OXFORD

Briefings in Bioinformatics, 2022, **23**, 1–12
https://doi.org/10.1093/bib/bbab471
Problem Solving Protocol

A systematic comparison of novel and existing differential analysis methods for CyTOF data

Lis Arend†, Judith Bernett†, Quirin Manz†, Melissa Klug, Olga Lazareva, Jan Baumbach, Dario Bongiovanni and Markus List

Corresponding author: Markus List, Chair of Experimental Bioinformatics, Technical University of Munich, Maximus-von-Imhof-Forum 3, Freising 85354, Germany. Tel: +49-8161-71-2761; E-mail: markus.list@wzw.tum.de
†These authors contributed equally to this work.

Abstract

Cytometry techniques are widely used to discover cellular characteristics at single-cell resolution. Many data analysis methods for cytometry data focus solely on identifying subpopulations via clustering and testing for differential cell abundance. For differential expression analysis of markers between conditions, only few tools exist. These tools either reduce the data distribution to medians, discarding valuable information, or have underlying assumptions that may not hold for all expression patterns. Here, we systematically evaluated existing and novel approaches for differential expression analysis on real and simulated CyTOF data. We found that methods using median marker expressions compute fast and reliable results when the data are not strongly zero-inflated. Methods using all data detect changes in strongly zero-inflated markers, but partially suffer from overprediction or cannot handle big datasets. We present a new method, CyEMD, based on calculating the earth mover's distance between expression distributions that can handle strong zero-inflation without being too sensitive. Additionally, we developed CYANUS – CYtometry ANalysis Using Shiny – a user-friendly R Shiny App allowing the user to analyze cytometry data with state-of-the-art tools, including well-performing methods from our comparison. A public web interface is available at https://exbio.wzw.tum.de/cyanus/.

Keywords: cytometry, CyTOF, differential expression analysis, benchmark

Introduction

High-dimensional time-of-flight mass cytometry (CyTOF) is a powerful tool to unveil new cell subtypes, functions and biomarkers in many fields, e.g. the discovery of disease-associated immunologic changes in cancer [1]. Cytometry experiments rely on a panel of antibodies that are associated with a specific experimental condition or phenotype of interest. The analysis of cytometry data starts by clustering cells into cell subpopulations using type markers, followed by a differential expression analysis between and within clusters [2]. Several methods have been developed for testing clusters representing cell populations for differential abundance (DA) between conditions [3–5]. However, many experiments aim to detect differential states (DS) using state markers, i.e. differential expression of markers between conditions and within cell populations (see Figure 1A). Clustering and differential expression can be considered complementary, e.g. state markers can be differentially expressed in specific clusters but not overall and thus differential expression analysis is essential to characterize clustering results.

Diffcyt [5] presents two methods for differential expression detection, a linear mixed effect model (LMM) and an adaptation of limma [6]. For both approaches, the data are reduced to median marker expressions per sample and per cluster when comparing conditions. Alternatively, CytoGLMM [7] uses (bootstrapped) generalized linear models of expression values across markers.

Methods that rely solely on median marker expression are oblivious to biases, whereas a comparison of hundreds of thousands of cells per patient is computationally infeasible. The optimal solution may thus depend on the properties of the data. Here, we assess statistical tests relying on medians, logistic regression, two techniques modeling the expression distributions and a method using the earth mover's distance (EMD; see Figure 1B) and evaluate them on semi-simulated, simulated and real datasets resembling several experimental scenarios: globally visible differences in various magnitudes, patient-specific effects on paired data,

Lis Arend, Judith Bernett, and Quirin Manz are Bioinformatics (MSc) students at the Technical University of Munich and the Ludwig-Maximilians Universität.
Melissa Klug is a PhD candidate at the Technical University of Munich.
Olga Lazareva is doctoral fellow at the Bavarian Research Institute for Digital Transformation and a PhD candidate at the Technical University of Munich.
Jan Baumbach is professor and chair of Computational Systems Biology at the University of Hamburg. He obtained his PhD in Computer Science from Bielefeld University.
Dario Bongiovanni is a clinician scientist and postdoc at University hospital rechts der ISAR, Munich, Germany.
Markus List obtained his PhD at the University of Southern Denmark and worked as a postdoctoral fellow at the Max Planck Institute for Informatics before starting his group Big Data in BioMedicine at the Technical University of Munich.
Received: August 24, 2021. **Revised:** September 30, 2021. **Accepted:** October 13, 2021

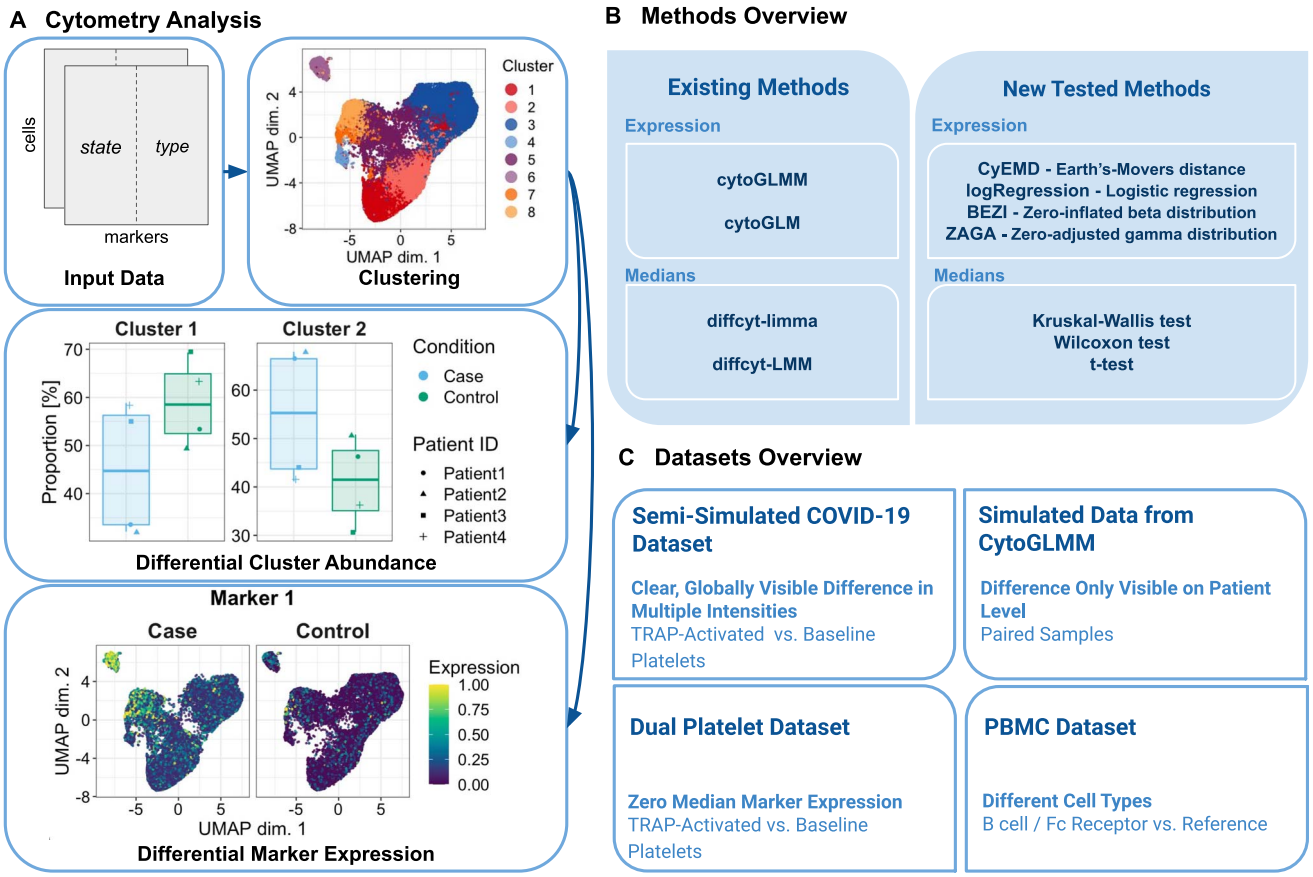

Figure 1. (**A**) Schematic overview of a differential analysis workflow for cytometry data. In a cytometry experiment, the abundance of state (condition) and type (lineage) markers are measured for each cell. Usually, cells are clustered using type markers to identify cell subpopulations. When differential cluster abundance is analyzed, the proportion of cell types between conditions is compared (e.g. condition 2 stimulates the production of cell subpopulation 1). When differential marker expression is analyzed, marker expression is compared between conditions within each cluster (e.g. Marker 1 is more highly expressed in condition 1 in clusters 6 and 8). (**B**) Overview of the methods compared in this study. (**C**) Overview of the datasets used in this study. One simulated, one semi-simulated and two real CyTOF datasets were used to evaluate the methods.

highly zero-inflated marker expressions and an immune dataset composed of multiple cell types (see Figure 1C).

In addition, we present CYtometry ANalysis Using Shiny (CYANUS), a user-friendly R Shiny App available at https://exbio.wzw.tum.de/cyanus/. In contrast to existing cytometry analysis platforms like Cytobank [8] or OMIQ [9], we provide an open-source platform allowing researchers to analyze normalized, gated cytometry data. To this end, we integrated state-of-the-art methods from CATALYST [10] for preprocessing, visualization and clustering. We further integrated methods for differential marker expression and abundance analysis showing good performance in our benchmark.

Methods

Data description

For the evaluation of the differential expression methods, we worked with four different datasets. The methods were tested on one semi-simulated, one simulated and two real CyTOF datasets (Figure 1C).

Semi-simulated COVID-19 data

The semi-simulated COVID-19 platelet dataset is derived from [11] and comprises CyTOF data of 8 symptomatic SARS-CoV-2-infected patients and 11 healthy donors. A baseline sample (non-stimulated platelets) and one sample stimulated with thrombin receptor-activating peptide (TRAP) were prepared for each donor. To study the sensitivity of the methods to changes in the expression patterns, baseline healthy samples were randomly split in half. Half of a sample was used for randomly spiking in the expression values for the four known activation markers (CD62P, CD63, CD107a CD154) from the activated sample of the corresponding patient. Because this leads to very clear, well distinguishable results, we reduced the differences in expression between baseline and spike expressions for the four markers using the following formula:

$$c_{m,x_i} := 5\sinh\left[\operatorname{asinh}\left(\frac{c_{m,y_i}}{5}\right) - \alpha\left(\operatorname{asinh}\left(\frac{c_{m,y_i}}{5}\right) - \operatorname{asinh}\left(\frac{c_{m,x_i}}{5}\right)\right)\right] \quad (1)$$

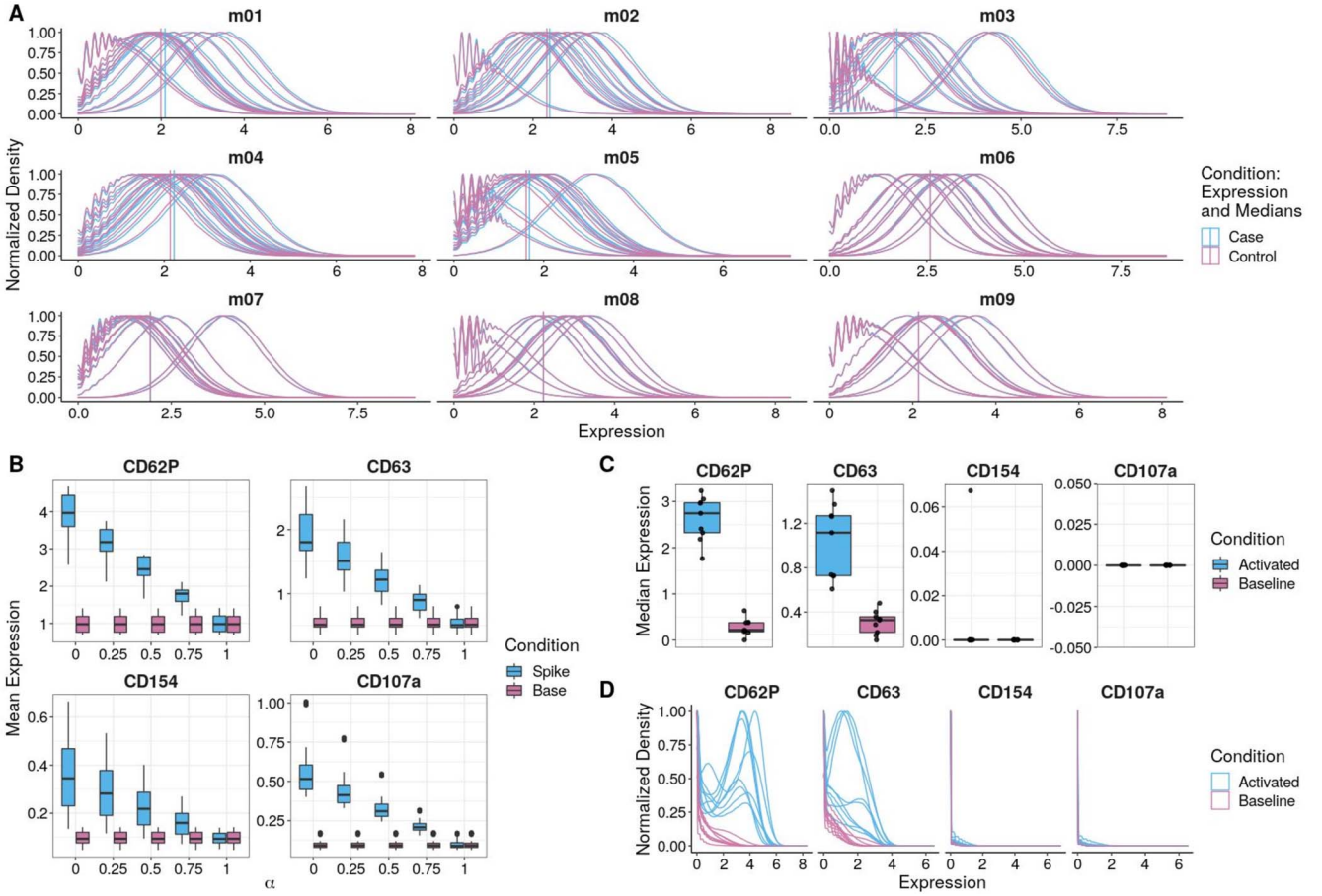

Figure 2. Marker expressions of the simulated CytoGLMM (**A**), semi-simulated COVID-19 (**B**) and the dual platelet datasets (**C**, **D**). **(A)** Normalized density of the markers m01-m09 of the dataset simulated using the CytoGLMM data generation process by [7]. The markers m01-m05 are simulated to be differentially expressed in such a way that the expression differs slightly but consistently for each patient. Meanwhile, the median marker expressions of the whole dataset, marked by the vertical lines, do not differ significantly. **(B)** Mean expressions for the four spiked-in activation markers at different intensities. For α=0 (full intensity), the originally measured expressions of the corresponding activated sample were used. Subsequently, α was repeatedly increased by 0.25 in order to reduce the difference between the spiked and the base condition so that the differences would become harder to detect. α=1 was used as control dataset. **(C)** Median expression of state markers of the dual platelet dataset. Markers CD62P and CD63 are higher expressed in the activated condition. The median marker expression of CD107a and CD154 is zero, except for one sample. **(D)** Normalized density of state markers of the dual platelet dataset. CD107a and CD154 show a small difference in the expression.

where m is the marker, c_{m,x_i} is the raw value measured for the baseline sample for cell x_i, c_{m,y_j} is the raw value measured for the activated sample for cell y_j, $X = x_1, \ldots, x_{N/2}$ are the indices of the baseline cells whose expression was randomly replaced and $Y = y_1, \ldots, y_{N/2}$ are the indices of the activated cells whose values were used for spiking with α 0 (full intensity), 0.25, 0.5, 0.75 and 1.0 (control) (see Figure 2B). Each dataset contains 11 paired samples with 4 052 622 cells in total (see Supplementary Table 7 for the number of cells per sample). This approach was inspired by the diffcyt benchmarking strategy [5]. In contrast to their approach, we did not use differences in means and standard deviations between the two conditions for reducing the signal but the actual differences between c_{m,x_i} and c_{m,y_j}.

Simulated CytoGLMM data

To investigate the sensitivity to paired differences in expression, we used a customized version of the data simulation process described by [7]. The algorithm samples from a Poisson generalized linear model (GLM) with an underlying hierarchical model combining effects on cell and donor-level for two conditions (Figure 2A). We simulated 20 markers, of which five are differentially expressed, in 22 paired samples from 11 patients with 200 000 cells per sample.

Dual platelet data

We used a CyTOF platelet dataset originating from the University Hospital rechts der Isar, Munich, Germany, consisting of platelet heterogeneity measurements of patients with chronic coronary syndrome receiving dual anti-thrombotic therapy. The dataset contains 4 491 504 cells and includes 18 paired samples from 9 donors in two conditions: non-stimulated and stimulated (TRAP). For the exact number of cells per sample, refer to Supplementary Table 8. The panel containing 22 protein markers (see Supplementary Table 9) includes

four well-known platelet activation markers [12]. Two of the platelet activation markers, CD63 and CD62P, are known to be highly upregulated after TRAP stimulation, whereas CD107a and CD154 are upregulated less strongly (see Figures 2C and D).

Peripheral blood mononuclear cell data

The peripheral blood mononuclear cells (PBMCs) dataset originating from [13] consists of samples from 8 healthy donors in 12 conditions. Nowicka et al. [2] performed a complete CyTOF analysis on a subset of this data containing the reference and one stimulated condition. In the stimulated condition, the cells were cross-linked with B cell receptor/Fc receptor for 30 min. This subset consists of 172 791 cells in 16 paired samples from 8 patients (see Supplementary Table 10). Nowicka et al. [2] manually merged 20 clusters obtained via meta clustering into 8 cell populations which were made publicly available. In this study, this annotated and well-described subset was used.

Downsampling of artificial datasets

For sensitivity and runtime analysis, the spiked COVID-19 and the simulated CytoGLMM data were subsampled to 1000, 2000, 5000, 10 000, 15 000 and 20 000 cells per patient such that the smaller sets are always subsets of the bigger ones. The same cells were used in the COVID-19 dataset for different α values to ensure a fair comparison.

Effect size

To quantify the difference between marker expressions, we computed *Cohen's d* [14] for each marker in every dataset using the `rstatix` R package [15]. The effect size was calculated overall (on the whole expression) and grouped (based on the median marker expression of the paired samples). The overall effect size compares marker intensities between two conditions by using their mean and (shared) standard deviation:

$$d = \frac{\mu_1 - \mu_2}{\sigma} \tag{2}$$

Patient-level differences can be captured with a paired effect-size estimation (grouped effect size between the medians of each sample) and can be tested for significance with a paired *t*-test. We used Hedges' correction to adjust for the small sample size.

Differential analysis

The following differential expression detection methods were tested on the datasets mentioned above. Constructing the ground truth in the (semi-)simulated datasets allowed us to evaluate the methods using sensitivity, specificity and the F1-score. Furthermore, there are well-studied activation markers in platelets, allowing us to assess the methods on real biological data using the dual platelet dataset. Finally, we included the PBMC dataset which contains multiple cell types and was previously used as a benchmark CyTOF dataset [2, 5].

Diffcyt methods

The diffcyt-limma method fits a linear model for each marker-cluster combination, predicting the sample medians from the conditions. The LMM method builds a linear mixed-effects model and can therefore handle random effects in contrast to the limma method where a grouping variable can be included only as an additional fixed effect [5].

CytoGLMM methods

The CytoGLMM methods fit a generalized mixed model predicting the conditions from the whole expression vectors. The package contains two methods, CytoGLMM and CytoGLM. The former can only handle grouped data since it relies on a random effect like patient ID whereas the latter can also handle unpaired data. CytoGLM builds a bootstrapped generalized linear model, whereas CytoGLMM builds a generalized linear mixed model [7]. We used 500 bootstrap replications.

Statistical tests on medians

To test whether the diffcyt approach could be simplified, we included a *t*-test and two non-parametric statistical tests on marker medians: the Wilcoxon rank-sum/signed-rank test and the Kruskal–Wallis test.

Logistic regression

As a simpler alternative to CytoGLMM, we fitted a univariate logistic regression models per marker and cluster. A multivariate approach was omitted since the markers are not statistically independent. CytoGLMM partially evades this problem by fitting a hierarchical model containing random slopes and intercepts for the grouping variable (patient ID) which assumes dependent errors.

Approaches modeling the expression: BEZI and ZAGA

As CyTOF data can be strongly zero-inflated [16], we fit a zero-adjusted gamma distribution (ZAGA) [17]. A common choice for single-cell RNA-seq data is the negative binomial distribution [18], which is not suitable for CyTOF data as it requires discrete values. We use a zero-inflated beta distribution (BEZI) as a conjugate to negative binomial distribution. We consider the condition as explanatory variable and tested model coefficients for equality to zero via the gamlss and the gamlss.dist packages [19, 20]. To model changes on a patient level for paired data, random intercepts were included.

A model-free approach: CyEMD

The EMD is a distance metric for comparing normalized density histograms. It has previously been shown, that the EMD provides reliable and robust results in single-cell RNA-seq differential expression analysis [21]. CyEMD uses the EMD to compare differences in cytometry expressions between groups, either overall or cluster-wise. Each marker expression profile is represented as a

normalized distribution histogram. Since the expression densities in CyTOF data can have different ranges for distinct values, we use a flexible bin width estimated by the Freedman-Diaconis rule. Significance is determined via permutation test (500 permutations). We permute the condition labels sample-wise to obtain a *P*-value for each marker. More information on the formal definition of the EMD can be found in the Supplementary Methods.

Multiple testing correction

All *P*-values have been adjusted per method and dataset using the Benjamini-Hochberg method at $\alpha < 0.05$.
For more detailed explanations on all methods, we refer to Supplementary Methods.

Results

In this work, we compared existing approaches for differential marker expression analysis with simple and advanced novel approaches that either rely on the median or on full marker expression data using one semi-simulated, one simulated, and two real datasets.

We hypothesize that when reducing the datasets to their medians as in diffcyt, simple statistical tests such as the Wilcoxon rank-sum/signed-rank test, Kruskal–Wallis, or (paired) *t*-test could be effective. We further use a univariate logistic regression to examine whether the CytoGLMM approach could be simplified. To explore whether using the entire distribution of the dataset is beneficial, we modeled the expression data by fitting a BEZI and a ZAGA, respectively. We further used the EMD to compare normalized distributions for each marker (and cluster) between groups (CyEMD).

Statistical tests may report significant differences that are not meaningful due to their negligible effect size. To account for this, we computed, for all results, the overall (global) and grouped (accounting for patients or other groups) effect size. It should be noted that the grouped effect size must be treated with caution due to the small number of samples (see Methods).

Semi-simulated COVID-19 dataset with clean, globally visible difference between conditions

For the semi-simulated COVID-19 platelet dataset, we created an artificial signal to introduce differential expression of CD63, CD62P, CD107a and CD154 by using the differences between the unstimulated and TRAP-stimulated samples of the original experiment. The four markers whose expression was used for creating the signal, hereinafter referred to as state markers, are known platelet activation markers [12].

All other markers (i.e. type markers) detected by any method can be classified as false positives, since the baseline expression values were not modified.

To examine the sensitivity of the methods, we reduced the differences in expression between the baseline and the spike condition step-wise via a parameter α (Equation 1) which indicates by what percentage the difference between the spiked-in and the baseline expression values is reduced.

The differences are visible on a global level for CD63, CD62P, CD107a and moderate for CD154. Although the other 18 markers have a negligible overall effect size, six show a small and one a moderate grouped effect size. Supplementary Figures 5 and 6 show the results containing all downsampled datasets for activation (state) markers and other (type) markers, respectively.

Table 1 gives an overview of the methods' performance across all COVID-19 datasets measured by the F1-score. Sensitivity, specificity and precision on the same datasets can be found in Supplementary Tables 4, 5 and 6, respectively. The methods relying on the median marker expression tend to perform better with an increasing number of cells. The opposite is the case for both methods from the CytoGLMM package, as well as BEZI, ZAGA and the logistic regression.

The diffcyt methods can find all activation markers regardless of sample size and signal intensity. In the negative controls, both methods find markers in the small downsampled datasets (1000 and 2000 cells per patient).

The Kruskal–Wallis test correctly detects all of the state markers and none of the type markers across all sample sizes and α values. The Wilcoxon signed-rank test misses CD154 for $\alpha = 0$ regardless of the sample size. In the negative controls, the Wilcoxon test and the *t*-test find one type marker for $n = 2000$ and the *t*-test finds one type marker for $n = 1000$. This observation can be made for all α values, except for α=1.

The CytoGLMM methods find many false positive type markers across all α values (except for α=1). The number of false positives rises with increasing sample size. For the downsampled datasets, more type markers are found for higher α values. Additionally, CD154 cannot be detected by both CytoGLMM methods for $\alpha = 0$, as well as by CytoGLM for α= 0.25.

BEZI fails to find different subsets of the state markers across all sample sizes and α values, either due to convergence errors (for datasets bigger than 5000 cells/patient) or because they did not pass the significance threshold of 0.05. Additionally, BEZI classifies PEAR as differentially expressed for all α values in the datasets that were not subsampled. ZAGA and the univariate logistic regression also find PEAR in this dataset but not for α=1. Similar to the CytoGLMM methods, ZAGA fails to find CD154 for smaller datasets when α is set to 0.25.

CyEMD was able to classify all markers correctly.

Simulated data from CytoGLMM package with differences only visible on patient-level

The CytoGLMM simulation leads to patient-wise differences of markers m01–m05 that are only visible considering the grouped rather than the overall effect size (see Supplementary Figure 7).

Table 1. Methods' performance measured by F1 scores on the semi-simulated COVID-19 dataset. The means and standard deviations of the scores are reported across the multiple α values

Number of cells	1000	2000	5000	10 000	15 000	20 000	4 052 622
diffcyt-DS-limma	0.89 +/− 0	0.89 +/− 0	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**
diffcyt-DS-LMM	0.62 +/− 0	0.67 +/− 0	0.73 +/− 0	0.73 +/− 0	0.73 +/− 0	0.73 +/− 0	0.73 +/− 0
t-test	0.89 +/− 0	0.89 +/− 0	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**
Wilcoxon test	0.96 +/− 0.07	0.85 +/− 0.07	0.96 +/− 0.07	0.96 +/− 0.07	0.96 +/− 0.07	0.96 +/− 0.07	0.96 +/− 0.07
Kruskal–Wallis test	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**
CytoGLM	0.79 +/− 0.05	0.69 +/− 0.14	0.5 +/− 0.07	0.47 +/− 0.09	0.48 +/− 0.13	0.48 +/− 0.12	0.4 +/− 0.05
CytoGLMM	0.74 +/− 0.06	0.47 +/− 0.03	0.38 +/− 0.02	0.36 +/− 0.03	0.34 +/− 0.04	0.33 +/− 0.03	0.37 +/− 0.04
logRegression	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	0.89 +/− 0
ZAGA	0.96 +/− 0.07	0.85 +/− 0.07	0.96 +/− 0.07	0.93 +/− 0.08	0.96 +/− 0.07	0.85 +/− 0.07	0.89 +/− 0
BEZI	0.93 +/− 0.08	0.96 +/− 0.07	0.96 +/− 0.07	0.88 +/− 0.16	0.96 +/− 0.07	0.58 +/− 0.06	0.28 +/− 0.09
CyEMD	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**

Highest values per dataset are shown in bold.

Table 2. Methods' performance on the simulated CytoGLMM dataset. Sensitivity, specificity, precision and F1 score are shown for each method. Means and standard deviations of the scores are reported across the multiple numbers of cells. If no positive classification was made, precision and F1 score cannot be computed and are marked as NaN in the table

	Sensitivity	Specificity	Precision	F1 score
diffcyt-DS-limma	0.97 +/− 0.08	0.99 +/− 0.03	0.98 +/− 0.06	0.97 +/− 0.05
diffcyt-DS-LMM	**1 +/− 0**	0.96 +/− 0.04	0.9 +/− 0.09	0.95 +/− 0.05
t-test	0.97 +/− 0.08	**1 +/− 0**	**1 +/− 0**	0.98 +/− 0.04
Wilcoxon test	0.49 +/− 0.34	**1 +/− 0**	**1 +/− 0**	0.81 +/− 0.08
Kruskal–Wallis test	0 +/− 0	**1 +/− 0**	NaN	NaN
CytoGLM	0.8 +/− 0.38	**1 +/− 0**	**1 +/− 0**	0.96 +/− 0.1
CytoGLMM	0.97 +/− 0.08	0.97 +/− 0.05	0.94 +/− 0.12	0.95 +/− 0.07
logRegression	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**	**1 +/− 0**
ZAGA	0.91 +/− 0.16	0.91 +/− 0.12	0.83 +/− 0.19	0.85 +/− 0.14
BEZI	0.86 +/− 0.15	0.93 +/− 0.07	0.84 +/− 0.16	0.83 +/− 0.1
CyEMD	0 +/− 0	**1 +/− 0**	NaN	NaN

Highest values per performance measurement are shown in bold.

Table 2 shows an overview of performance measurements on all subsets of this dataset. For more detailed results, we refer to Supplementary Figure 7.

The two methods that cannot perform a paired analysis, CyEMD and the Kruskal–Wallis test on marker expression medians, do not find any marker to be differentially expressed.

The diffcyt methods have a high performance and gain power for greater numbers of cells. This effect can also be observed for most other methods. CytoGLMM's and CytoGLM's scores are close to 1 except for the sensitivity scores for CytoGLM, which vary more strongly since some of the differentially expressed markers cannot be detected for low cell counts. BEZI and ZAGA lose performance mostly because the algorithms do not converge. Apart from that, they yield high scores. Only the univariate logistic regression can correctly identify all differentially expressed markers without a false positive discovery.

Dual platelet dataset with zero median marker expression

This dataset was generated by collecting two samples from each participant and stimulating one of the two samples with TRAP to activate the platelets. Therefore, we expected to find platelet activation (state) markers like CD63, CD62P, CD154 and CD107a to be differentially expressed between the two conditions. Figure 2C shows that CD154 and CD107a have a median marker expression of zero, posing a challenge for the methods using only marker medians.

We tested our methods twice on this dataset. For the first run, the patient ID was included as a grouping variable, whereas the second analysis was unpaired (see Figure 3). We used the Wilcoxon rank-sum test and the Wilcoxon signed-rank test in the unpaired and paired design, respectively.

ZAGA, BEZI and the univariate logistic regression classify all markers as significant or do not converge. The issues of these three methods are examined thoroughly in the discussion.

The five algorithms (diffcyt-limma, diffcyt-LMM, *t*-test, Wilcoxon test and Kruskal–Wallis test) using only median expressions find the two state markers CD63 and CD62P but are not able to find the zero-inflated markers CD154 and CD107a. In the unpaired run, no type markers are found by these methods. In the analysis with patient ID as grouping variable, the Wilcoxon signed-rank test, *t*-test and both diffcyt methods find PAR1, PEAR and CD69 to be significantly differentially expressed between the non-stimulated and stimulated samples. CD42a was also found by the *t*-test and both diffcyt methods. Each of the four markers has a large grouped effect size.

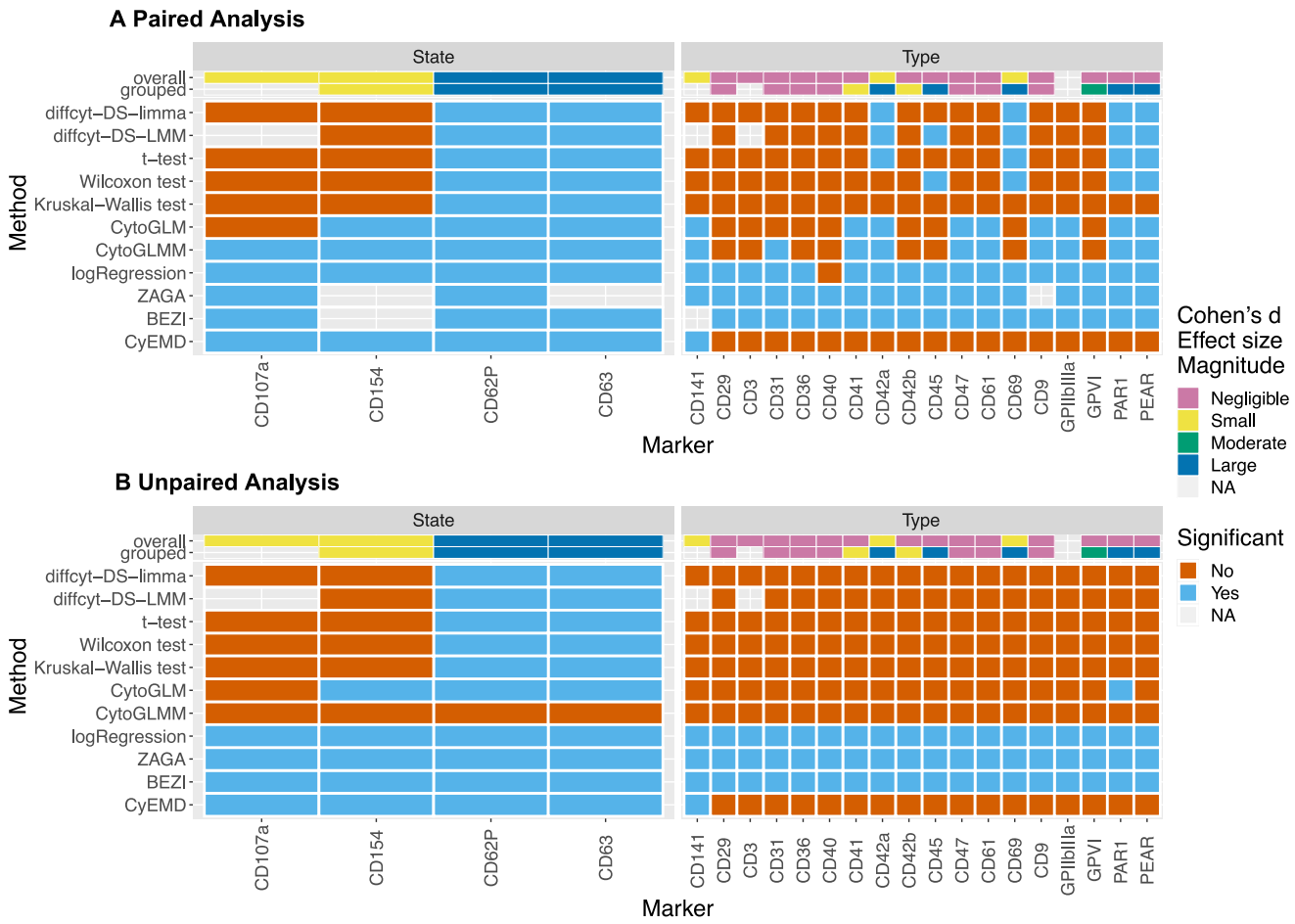

Figure 3. Method results for the dual dataset with patient id as grouping variable (**A**) and without any grouping variable (**B**). Results colored in blue if the adjusted *P*-value < 0.05, else in red. Uncolored tiles mean convergence errors of the method for the specific marker. The overall and grouped effect size magnitudes per marker are shown at the top. Overall effect size refers to *Cohen's d* magnitudes using all expression data between two conditions. The magnitudes indicated by grouped effect size are computed in a paired fashion on the median marker expressions per sample. Wilcoxon test refers to the Wilcoxon signed-rank test and the Wilcoxon rank-sum test for the paired and the unpaired analysis, respectively. Markers are divided into their marker class (state and type).

Two of the three methods using whole marker expression, CyEMD and CytoGLMM, classify all four state markers as significant. CytoGLM only misses CD107a which has a small overall effect size. Since CytoGLMM cannot be run without a random effect, its result for this run is not reliable. Looking at the results for the type markers, CyEMD finds CD141 and CytoGLM finds PAR1 when no paired analysis is performed. After including the patient ID as grouping variable, additional markers are found by all methods able to incorporate this information. CytoGLMM, CytoGLM and CyEMD all detect CD141 (small overall effect size). The two methods of the CytoGLMM package find several additional markers: CD41, CD61, PAR1, GPIIbIIIa, CD141, CD9, PEAR, CD47, CD31 and CD42a. Although PAR1, PEAR and CD42a are also found by other methods (as mentioned above), some of these markers (CD61, CD47, CD9 and CD31) have negligible effect sizes which is why we classify them as false positives (see Supplementary Figure 8).

PBMC dataset with different cell types

Since our first real dataset, the dual platelets dataset, only contains one cell type, we also evaluated the different approaches on the PBMC dataset by Bodenmiller et al. [13], which contains eight immune cell types annotated by Nowicka et al [2]. For each cluster of cell types, we compared the reference condition against cells that were cross-linked with B cell receptor/Fc receptor (BCR/FcR-XL). We expected to find pS6 differentially expressed as reported by [13] (see Supplementary Figure 9).

Many markers were significant across all clusters. Independent of the method, overall and grouped effect sizes were large for numerous markers in all clusters (see Supplementary Figure 9).

Of all possible 192 marker-cluster combinations (24 markers in 8 cell types), the univariate logistic regression, BEZI and ZAGA find the most markers to be differentially expressed (168, 156 and 155, respectively). The methods that are not able to include a grouping variable, CyEMD

Table 3. Runtime of the methods on the different datasets. The methods that reduce the data to medians and CytoGLMM have very low runtime requirements, whereas CytoGLM and BEZI are slow on big datasets. The univariate logistic regression, CyEMD and ZAGA have moderate runtimes

	Semi-simulated COVID-19	Simulated CytoGLMM	Paired dual platelets	Unpaired dual platelets	PBMC
Number of cells	4 052 622	4 400 000	4 491 504	4 491 504	906 815
diffcyt-DS-LMM	**26 +/− 2 s**	**29 s**	**35 s**	**33 s**	**5 s**
diffcyt-DS-limma	29 +/− 5 s	38 s	47 s	41 s	**5 s**
t-test	1.03 +/− 0.06 min	1.06 min	1.03 min	1.09 min	28 s
Kruskal–Wallis test	1.04 +/− 0.09 min	1.01 min	1.08 min	1.04 min	31 s
Wilcoxon test	1.04 +/− 0.06 min	1.07 min	1.11 min	1.07 min	29 s
CytoGLMM	1.92 +/− 0.41 min	1.18 min	2.02 min	7.82 min	11 s
CyEMD	1.9 +/− 0.2 h	2.1 h	2.0 h	1.9 h	6.66 min
logRegression	2.5 +/− 0.2 h	2.2 h	2.6 h	3.62 min	49.73 min
ZAGA	2.7 +/− 0.4 h	2.3 h	2.1 h	49.53 min	5.9 min
CytoGLM	6.5 +/− 1.6 h	4.0 h	6.5 h	7.8 h	15.92 min
BEZI	9.8 +/− 1.6 h	9.7 h	9.7 h	4.7 h	26.13 min

Lowest runtimes per dataset are shown in bold.

and the Kruskal–Wallis test, find the least markers to be differentially expressed (86 and 88, respectively). In contrast to the dual dataset results, CytoGLMM and CytoGLM do not produce more positive predictions than the statistical tests, CyEMD or the diffcyt methods.

Runtime

The runtimes for the complete datasets are shown in Table 3. For the runtimes of the subsampled datasets, please refer to Supplementary Figure 9.

The diffcyt methods outperform all other methods in terms of runtime. Methods that use median marker expressions are fast, independent of sample size. CytoGLMM and the unpaired logistic regression are quick as well, even though they take the whole distribution into account.

The paired univariate logistic regression, CyEMD and ZAGA have moderate runtimes, whereas CytoGLM and BEZI often run more than 6 h on big datasets.

Discussion

Semi-simulated COVID-19 dataset with clean, globally visible difference between conditions

Seven of the markers that were not spiked in appear to have a small or even moderate grouped effect size owing to the small standard deviation (see Supplementary Equation 5). We recommend a paired *t*-test, which reveals that these differences are not significant.

The diffcyt methods and the statistical tests perform well, especially for larger sample sizes. We hypothesize that the markers that are found in the small negative control datasets were detected because of noise in the measured data. Due to the law of large numbers, the median becomes more reliable for higher cell counts. Therefore, methods that reduce the expression data to medians become more stable with growing dataset size.

The CytoGLMM methods produce a high number of false positives for all α values except for $\alpha = 1$, especially with rising sample size. A possible explanation could be that the multivariate generalized mixed effect models become too sensitive to small changes when there are only few bigger differences (here, CD62P and CD63) because all markers are included as explanatory variables. Therefore, the condition is modeled as a result of various small changes which are present because of the semi-simulated nature of the data. The increasing sample size seems to reduce the magnitude of the *P*-values.

The Wilcoxon signed-rank test and the CytoGLMM methods miss CD154 for $\alpha = 0$ and $\alpha = 0.25$ but find it for the other α values, since the medians of the spike condition are higher for $\alpha = 0.25$, 0.5 and 0.75 than for $\alpha = 0$ in two patients (see Supplementary Figure 11).

BEZI shows high sensitivity for large sample sizes (see dual dataset). ZAGA and the univariate logistic regression yield reliable results for datasets with a clean, globally visible difference, especially for smaller sample sizes.

Simulated data from CytoGLMM package with differences only visible on patient-level

Because these data are paired, we expect that only methods that can handle paired data can detect the differentially expressed markers between the two conditions. This is confirmed as all methods except for CyEMD and the unpaired Kruskal–Wallis test detect the differential expression and can be sensitive to small changes in expression that are only detectable at the patient level.

CytoGLMM's false detection of one marker suggests an over-sensitivity further described in the next section.

Dual platelet dataset with zero median marker expression

The results for this dataset clearly show the problem of reducing the data on median marker expressions to perform differential expression analysis. Methods taking the whole marker expression into account find markers with zero-median marker expression, whereas methods working on the medians are not able to find these.

The markers PAR1 and PEAR are detected by several methods. Although PEAR has a higher expression in the

stimulated condition, PAR1 is less expressed in this condition (see Supplementary Figure 12). In literature, the PEAR receptor has been described to be increased on the platelet membrane after stimulation with several activators [22], whereas the effect on PAR1 expression after stimulation depends on the agonist. Studies using a PAR1-AP are in line with our findings and show a decreased amount of the PAR1 receptor on the platelet surface after stimulation [23].

CD69, which is found by limma, LMM, the Wilcoxon test and the *t*-test, shows a higher signal after stimulation. Several studies have observed a similar trend for CD69 increase upon stimulation [24, 25]. CD42a is detected by the two diffcyt methods, the CytoGLM/M methods, and the *t*-test and shows a decreasing trend after TRAP stimulation. This also has been previously shown in platelets using CyTOF [12]. Several other studies examined a decrease of CD42a expression after stimulation with activators adenosine-5'-diphosphate (ADP) [26] and collagen [27]. The biological reason behind the differential expression of the two markers CD141 and CD45 remains unclear. In general, CD141 is not found to be expressed on platelets [11], whereas CD45 has shown to be present on the surface of several platelets [28].

The application of ZAGA, BEZI and the univariate logistic regression is unfeasible for a real dataset of this size. CytoGLMM and CytoGLM produce at least three false positives due to their high sensitivity. Additionally, CytoGLM misses one of the two highly zero-inflated activation markers. The diffcyt methods, the *t*-test and the Wilcoxon signed-rank test perform fast and yield reliable results but miss the two activation markers that have a median of zero. The Kruskal–Wallis test performs worse than the Wilcoxon signed-rank test on this dataset because it is not able to handle paired data and could therefore not detect markers like PAR1, PEAR, CD69 or CD42a. Lastly, CyEMD detects the globally visible changes for the activation markers and CD141 but fails to detect any of the changes that can only be seen on the patient level as seen in Figure 3.

PBMC dataset with different cell types

The evaluation of this dataset is limited by the number of cells per sample and cluster (see Supplementary Figure 13). For cell types with less than 1000 cells per sample, noise is distorting the analysis.

When Weber et al. [5] evaluated their diffcyt methods on this dataset, they could confirm that pS6 is differentially expressed in B-cells, which is shown by all tested methods (Supplementary Figure 9). The diffcyt methods also identified pS6 as differentially expressed in other cell types [2], which was confirmed by all methods in all cell types except for dendritic cells.

In contrast to the dual platelet dataset, the univariate logistic regression, BEZI and ZAGA were not suffering from a clear over-identification of markers. Because the PBMC dataset is rather small (172 791 cells in total versus 4 491 504 in the dual platelet dataset), we hypothesize that the higher the number of cells, the less suitable these three methods become. This is due to the influence of large sample sizes on the magnitude of the *P*-values [29].

Compared to the dual platelet dataset, the CytoGLM/M methods did not identify more markers as significantly differentially expressed than the other methods, even though there are more markers with a large effect size.

Conclusion and outlook

Existing approaches for differential marker expression analysis were compared with simple and advanced novel approaches that rely either on median or on full marker expression data using two real, one semi-simulated and one simulated dataset.

We could not clearly interpret the results obtained on the PBMC dataset and further work is needed to understand which of the markers show biologically relevant differences in expression. Further, we did not assess robustness against batch effects, which, in principle, can be incorporated as a random effect or additional term in all methods using linear models. With the exception of CytoGLMM, differential expression analysis methods focus on individual markers and do not take interactions between the markers into account which could be further explored in future method development.

All in all, the diffcyt methods perform fast and yield good, trustworthy results when the median of the differentially expressed marker is not zero. Nevertheless, they did not outperform a simple, Wilcoxon signed-rank test or *t*-test on the medians. A clear advantage of the Wilcoxon/*t*-test over the Kruskal–Wallis test is the ability to compute a paired test statistic to reveal group-specific effects.

Regarding the cytoGLMM methods, we observe that small, individual changes can be detected as well as globally visible changes on very clean data, even when it is strongly zero-inflated. Additionally, cytoGLMM is fast, considering it takes the whole distribution into account. On the other hand, these methods classify many markers as differentially expressed, especially with growing dataset size. Therefore, we recommend checking for overlaps between cytoGLMM and other methods, making diagnostic plots and looking at the effect size magnitude when running cytoGLMM on larger, real datasets.

BEZI, ZAGA and the univariate logistic regression proved to be infeasible for larger, real datasets. Although the performance on the completely artificial CytoGLMM dataset was acceptable, the method performance dropped for the semi-simulated spike dataset and eventually only produced positive predictions on the real, dual platelets dataset. Additionally, BEZI is unacceptably slow.

Finally, our novel method CyEMD exploits the advantages of taking the whole marker expressions into

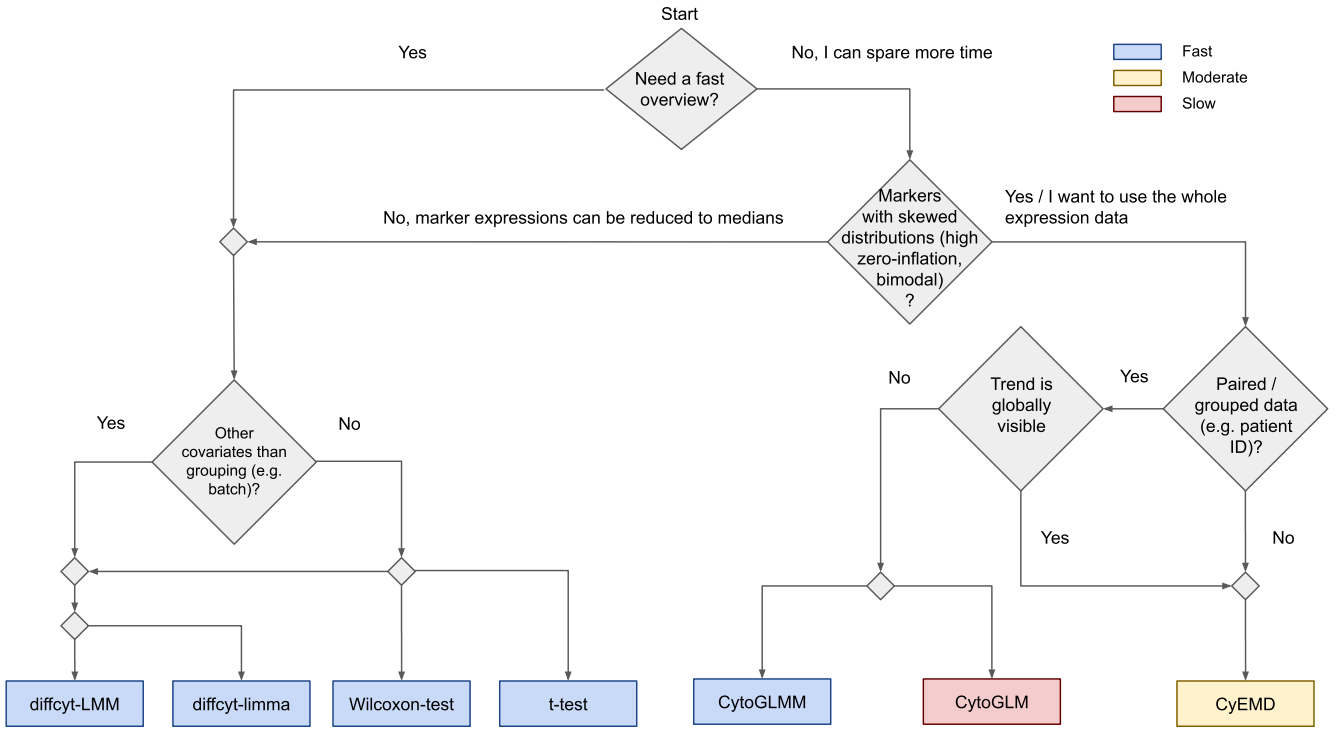

Figure 4. Overview of the methods suitable for CyTOF data. Several scenarios can occur while analyzing CyTOF data. This graph helps to identify the most suitable method and includes the runtime of the different methods.

account and still performs well on big datasets because it partitions the distribution into bins and computes *P*-values via permutation tests. We showed that the EMD approach can detect differentially expressed markers that are strongly zero-inflated in an acceptable amount of time. Additionally, the approach should be able to find differences in bimodal or skewed marker expressions, even when the medians are similar. A disadvantage to the EMD approach is that it cannot detect differentially expressed markers when the changes are only visible by comparing expressions group- or patient-wise.

Our results across datasets with different properties show that each of the tested methods comes with its own strengths and weaknesses. Taking factors like runtime, skewed distributions and sample groups into account, we offer a guideline for users to choose optimal methods for their analysis (Figure 4). However, often several methods are suitable for a given scenario and should be compared to obtain robust and interpretable results.

To make such a comparative analysis easily accessible, we integrated the diffcyt methods, the Wilcoxon rank-sum and signed-rank test, the *t*-test, the cytoGLMM methods and CyEMD into a user-friendly R Shiny App CYANUS available at https://exbio.wzw.tum.de/cyanus/. CYANUS allows the user to analyze gated and normalized cytometry data (i.e. flow cytometry as well as CyTOF) with state-of-the-art methods from CATALYST [10]. For DA analysis, we integrated the methods included in the diffcyt package. All differential analysis methods can be easily compared to each other, enabling thorough analysis of cytometry data exploiting the advantages of the various approaches.

Key Points

- A systematic comparison of differential expression methods for cytometry data is currently missing.
- We compared existing and novel approaches for differential marker expression analysis using simulated and real data sets.
- The choice of the optimal methods depends on the properties of the data set and we offer a guideline for method selection w.r.t. runtime, skewedness and sample groups.
- We present CyEMD, a model-free approach using the EMD to handle high zero-inflation without overpredicting.
- We developed CYANUS, a user-friendly web application for analyzing gated, normalized cytometry data.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data Availability

Analysis scripts and code are available at https://github.com/biomedbigdata/cyanus (GPLv3).

The COVID-19 dataset is available at flowrepository.org (FR-FCM-Z4AE). Access to the dual dataset (nine patients) is granted upon request. The original PBMC

dataset is published at www.cytobank.org/nolanlab. We followed the CyTOF workflow by [2] and downloaded the data using `HDCytoData` [30]. The manual cluster annotation of the CyTOF workflow can be downloaded from http://imlspenticton.uzh.ch/robinson_lab/cytofWorkflow/.

Acknowledgments

The authors thank Simona Ursu and Sarah Warth at the Core Facility Cytometry of the Ulm University Medical Facility for their support acquiring both platelet datasets. We thank Marc Rosenbaum, Dries van Hemelen, Gloria Martrus and Mayur Bakshi for excellent technical assistance, and Kilian Kirmes for testing and evaluating our app.

Funding

Contributions by O.L. are funded by the Bavarian State Ministry of Science and the Arts within the framework coordinated by the Bavarian Research Institute for Digital Transformation (bidt, Doctoral Fellow).
This work was supported by the German Center for Cardiovascular Research (DZHK) grant number 81X3600606 to D.B. J.B. and M.L. are grateful for financial support from BMBF grant Sys_CARE [grant number 01ZX1908A] of the Federal German Ministry of Research and Education.

References

1. Gadalla R, Noamani B, MacLeod BL, *et al.* Validation of cytof against flow cytometry for immunological studies and monitoring of human cancer clinical trials. *Front Oncol* 2019; **9**:415.
2. Nowicka M, Krieg C, Crowell HL, *et al.* CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research* May 2019; **6**.
3. Bruggner RV, Bodenmiller B, Dill DL, *et al.* Automated identification of stratifying signatures in cellular subpopulations. *Proc Natl Acad Sci* 2014; **111**(26): E2770–7.
4. Eirini Arvaniti and Manfred Claassen. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nat Commun*, **8**(1): 14825, April 2017.
5. Weber LM, Nowicka M, Soneson C, *et al.* Robinson. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications biology* 2019; **2**(1): 1–11.
6. Ritchie ME, Belinda Phipson DI, Wu YH, *et al.* limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res* 2015; **43**(7): e47–7.
7. Seiler C, Ferreira A-M, Kronstad LM, *et al.* Cytoglmm: conditional differential analysis for flow and mass cytometry experiments. *BMC bioinformatics* 2021; **22**(1): 1–14.
8. Kotecha N, Krutzik PO, Irish JM. Web-based analysis and publication of flow cytometry experiments. *Curr Protoc Cytom* Chapter 10:Unit10.17July 2010.
9. Belkina AC, Ciccolella CO, Anno R, *et al.* Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat Commun* 2019; **10**(1): 1–12.
10. Crowell HL, Zanotelli VRT, Chevrier S, *et al.* CATALYST: Cytometry dATa anALYSis Tools. 2021; R package version 1.14.1.
11. Bongiovanni D, Klug M, Lazareva O, Weidlich S, Biasi M, Ursu S, Warth S, Buske C, Lukas M, Spinner CD, von Scheidt M, Condorelli G, Baumbach J, Laugwitz K-L, List M, and Bernlochner I. SARS-CoV-2 infection is associated with a pro-thrombotic platelet phenotype. *Cell Death Dis*, 2021; **12**(1): 1–10.
12. Blair TA, Michelson AD, Frelinger AL. Mass cytometry reveals distinct platelet subtypes in healthy subjects and novel alterations in surface glycoproteins in glanzmann thrombasthenia. *Sci Rep* July 2018; **8**(1): 1–13.
13. Bodenmiller B, Zunder ER, Finck R, Chen TJ, Savig ES, Bruggner RV, Simonds EF, Bendall SC, Sachs K, Krutzik PO, and Nolan GP. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nat Biotechnol*, 2012; **30**(9): 858–67.
14. Cohen J. *Statistical power analysis for the behavioral sciences*. Academic press, 1977.
15. Kassambara A. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*, 2021, R package version 0.7.0.
16. Papoutsoglou G, Lagani V, Schmidt A, *et al.* Challenges in the multivariate analysis of mass cytometry data: The effect of randomization. *Cytometry A* 2019; **95**(11): 1178–90.
17. deTorrenté L, Zimmerman S, Suzuki M, *et al.* The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data. *BMC bioinformatics* 2020; **21**(21): 1–18.
18. He L, Davila-Velderrain J, Sumida TS, *et al.* Nebula is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Communications biology* 2021; **4**(1): 1–17.
19. Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics* 2005; **54**:507–54.
20. Stasinopoulos M, Rigby R. *gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape*, 2021, R package version 5.3-2.
21. Wang T, Nabavi S. Sigemd: A powerful method for differential gene expression analysis in single-cell rna sequencing data. *Methods* 2018; **145**:25–32.
22. Kauskot A, Di Michele M, Loyen S, Freson K, Verhamme P, and Hoylaerts MF. A novel mechanism of sustained platelet αiibβ3 activation via pear1. *Blood, The Journal of the American Society of Hematology* 2012:**119**(17): 4056–65.
23. Ramström S, Öberg KV, Åkerström F, *et al.* Platelet par1 receptor density-correlation to platelet activation response and changes in exposure after platelet activation. *Thromb Res* 2008; **121**(5): 681–8.
24. Testi R, Pulcinelli F, Frati L, *et al.* Cd69 is expressed on platelets and mediates platelet activation and aggregation. *J Exp Med* 1990; **172**(3): 701–7.
25. Testi R, Pulcinelli FM, Cifone MG, *et al.* Preferential involvement of a phospholipase a2-dependent pathway in cd69-mediated platelet activation. *The Journal of Immunology* 1992; **148**(9): 2867–71.
26. Braune S, Walter M, Lendlein A, *et al.* Changes in platelet morphology and function during 24 hours of storage. *Clin Hemorheol Microcirc* 2014; **58**(1): 159–70.
27. Hagberg IA, Roald HE, Lyberg T. Platelet activation in flowing blood passing growing arterial thrombi. *Arterioscler Thromb Vasc Biol* 1997; **17**(7): 1331–6.
28. Gabbasov Z, Ivanova O, Kogan-Yasny V, *et al.* Activated platelet chemiluminescence and presence of cd45+ platelets in patients with acute myocardial infarction. *Platelets* 2014; **25**(6): 405–8.

29. Lin M, Lucas Jr HC, Shmueli G. Research commentary-too big to fail: large samples and the p-value problem. *Information Systems Research* 2013; **24**(4): 906–17.
30. Weber LM, Soneson C. Hdcytodata: collection of high-dimensional cytometry benchmark datasets in bioconductor object formats. *F1000Research* 2019;8.
31. Hedges LV, Olkin I. Statistical methods for meta-analysis. *Academic press* 1985.
32. Ospina R, Ferrari SLP. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis* 2012; **56**(6): 1609–23.
33. Rigby RA, Stasinopoulos MD, Heller GZ, *et al. Distribution for modelling location, scale, and shape: using GAMLSS in R*. Bocat Raton, Florida: Chapman & Hall/CRC: the R series. CRC Press, 2020.
34. Rubner Y, Tomasi C, Guibas LJ. A metric for distributions with applications to image databases. In: *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*. IEEE, 1998, 59–66.
35. Rubner Y, Tomasi C, Guibas LJ. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 2000; **40**(2): 99–121.
36. Freedman D, Diaconis P. On the histogram as a density estimator: L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie sund verwandte Gebiete* 1981; **57**(4): 453–76.