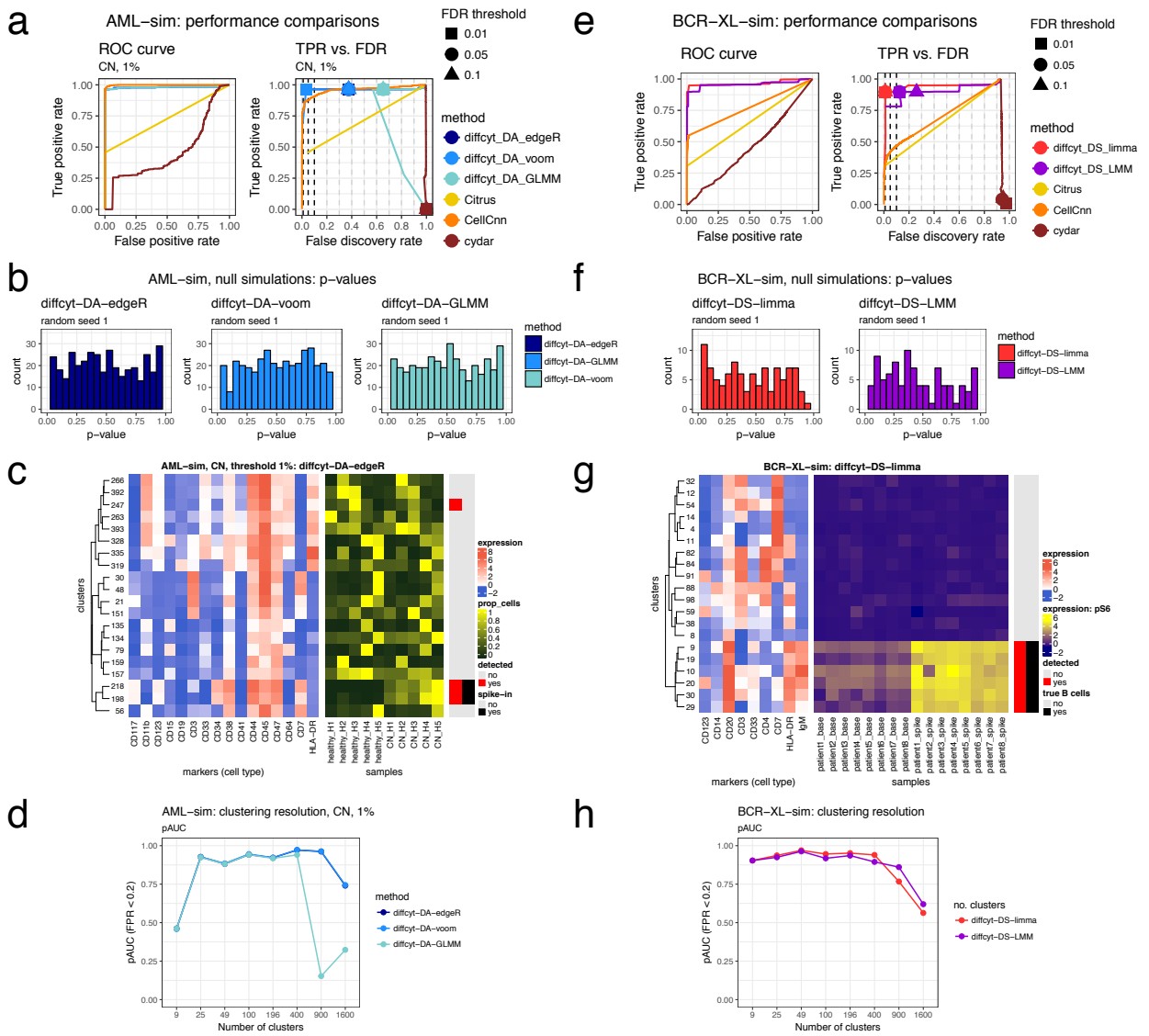


Figure 2. Results of performance evaluations and benchmarking for semi-simulated datasets. (a, e) Performance metrics for (a) dataset AML-sim, testing for differential abundance (DA) of cell populations; and (e) dataset BCR-XL-sim, testing for differential states (DS) within cell populations. Panels show (i) receiver operating characteristic (ROC) curves, and (ii) true positive rate (TPR) vs. false discovery rate (FDR) (also indicating observed TPR and FDR at FDR cutoffs 1%, 5%, and 10%). For dataset AML-sim, representative results for one condition (CN vs. healthy) and abundance threshold (1%) are shown (complete results are included in Supplementary Figure 3). (b, f) Results for additional 'null simulations', where no true spike-in signal was included; p-value distributions are approximately uniform (additional replicates are included in Supplementary Figures 4 and 15). (c, g) Heatmaps displaying phenotypes (expression profiles) of detected and true differential clusters, along with the signal of interest (relative cluster abundances or expression of signaling marker pS6, by sample), for methods `diffcyt-DA-edgeR` and `diffcyt-DS-limma`. Expression values represent median arcsinh-transformed expression per cluster across all samples (left panels) or by individual samples (right panel in (g)); rows (clusters) are grouped by hierarchical clustering with Euclidean distance and average linkage. Each heatmap shows the top 20 most highly significant clusters for that method. Vertical annotation indicates detected significant clusters at 10% FDR (red) and clusters containing >50% true spiked-in cells (black). (Additional heatmaps are included in Supplementary Figures 5 and 16). (d, h) Results for varying clustering resolution (between 9 and 1,600 clusters per dataset); showing partial area under ROC curves (pAUC) for false positive rates (FPR) <0.2 (additional figures are included in Supplementary Figure 6). Performance metric plots generated using `iCOBRA` [24]; heatmaps generated using `ComplexHeatmap` [25].

In order to demonstrate our methods on experimental data, we re-analyzed a dataset from a recent study using mass cytometry to characterize immune cell subsets in peripheral blood from melanoma patients treated with anti-PD-1 immunotherapy [26] ('Anti-PD-1' dataset). In this study, differential signals were detected for a number of cell populations, both in response to treatment and in 'baseline' comparisons before treatment, between groups of patients classified as 'responders' and 'non-responders' to treatment. One key result was the identification of a small subpopulation of monocytes, with frequency in 'baseline' samples (prior to treatment) strongly associated with responder status. The relatively rare frequency made this population difficult to detect; in addition, the dataset contained a strong batch effect due to sample acquisition on two different days [26]. Using method `diffcyt-DA-edgeR` to perform a differential comparison between baseline samples from the responder and non-responder patients (and taking into account the batch effect), we correctly identified three significant differentially abundant clusters (at an FDR cutoff of 10%) with phenotypes that closely matched the subpopulation of monocytes detected in the original study (CD14$^+$ CD33$^+$ HLA-DR^{hi} ICAM-1$^+$ CD64$^+$ CD141$^+$ CD86$^+$ CD11c$^+$ CD38$^+$ PD-L1$^+$ CD11b$^+$ monocytes) (clusters 317, 358, and 380; Supplementary Figure 26). One additional cluster with a different phenotype was also detected (cluster 308). However, these results were sensitive to the choice of random seed for the clustering: in 5 additional runs using different random seeds, we detected between 0 and 4 significant differentially abundant clusters (at 10% FDR) per run; clusters matching the expected phenotype were detected in 4 out of the 5 runs (Supplementary Figure 27).

For a second evaluation on experimental data, we re-analyzed the original (unmodified) data from the BCR-XL stimulation condition in [23] ('BCR-XL' dataset). This dataset contains strong differential signals for several signaling state markers in several cell populations, as previously described [12, 23]. Using method `diffcyt-DS-limma`, we reproduced several of the major known signals, including strong differential expression of: (i) pS6, pPlcg2, pErk, and pAkt (elevated), and pNFkB (reduced, in BCR-XL stimulated condition) in B cells (identified by expression of CD20); (ii) pBtk and pNFkB in CD4$^+$ T cells (identified by expression of CD3 and CD4); and (iii) pBtk, pNFkB, and pSlp76 in natural killer (NK) cells (identified by

expression of CD7). Here, phenotypes can be identified either by marker expression profiles (Supplementary Figure 28) or using reference population labels (Supplementary Figure 29).

4 Discussion

We have presented a new computational framework for performing flexible differential discovery analyses in high-dimensional cytometry data. Our methods are designed for two related but distinct discovery tasks: detecting differentially abundant cell populations, including rare populations; and detecting differential expression of functional or other cell state markers within cell populations. Compared to existing approaches, our methods provide improved detection performance on semi-simulated benchmark datasets, along with fast runtimes. We have also successfully recovered known differential signals in re-analyses of two published experimental datasets, including differential abundance of a highly specific rare population. Our methods can account for complex experimental designs, including batch effects, paired designs, and continuous covariates. In addition, the set of protein markers may be split into cell type and cell state markers, facilitating biological interpretability. Visualizations such as heatmaps can be used to interpret the high-resolution clustering results (for example, to judge whether groups of clusters form larger populations, and to identify the phenotype of detected clusters). Methods `diffcyt-DA-edgeR` (for DA tests) and `diffcyt-DS-limma` (for DS tests) achieved the best performance and fastest runtimes overall (Figure 2); we recommend these as the default choices.

One limitation of our framework is that groups of similar clusters cannot be automatically merged into larger cell populations with a consistent phenotype. For example, the clear group of detected clusters in Figure 2(g) would ideally be merged into a single population representing B cells. However, this is a difficult computational problem, since the optimal resolution depends on the biological setting, and any automatic merging must avoid merging rare cell populations into larger ones. Our high-resolution clustering approach instead provides a tractable 'middle ground' between discrete clustering and a continuum of cell populations; we return results directly at the level of high-resolution clusters, and let the user interpret them via visualizations. A related issue concerns the identification of cell population

phenotypes: our approach relies on visualizations and manual annotation of populations, which necessarily involves some subjectivity. Recently, several new methods have been published for automated labeling of cell populations [27], identification of simplified gating strategies to describe cell populations of interest [28, 29], or to compare cluster phenotypes [30]. These methods could be integrated within our framework to interpret detected differential clusters in a more automated manner. Similarly, clustering algorithms that generate biologically interpretable clusters could be used to improve interpretability [31].

A further limitation relates to batch effects: in datasets with strong batch effects, the high-resolution clustering may separate across batches, making it more difficult to distinguish the signal of interest. Aligning cell populations across batches is an active area of research in single-cell analysis (e.g. [32–35]); ideally, these methods will be integrated with frameworks for downstream differential analyses. Another issue concerns our strategy of summarizing cell state marker signals into median values. This strategy has advantages of simplicity, ease of interpretation, and fast runtimes. However, some information is necessarily lost, especially for markers with multi-modal distributions; good frameworks for flexible comparisons of full distributions are currently lacking. Additionally, splitting markers into groups representing cell type and cell state may be seen as a disadvantage in applications where this distinction is not clear. However, this step is optional: it is possible to run our methods using all markers for clustering (i.e. treating all markers as cell type) and testing for differential abundance (Supplementary Figure 18). Methods are also available to automatically group markers [12, 22, 36], although these should be used with care to ensure that cell population definitions are biologically plausible.

The main user parameter in our methods is the number of clusters. The optimal value depends on several factors, including the size of the dataset (number of cells and samples), the expected relative abundances of cell populations of interest, and the number of markers used to define cell populations. The number of clusters determines the number of statistical tests, and affects power through the multiple testing penalty and the counts per cluster. We recommend higher numbers of clusters when rare cell populations are of interest (for example, we used 400 clusters for the AML-sim dataset, and 100 clusters for the BCR-XL-sim dataset). Ultimately, this is a subjective choice for the user, which may also be explored interactively;

however, in our evaluations, good results were obtained over a range of values (Figure 2(d, h)). Most computational methods include one or more parameters that can be adjusted by the user; in our view, one of the advantages of our approach is that the number of clusters is an intuitive parameter, with values that can be easily interpreted.

In general, we note that our methods are designed for 'discovery' analyses: all results should be explored and interpreted using visualizations, and any generated hypotheses must ultimately be validated with targeted confirmatory experiments. Our methods are implemented in the open-source R package `diffcyt`, available from Bioconductor (http://bioconductor.org/packages/diffcyt). The package includes comprehensive documentation and code examples, including an extended workflow vignette. Code to reproduce all analyses and figures from our benchmarking evaluations is available from GitHub (https://github.com/lmweber/diffcyt-evaluations), and data files from the benchmarking datasets are available from FlowRepository (FR-FCM-ZYL8) [37], allowing other researchers to extend and build on our analyses.

5 Supplementary Material

Supplementary Material includes additional details on the `diffcyt` methodology (Supplementary Note 1), comparisons with existing methods (Supplementary Note 2), benchmark datasets (Supplementary Note 3), and supplementary results (Supplementary Note 4). This includes Supplementary Tables 1–5 and Supplementary Figures 1–29, referred to in the main text. The Supplementary Material is available online as a single PDF document; this is intended as a reference for consultation by readers interested in methodological details required for reproducing or extending results.

6 Acknowledgments

The authors thank members of the Robinson Lab at the University of Zurich for valuable feedback. LMW was supported by a Forschungskredit (Candoc) grant from the University of Zurich (FK-17-100).

7 Author contributions

LMW and MDR developed methods, designed analyses, and wrote the manuscript. LMW implemented methods and performed analyses. MN developed methods and assisted with interpretation. CS assisted with designing analyses and interpretation. All authors read and approved the final manuscript.

8 Competing interests

The authors declare no competing interests.

References

[1] Saeys, Y., Van Gassen, S., and Lambrecht, B. N. (2016). Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, 16:449–462.

[2] Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E.-a. D., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I., Balderas, R. S., Plevritis, S. K., Sachs, K., Pe'er, D., Tanner, S. D., and Nolan, G. P. (2011). Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science*, 332:687–696.

[3] Shahi, P., Kim, S. C., Haliburton, J. R., Gartner, Z. J., and Abate, A. R. (2017). Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Scientific Reports*, 7(44447).

[4] Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868.

[5] Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., Moore, R., McClanahan, T. K., Sadekova, S., and Klappenbach, J. A. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, 35(10):936–939.

[6] Aghaeepour, N., Finak, G., The FlowCAP Consortium, The DREAM Consortium, Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., and Scheuermann, R. H. (2013). Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods*, 10(3):228–238.

[7] Weber, L. M. and Robinson, M. D. (2016). Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data. *Cytometry Part A*, 89A:1084–1096.

[8] Aghaeepour, N., Chattopadhyay, P., Chikina, M., Dhaene, T., Van Gassen, S., Kursa, M., Lambrecht, B. N., Malek, M., McLachlan, G. J., Qian, Y., Qiu, P., Saeys, Y., Stanton, R., Tong, D., Vens, C., Walkowiak, S., Wang, K., Finak, G., Gottardo, R., Mosmann, T., Nolan, G. P., Scheuermann, R. H., and Brinkman, R. R. (2016). A Benchmark for Evaluation of Algorithms for Identification of Cellular Correlates of Clinical Outcomes. *Cytometry Part A*, 89A:16–21.

[9] Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J., and Nolan, G. P. (2014). Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences of the United States of America*, pages E2770–E2777.

[10] Arvaniti, E. and Claassen, M. (2017). Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature Communications*, 8(14825):1–10.

[11] Lun, A. T. L., Richard, A. C., and Marioni, J. C. (2017). Testing for differential abundance in mass cytometry data. *Nature Methods*, 14(7):707–709.

[12] Nowicka, M., Krieg, C., Weber, L. M., Hartmann, F. J., Guglietta, S., Becher, B., Levesque, M. P., and Robinson, M. D. (2017). CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research*, version 2.

[13] Fonseka, C. Y., Rao, D. A., Teslovich, N. C., Korsunsky, I., Hannes, S. K., Slowikowski, K., Gurish, M. F., Donlin, L. T., Lederer, J. A., Weinblatt, M. E., Massarotti, E. M., Coblyn, J. S., Helfgott, S. M., Todd, D. J., Bykerk, V. P., Karlson, E. W., Ermann, J., Lee, Y. C., Brenner, M. B., and Raychaudhuri, S. (2018). Mixed-effects association of single cells identifies an expanded effector CD4+ T cell subset in rheumatoid arthritis. *Science Translational Medicine*, 10:eaaq0305.

[14] Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., and Saeys, Y. (2015). FlowSOM: Using Self-Organizing Maps for Visualization and Interpretation of Cytometry Data. *Cytometry Part A*, 87A:636–645.

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

[16] McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297.

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

[18] Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15:R29.

[19] Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 34(11):1145–1160.

[20] Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J. T., Theis, F. J., Uhlen, M., Oudenaarden, A. V., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N., and Human Cell Atlas Meeting Participants (2017). The Human Cell Atlas. *eLIFE*, 6(e27041):1–30.

[21] Zeng, H. and Sanes, J. R. (2017). Neuronal cell-type classification: challenges, opportunities and the path forward. *Nature Reviews Neuroscience*, 18:530–546.

[22] Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., Finck, R., Gedman, A. L., Radtke, I., Downing, J. R., Pe'er, D., and Nolan, G. P. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162:184–197.

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

[24] Soneson, C. and Robinson, M. D. (2016). iCOBRA: open, reproducible, standardized and live method benchmarking. *Nature Methods*, 13(4):283.

[25] Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849.

[26] Krieg, C., Nowicka, M., Guglietta, S., Schindler, S., Hartmann, F. J., Weber, L. M., Dummer, R., Robinson, M. D., Levesque, M. P., and Becher, B. (2018). High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nature Medicine*, 24(2):144–153.

[27] Abdelaal, T., van Unen, V., Höllt, T., Koning, F., Reinders, M. J., and Mahfouz, A. (2018). Predicting cell types in single cell mass cytometry data. *bioRxiv*.

[28] Becht, E., Simoni, Y., Coustan-Smith, E., Evrard, M., Cheng, Y., Ng, L. G., Campana, D., and Newell, E. W. (2018). Reverse-engineering flow-cytometry gating strategies for phenotypic labelling and high-performance cell sorting. *Bioinformatics*.

[29] Aghaeepour, N., Simonds, E. F., Knapp, D. J. H. F., Bruggner, R. V., Sachs, K., Culos, A., Gherardini, P. F., Samusik, N., Fragiadakis, G. K., Bendall, S. C., Gaudilliere, B., Angst, M. S., Eaves, C. J., Weiss, W. A., Fantl, W. J., and Nolan, G. P. (2018). GateFinder: projection-based gating strategy optimization for flow and mass cytometry. *Bioinformatics*.

[30] Platon, L., Pejoski, D., Gautreau, G., Targat, B., Le Grand, R., Beignon, A.-S., and Tchitchek, N. (2018). A computational approach for phenotypic comparisons of cell populations in high-dimensional cytometry data. *Methods*, 132:66–75.

[31] Commenges, D., Alkhassim, C., Gottardo, R., Hejblum, B., and Thiébaut, R. (2018). cytometree: A Binary Tree Algorithm for Automatic Gating in Cytometry Analysis. *Cytometry Part A*, 93A:1132–1140.

[32] Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*.

[33] Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*.

[34] Orlova, D. Y., Meehan, S., Parks, D., Moore, W. A., Meehan, C., Zhao, Q., Ghosn, E. E. B., Herzenberg, L. A., and Walther, G. (2018). QFMatch: multidimensional flow and mass cytometry samples alignment. *Scientific Reports*, 8(3291):1–14.

[35] Li, Y. H., Li, D., Samusik, N., Wang, X., Guan, L., Nolan, G. P., and Wong, W. H. (2017). Scalable multi-sample single-cell data analysis by Partition-Assisted Clustering and Multiple Alignments of Networks. *PLoS Computational Biology*, 13(12):1–37.

[36] Diggins, K. E., Greenplate, A. R., Leelatian, N., Wogsland, C. E., and Irish, J. M. (2017). Characterizing cell subsets using marker enrichment modeling. *Nature Methods*, 14(3):275–278.

[37] Spidlen, J., Breuer, K., Rosenberg, C., Kotecha, N., and Brinkman, R. R. (2012). FlowRepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytometry Part A*, 81A:727–731.