CytoGLMM: Conditional Differential Analysis for Flow and Mass Cytometry Experiments

Christof Seiler[123], Anne-Maud Ferreira[3], Lisa M. Kronstad[457], Laura J. Simpson[45], Mathieu Le Gars[45], Elena Vendrame[45], Catherine A. Blish[456], and Susan Holmes[3]

[1]Department of Data Science and Knowledge Engineering, Maastricht University
[2]Mathematics Centre Maastricht, Maastricht University
[3]Department of Statistics, Stanford University
[4]Immunology Program, Stanford University School of Medicine
[5]Department of Medicine, Stanford University School of Medicine
[6]Chan Zuckerberg Biohub, San Francisco
[7]Department of Microbiology and Immunology, Midwestern University

December 09, 2020

Abstract

Background: Flow and mass cytometry are important modern immunology tools for measuring expression levels of multiple proteins on single cells. The goal is to better understand the mechanisms of responses on a single cell basis by studying differential expression of proteins. We focus on cell-specific differential analysis and one fixed cell type. In contrast, most current methods learn cell types and perform differential analysis jointly. Our narrower field of application allows us to define a more specific statistical model with easier to control statistical guarantees. **Results:** Differential analysis of marker expressions can be difficult due to marker correlations and inter-individual heterogeneity, particularly for studies of human immunology. We address these challenges with two multiple regression strategies: A bootstrapped generalized linear model and a generalized linear mixed model. On simulated datasets, we compare the robustness towards marker correlations and heterogeneity of both strategies. For paired experiments, we find that both strategies maintain the target false discovery rate under medium correlations and that mixed models are statistically more powerful under the correct model specification. For unpaired experiments, our results indicate that much larger patient sample sizes are required to detect differences. We illustrate the CytoGLMM *R* package and workflow for both strategies on a pregnancy dataset. **Conclusions:** Our approach to find differential proteins in flow and mass cytometry data reduces biases arising from maker correlations and safeguards against false discoveries induced by patient heterogeneity.

1 Introduction

Flow (Saeys, Van Gassen, and Lambrecht 2016) and mass cytometry (Bendall et al. 2011) allow researchers to simultaneously assess expression patterns of a large number of proteins on individual cells, allowing deep interrogation of cellular responses. The goal of such studies is to better understand the mechanisms of responses on a single cell basis by defining protein expression patterns that are associated with a particular stimulus or experimental condition. Finding differentially expressed proteins can help identify how cells function across experimental conditions.

Statistical workflows to analyze data generated by flow and mass cytometry usually begin by clustering cells into both known and novel cell types. Many cluster algorithms are available (Lo et al. 2009; Finak et al. 2009; Qian et al. 2010; Zare et al. 2010; Aghaeepour et al. 2011; Qiu et al. 2011; Ge and Sealfon 2012; Shekhar et al. 2014; Becher et al. 2014; Naim et al. 2014; Meehan et al. 2014; Van Gassen et al. 2015; Sörensen et al. 2015; Levine et al. 2015; Chen et al. 2016; Samusik et al. 2016; Lee et al. 2017; Li et al.

2017; Theorell, Bryceson, and Theorell 2019; Abdelaal et al. 2019) and Weber and Robinson (2016) provide an informative benchmark comparison study of most of these algorithms. The cluster step is followed by a differential expression analysis between and within cell types. The most popular differential analysis tools are: `Citrus` (Bruggner et al. 2014), the `Bioconductor workflow` by Nowicka et al. (2017), `cydar` (Lun, Richard, and Marioni 2017), `CellCnn` (Arvaniti and Claassen 2017), and `diffcyt` (Weber et al. 2019).

We can classify differential analysis methods into marginal regression—analyses that focus on individual markers—and multiple regression—analyses that work on multiple markers simultaneously. The `Bioconductor workflow` by Nowicka et al. (2017), `cydar`, and `diffcyt` are marginal regression methods. The advantage of marginal regression approaches is that they allow for flexible experimental designs. The main disadvantage of this approach is the separate testing for differential expression for each protein—when studying a specific protein marker all the other markers are ignored. Therefore these methods are susceptible to biases induced by marker correlations.

`Citrus` and `CellCnn` are multiple regression methods. The advantage is that they can provide a conditional interpretation of the effect of a protein onto the outcome, and thus reduce the bias coming from marker correlations. The disadvantage is that `Citrus` summarizes protein expressions by taking the median for each cell type which can lead to a decrease in statistical power. The power decrease comes from the reduction in cell sample size from thousands of cells to one cell per sample. On the other hand, `cydar` uses a neural network for which it is currently unclear how to build confidence intervals, derive p-values, and control the number of falsely reported markers.

It is helpful to consider an example to further illustrate the differences between the marginal and the multiple regression method. Consider two intracellular proteins, A and B, that are part of the same signal transduction pathway. Assume that applying a stimulus to A activates B. Further assume that the stimulus does not directly activate B. If we performed separate differential analyses on protein A and B, we would observe differential expressions for both A and B, even though only A had been directly activated. In contrast, a multiple regression method would report A as differentially expressed given B, and B as not differentially expressed given A.

`CytoGLMM` implements multiple regression that accounts for marker correlations without the aforementioned limitations. The main difference between our method and current methods is that we focus on cell-specific differential analysis and one fixed cell type, whereas current methods (`Citrus`, `CellCnn`, `cydar`, and `diffcyt`) learn cell types and perform differential analysis jointly. The narrower field of application allows us to define a more specific statistical model with easier to control statistical guarantees. Only the `Bioconductor workflow` by Nowicka et al. (2017) focuses on specific cell types, but as mentioned before, they employ marginal regression which makes comparison to our multiple regression method difficult—as the two methods have different aims.

We present two versions of multiple regression: (i) A Generalized Linear Model (GLM) for unpaired samples. A GLM is a regression model that allows for a response and error terms that follow different distributions than the normal. (ii) A restricted Generalized Linear Mixed Model (GLMM), which is a GLM that allows for random and fixed effects, for paired samples—when the same donor provides two samples, one for each condition. GLMs and GLMMs are generalizations of least squares to data from the exponential family. In our case, we will use logistic regression to model the experimental condition with a Bernoulli distribution and link it to a linear model of marker expressions with the logit function.

Our models depart from the classic model where the marker expressions are the response variables. In our GLMs, the experimental condition Y is independent of the jth marker expression X_j given the other markers X_{-j} (all makers $X_1, \ldots, X_P$ except the marker X_j) if and only if the jth regression coefficient is zero (for a mathematical proof of this statement see Proposition 2.2 in Candès et al. (2018)). In contrast, the usual marginal regression analysis does not allow for such conditional statements. For instance, it would not allow us to rule out markers that are merely correlated with other makers but are independent of the experimental condition—as illustrated with the example earlier.

In summary, our two main contributions are:

1. We present a conditional differential analysis to avoid biases arising from marker correlations by using

multiple regression instead of marginal regression.

2. We present two multiple regression strategies that work with the unsummarized expression data to maximize statistical power and account for patient heterogeneity to safeguard against false discoveries: (i) GLMs with a patient-level bootstrap, and (ii) GLMMs with a patient-level random effect.

In Section 2, we review the statistical background for GLMs and GLMMs. Section 3 evaluates the statistical properties of both strategies implemented in our *R* package `CytoGLMM` on different simulated datasets, and illustrates the full workflow for real pregnancy data. In Section 4, we discuss our results in terms of biases and confounders.

2 Methods

2.1 Preprocessing

We recommend that marker expressions be corrected for batch effects (Nowicka et al. 2017; Chevrier et al. 2018; Schuyler et al. 2019; Van Gassen et al. 2020; Trussart et al. 2020) and transformed using variance stabilizing transformations to account for heteroskedasticity, for instance with a hyperbolic sine transformation with the cofactor set to 150 for flow cytometry, and 5 for mass cytometry (Bendall et al. 2011). This transformation assumes a two-component model for the measurement error (Rocke and Lorenzato 1995; Huber et al. 2003) where small counts are less noisy than large counts. Intuitively, this corresponds to a noise model with additive and multiplicative noise depending on the magnitude of the marker expression; see (Holmes and Huber 2019) for details.

2.2 Generalized Linear Model (GLM)

The goal of the GLM is to find protein expression patterns that are associated with the condition of interest, such as a response to a stimulus. We set up the GLM to predict the experimental condition from protein marker expressions, thus our experimental conditions are response variables and marker expressions are explanatory variables. The response variable Y_i is a binary variable encoding experimental condition as zero or one. The response variable can be modeled as a Bernoulli random variable with probability π_i for each cell. Then we use the logit link to relate the linear model to binary responses. The linear model predicts the logarithm of the odds of the ith cell being $Y_i = 1$ instead of $Y_i = 0$. The linear model has one coefficient per protein marker $\beta_1, \dots, \beta_P$ and an intercept β_0. If π_i is 0.5 then the cell could have come from either $Y_i = 1$ or $Y_i = 0$ with equal probability. If π_i is either very close to one or zero, then the cell is strongly representative of a cell observed under $Y_i = 1$ or $Y_i = 0$, respectively. We observe the protein marker expressions $\boldsymbol{x}_i$. For each cell we measure P protein markers.

The response probabilities π_i are not observed directly, only $Y_i = y_i$ and $\boldsymbol{x}_i$ are observed. Note that $\boldsymbol{x}_i$ is observed with errors. Here, we make the approximating assumption that the covariates are fixed. Our results will show that this assumption is conservative and introduces a regularization of the estimated coefficients. We estimate π_i from the data using maximum likelihood with the function `glm` in *R*. Our logistic regression model, which is part of a general class of GLMs, can be summarized in the following form:

$$
\begin{aligned}
Y_i &\sim \text{Bernoulli}(\pi_i), \\
\log\left(\frac{\pi_i}{1-\pi_i}\right) &= \boldsymbol{x}_i^T \boldsymbol{\beta}.
\end{aligned}
$$

For likelihood inference, we use the nonparametric bootstrap and resample entire donors with replacement to preserve the cluster structure. At the cell-level, we resample cells with replacement within each donor. We build percentile confidence intervals and compute p-values by inverting the intervals assuming two-sided intervals with equal tails (Efron and Tibshirani 1994). We use the Benjamini-Hochberg (BH) (Benjamini and Hochberg 1995) and Benjamini–Yekutieli (BY) (Benjamini and Yekutieli 2001) procedures to control the False Discovery Rate (FDR). We refer to GLM with BH control as GLM-BH, and with BY control as GLM-BY.

2.3 Generalized Linear Mixed Model (GLMM)

We make additional modeling assumptions by adding a random effect term in the standard logistic regression model to account for the subject effect. The covariates $\boldsymbol{x}_{ij}$ are the same as in the fixed effects GLM, except now we have an additional index j that indicates from which donor the cell was taken. Each cell i maps to a donor j. The additional term $\boldsymbol{u}_j$ represents regression coefficients that vary by donor. The statistical model can be summarized as,

$$Y_{ij} \sim \text{Bernoulli}(\pi_{ij}),$$
$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = x_{ij}^T\boldsymbol{\beta} + x_{ij}^T\boldsymbol{u}_j,$$

with a multivariate normal distribution and covariance matrix $\boldsymbol{\Sigma}$ for the random effect term $\boldsymbol{u}_j$,

$$\boldsymbol{u}_j \,|\, \boldsymbol{\Sigma} \sim \text{Normal}\left(\mathbf{0}, \boldsymbol{\Sigma}\right).$$

Analog to our GLM, we make the approximating assumption that the covariates are fixed.

The mixed effect model is a compromise between two extremes. On the one hand, we could estimate separate regression coefficients for each donor. This corresponds to random effects modeled with a multivariate normal distribution with infinite standard deviations and no constraint on how coefficients are related to each other. On the other hand, we could pool all donors into one group and ignore the donor information. This corresponds to a GLM with no random effects, with no additional variability besides the fixed effect term. A compromise between these two extremes is to estimate the standard deviations of the random effects from data, allowing the regression model to learn from the other donors. Mixed effects procedures are related to empirical Bayes procedures (Weber et al. 2019). The first step of an empirical Bayes procedure would estimate the mean and covariance matrix of the random effect term. The second step would fix the random effect parameters at their estimated values and estimate the fixed effect parameters. In contrast, the mixed effect procedure estimates the parameters of both steps jointly. This is possible for flow and mass cytometry data because of the relatively small number of proteins.

We use the method of moments as implemented in the *R* package `mbest` to estimate the model parameters $\boldsymbol{\beta}$, $\boldsymbol{u}_j$, and $\boldsymbol{\Sigma}$. For likelihood inference, we use the asymptotic theory derived by Perry (2017). The author showed that the sampling distribution of the estimated parameters can be approximated by a normal distribution. We use this mathematical alternative to the bootstrap method to create approximate confidence intervals and p-values. As in the GLM case, we use the Benjamini-Hochberg (BH) and Benjamini–Yekutieli (BY) procedures to control the FDR. We refer to GLMM with BH control as GLMM-BH, and with BY control as GLMM-BY.

3 Results

We first evaluate these procedures for both paired and unpaired samples on simulated datasets. We then test them on a real pregnancy dataset.

3.1 Simulated Datasets

We generate simulated data with both cell and donor level variability. We allow for negative and positive correlations between markers and a wide range of correlation strengths. We simulate different scenarios ranging from weak to strong patient/cell variability. To make sure that we generate positive counts we use a Poisson noise model after transforming the generated expressions to positive real numbers using the exponential function. This is similar to using the log link function for Poisson GLMs. Overall, there are four main parameters: correlation ρ_B and standard deviation σ_B at the cell level, and correlation ρ_U and standard deviation σ_U at the donor level. Additionally, we can regulate the number of cells per sample and the number of donors per dataset. The differential expression signal is induced by shifting the mean vector on the logarithmic scale.

We study the differential expression of three out of 10 markers after simulating exposure of cells to an experimental condition with two levels: stimulated versus unstimulated cells. We consider one underlying data generating mechanisms described by a hierarchical model:

$$\begin{aligned} X_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \log(\lambda_{ij}) &= B_{ij} + U_j \\ B_{ij} &\sim \begin{cases} \text{Normal}(\boldsymbol{\delta}^{(0)}, \boldsymbol{\Sigma}_B) & \text{if } Y_{ij} = 0, \text{ cell unstimulated} \\ \text{Normal}(\boldsymbol{\delta}^{(1)}, \boldsymbol{\Sigma}_B) & \text{if } Y_{ij} = 1, \text{ cell stimulated} \end{cases} \\ U_j &\sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}_U). \end{aligned}$$

The stimulus activates three proteins and induces a difference in marker expression. We define the effect size to be the difference between expected expression levels of stimulated versus unstimulated cells on the log-scale. We choose a set $C = 3$ of active markers. All markers that belong to the active set, have a non-zero effect size, whereas, all markers that are not, have a zero effect size:

$$\begin{cases} \delta_p^{(1)} - \delta_p^{(0)} > 0 & \text{if protein } p \text{ is in activation set } p \in C \\ \delta_{p'}^{(1)} - \delta_{p'}^{(0)} = 0 & \text{if protein } p' \text{ is not in activation set } p' \notin C. \end{cases}$$

Both covariance matrices have an autoregressive structure,

$$\begin{aligned} \Omega_{rs} &= \rho^{|r-s|} \\ \boldsymbol{\Sigma} &= \text{diag}(\boldsymbol{\sigma})\, \boldsymbol{\Omega}\ \text{diag}(\boldsymbol{\sigma}). \end{aligned}$$

We regulate two separate correlation parameters: a cell-level ρ_B and a donor-level ρ_U coefficient. Non-zero ρ_B or ρ_U induce a correlation between condition and marker expression even for markers with a zero effect size.

We performed simulations with a variety of different parameters. All simulations have 16 samples. For paired samples, those 16 samples come from 8 donors. For unpaired samples, those 16 samples come from 16 donors. Each sample has 1,000 cells. We compared the observed FDR and the power. The FDR measures the statistical type 1 errors, the expected proportion of falsely declared discoveries over the total number of reported discoveries. The statistical power represents the proportion of correctly reported discoveries over the total number of true discoveries.

Figures 1 and 2 show a summary averaged over 100 runs for paired sample and unpaired sample experiments with effect size $\delta_p^{(1)} - \delta_p^{(0)} = 1.8$ and $\delta_p^{(1)} - \delta_p^{(0)} = 15$, respectively, and varying standard deviation σ and correlation ρ parameters. The dashed lines indicate the target FDR of 0.05.

First, let's consider the paired samples experiment. The plots on the left show results when we vary cell and donor-level correlations at a fixed amount of cell $\sigma_B = 1$ and donor $\sigma_U = 1$ marker standard deviations. We observe only small differences across donor correlations ρ_U of a small increase of power with increasing correlation. In contrast, there are large increases of power as a function of cell correlations ρ_B. In the panel of plots on the right, we set both correlations to zero and vary the marker standard deviations. In this setting, we again observe major changes with increasing standard deviations at the cell-level σ_B: the larger the cell-level variability, the lower the power. This is also true for donor-level variability, though to a much lesser extent. FDR is controlled below its target level under medium cell-level marker correlations ($|\rho_B| \leq 0.4$) except when cell variability is at zero $\sigma_B = 0$, and donor variability is at one $\sigma_U = 1$. As expected, the BY procedure is more conservative than the BH procedure, that is both FDR and power are lower. Interestingly, power increases with cell-level correlations ρ_B, and is virtually unaffected by donor-level correlations ρ_U. Overall, GLMM methods are more powerful than GLM methods. Figure 3 shows simulations for power and FDR with varying numbers of paired samples. Both cell and donor standard deviations are set to $\sigma_B = \sigma_U = 1$, and correlations are set to $\rho_B = \rho_U = 0$. An efficiency gain is clearly visible when we compare how many paired samples are needed to achieve 80% power. We observe that for GLMMs we need 7 paired samples to exceed the 80% power threshold, whereas for GLMs we need 13 paired samples to achieve the same.

In the unpaired samples experiment, we only show GLM results as the GLMM results have zero power, there is no data to estimate the donor-level random effect term. We observe up to 20% FDR with a target FDR of 5%. To have non-zero power we need to increase the effect size to 15 (in comparison, for paired experiments

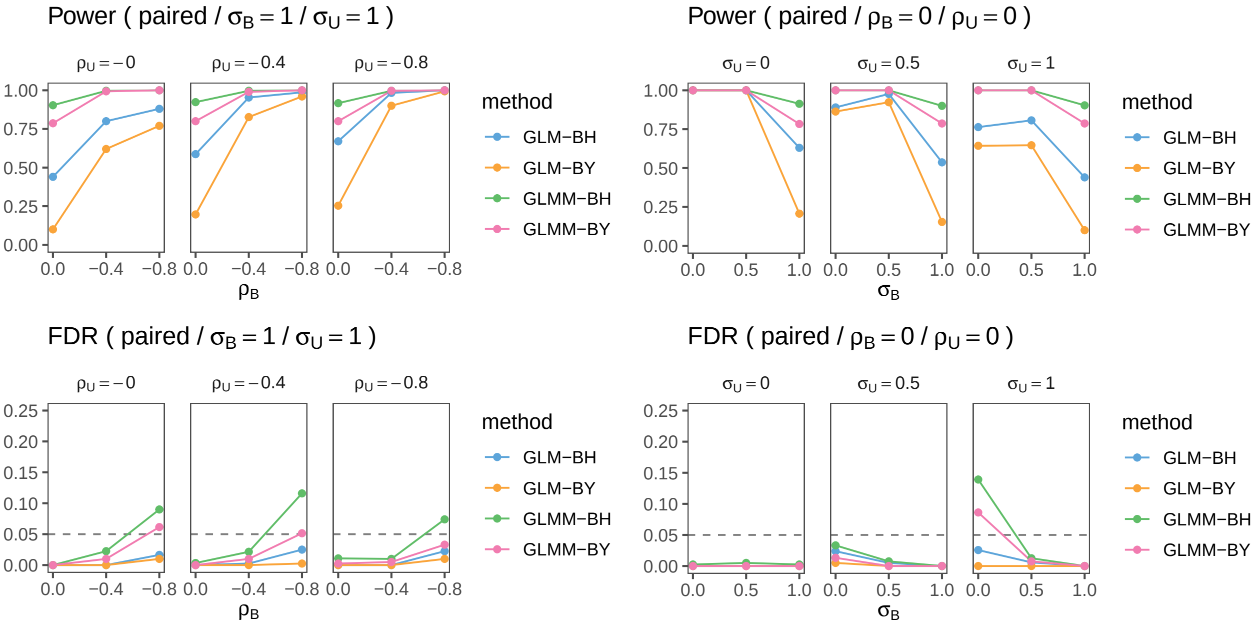


Figure 1: Summary of experiments with 1,000 cells per sample averaged over 100 runs. The horizontal dashed line represents the target FDR. Postfixes BH and BY stand for the respective FDR control procedure. Subscripts B and U indicate cell and donor-level standard deviation σ and correlation ρ, respectively.

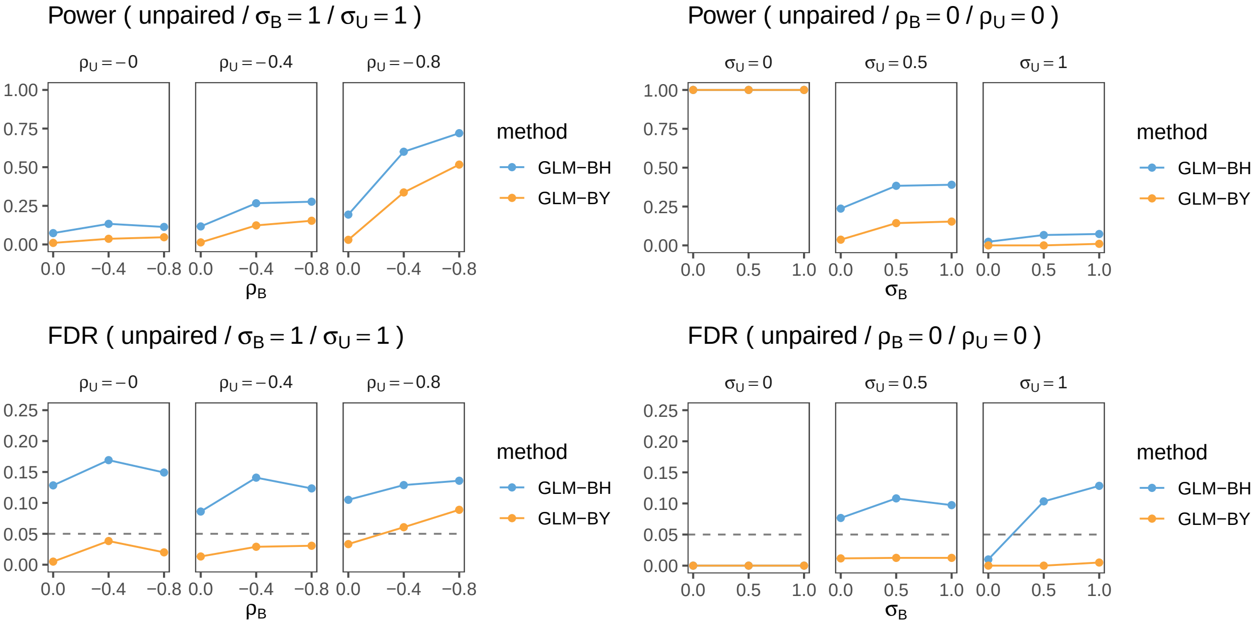


Figure 2: Summary of experiments with 1,000 cells per sample averaged over 100 runs. The horizontal dashed line represents the target FDR. Postfixes BH and BY stand for the respective FDR control procedure. Subscripts B and U indicate cell and donor-level standard deviation σ and correlation ρ, respectively.

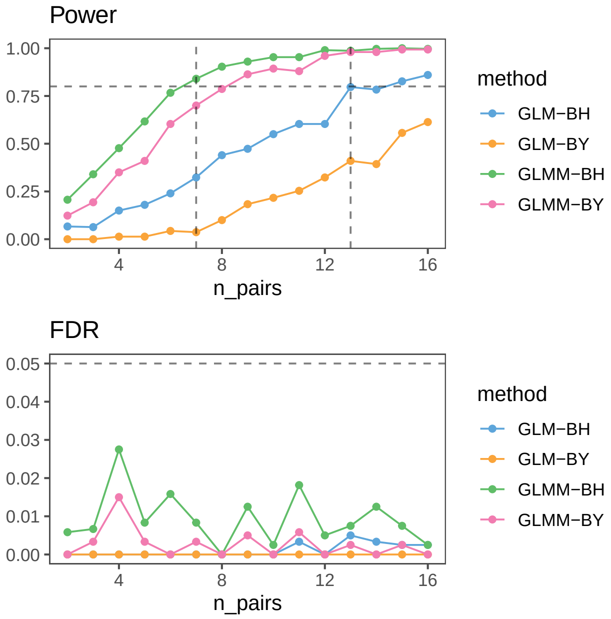


Figure 3: Summary of experiments with 1,000 cells per sample averaged over 100 runs. Power: The horizontal dashed line represents a power of 0.8. FDR: The horizontal dashed line represents the target FDR of 0.05.

the effect size is set to 1.8). Furthermore, FDR is only controlled under medium cell-level marker correlations using the more conservative BY procedure, with BH exceeding 0.05 in most scenarios except when we have zero donor-level variability $\sigma_U = 0$. As before, BY comes with a loss of power.

3.2 Experimental Dataset

We reanalyze a recently published dataset on the maternal immune system during pregnancy (Aghaeepour et al. 2017). The study provides a rich mass cytometry dataset collected at four time points during pregnancy in two cohorts. The authors isolated cells from blood samples and stimulated them with several activation factors. The goal was to explain how immune cells react to these stimuli, and how these reactions change throughout pregnancy. Findings from such experiments might identify immunological deviations implicated in pregnancy-related pathologies.

The authors collected data at early, mid, late pregnancy, and six weeks postpartum. Samples were left unstimulated or stimulated. Stimulation conditions included: interferon-α2A (IFNα), lipopolysaccharide, and a cocktail of interleukins (ILs) containing IL-2 and IL-6. They processed the samples on a CyTOF 2.0 mass cytometer instrument, and bead normalized the data to account for signal variation over time from changes in instrument performance (Finck et al. 2013).

In our analysis, we focus on comparing early (first trimester, $Y_i = 0$) with late (third trimester, $Y_i = 1$) pregnancy samples stimulated with IFNα in the first cohort of 16 women. We gate cells into cell types and organize them in a data frame. We follow the gating scheme detailed in (Aghaeepour et al. 2017) and define 12 cell types using the *R* package `openCyto` (Finak et al. 2014): memory CD4 positive T cells (CD4+Tmem), naive CD4 positive T cells (CD4+Tnaive), memory CD8 positive T cells (CD8+Tmem), naive CD8 positive T cells (CD8+Tnaive), $\gamma\delta$T cells (gdT), regulatory T memory cells (Tregsmem), regulatory T naive cells (Tregsnaive), B cells, classical monocytes (cMC), intermediate monocytes (intMC), non-classical monocytes (ncMC), and Natural Killer cells (NK). Out of the 32 protein markers measured on each cell, the authors defined 22 markers as gating markers, and 10 as functional markers. The functional markers are pSTAT1, pSTAT3, pSTAT5, pNFκB, total IκB, pMAPKAPK2, pP38, prpS6, pERK1/2, and pCREB (in plots Greek symbols are replaced by Latin symbols).

We plot the maximum likelihood (GLM) and the method of moments estimates (GLMM) with 95% confidence intervals for the fixed effects $\boldsymbol{\beta}$ (Figure 4). The estimates are on the log-odds scale. We see that pSTAT1 is a strong predictor of the third trimester. This means that one unit increase in the transformed marker expression makes it between $\exp(1) = 2.7$ to $\exp(1.5) = 4.5$ (95% confidence interval for GLMM) more likely to be a cell from the third trimester, while holding the other markers constant. pSTAT3 and pSTAT5 have negative coefficients. This means pSTAT3 and pSTAT5 predict the first trimester, while holding the other markers constant. Only pSTAT1, pSTAT3, and pSTAT5 are below an FDR of 0.05. Our results corroborate previous findings by Aghaeepour et al. (2017) reporting an increase of pSTAT1 during the third trimester for IFNα stimulated samples.

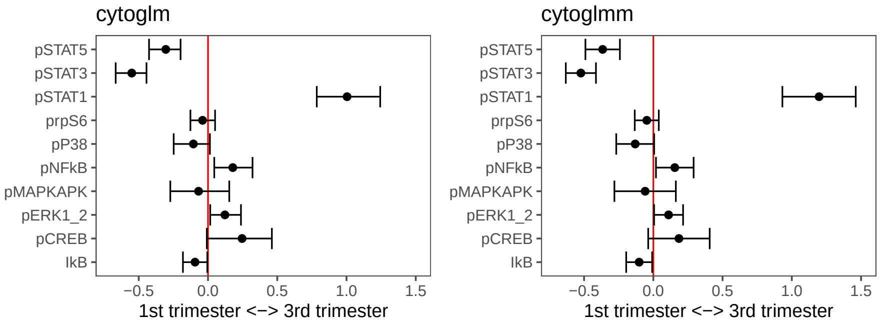


Figure 4: Methods comparison between bootstrap GLM (cytoglm) and GLMM (cytoglmm). The horizontal axes are on the log-odds scale. The vertical axes are the protein markers.

4 Discussion

Our new *R* package `CytoGLMM` is applicable to a wide range of cytometry studies. Besides comparisons on paired samples, where samples are available for the same subject under different experimental conditions, our `CytoGLMM` is applicable to unpaired samples, where samples are collected on two separate groups of individuals.

Our simulation experiments compare multiple regression GLM and GLMM, as implemented in `cytoglm` and `cytoglmm` in our *R* package. In simulated paired samples experiments, both GLMM-BH and GLMM-BY procedures control the FDR below the target FDR under cell-level marker correlations with an autoregressive structure with correlations up to ±0.4. GLMM methods are more powerful than GLM methods for paired samples. GLMM methods can account for the patient-to-patient variation in the model, whereas GLM methods treat this variation as noise, which results in noisier and thus less powerful estimates. For unpaired samples, we are forced to use the nonparametric bootstrap method for GLMs because there are no paired samples available to estimate the random effect term. In simulated unpaired experiments, only BY controls the target FDR. In practice, this means that we need a much higher donor samples size to detect a differential expression compared to paired experiments.

Overall, larger cell-level and donor-level correlations increase power and reduce the observed FDR. Hypothesis testing under arbitrary dependency structure is still an active research topic (Barber, Candès, and others 2015; Candès et al. 2018; Fithian and Lei 2020). What is easier to explain is the reduction in power and FDR as a function of increased cell-level variance. Research in measurement error models shows that increased uncertainty in measured covariates is linked to biased estimates. The coefficient estimates are regularized—shrinking them towards zero—which translates into more conservative p-values; for extensive literature on this topic see Fuller (1987) and Carroll et al. (2006). In GLMMs, donor-level correlations have only a weak impact on power and observed FDR because we explicitly model correlations with a random effect term.

In general, biases in coefficient estimates of GLMs and GLMMs can occur when we leave out proteins from the analysis. Assume that we would like to relate variable protein X to experimental condition Y. If there exists a second protein Z both related to X and Y, then Z is called a confounder, and not including it in the analysis can change the coefficients estimates. In the pregnancy data, if we removed pSTAT1 from the analysis, the confidence intervals of pSTAT3 and pSTAT5 could change. Such a difference is expected if pSTAT1 is a confounder. If pSTAT1 is not a confounder, the coefficient estimates for pSTAT3 and pSTAT5 will be the same whether pSTAT1 is included or not. The change of coefficients depending on what markers are included in the model can have strong effects. We have observed that in some real datasets that one marker can make other markers change their sign depending on whether we include them or not. In such cases, we recommend keeping all markers in the analysis to avoid introducing confounding biases.

Our simulations are limited to a Poisson mixed effect model for protein marker expression. Our conclusions are only valid with respect to this model. The real data generating process might be different. Two main caveats are to be noted. First, we can only encode an experimental design comparing two groups. Second, we require gated cell types. To reduce the person-to-person bias of manual gating, we employed the *R* package `openCyto` (Finak et al. 2014). The curse of dimensionality makes it challenging to scale this approach to very high dimensional gating schemes.

A possible alternative to GLMMs are Generalized Estimating Equations (GEEs). GEEs are statistically more efficient when the covariance structure of the residuals are known. In our case, the covariance structure is unknown and needs to be estimated from the data. In most immunology studies, we only have a few donors without a given covariance structure (e.g. no time dependency), resulting in a hard and possibly unstable covariance estimation problem, which could result in an overall loss of efficiency (Wakefield 2013).

We applied `CytoGLMM` in wide range of immunology studies: comparison between influenza strains (Kronstad et al. 2018), comparison between pregnant and non-pregnant women (Le Gars et al. 2019), comparison between healthy controls and HIV+ individuals (Vendrame et al. 2020), comparison between multiple sclerosis patients treated with daclizumab beta or placebo (Ranganath et al. 2020), and comparison between Beninese sex workers and healthy controls (Zhao et al. 2020). Our next step is extending `CytoGLMM` to include more complicated experimental designs; e.g. twin studies (Brodin et al. 2015).

Acknowledgments

This work was supported by the National Institutes of Health [U01AI131302 to C.A.B. and S.H., R56AI124788 to C.A.B. and S.H., R21AI130523 to C.A.B. and S.H., DP1DA046089 to C.A.B., R21AI130532 to C.A.B., R01AI133698 to C.A.B., R21AI135287 to C.A.B., 5T32AI007290-29 to L.M.K., TL1TR001084 to E.V., T32AI007502 to E.V., 1F32AI126674 to L.J.S.]; an A.P. Giannini fellowship [to L.M.K.]; and a Stanford Child Health Research Institute postdoctoral fellowship [to M.L.G.]. C.A.B. is the Tashia and John Morgridge Endowed Faculty Scholar in Pediatric Translational Medicine from the Maternal Child Health Research Institute, and a Chan Zuckerberg Investigator.

Supplementary Material

`CytoGLMM` *R* Package

Our *R* package is available on GitHub:

- https://github.com/christofseiler/CytoGLMM/

A vignette is available on our *R* package website:

- https://christofseiler.github.io/CytoGLMM/

`CytoGLMM` Workflow

Here we illustrate a complete `CytoGLMM` workflow in *R*. Prepare data frame marker counts and sample information.

```
str(df)
```

```
## tibble [2,503,107 x 13] (S3: tbl_df/tbl/data.frame)
##  $ donor    : Factor w/ 16 levels "PTLG001","PTLG002",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ condition: Factor w/ 2 levels "1st trimester",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ celltype : Factor w/ 12 levels "B","CD4+Tmem",..: 10 10 10 10 10 10 10 10 10 10 ...
##  $ pCREB    : num [1:2503107] 4 14 7 4 7 4 14 4 3 6 ...
##  $ pSTAT5   : num [1:2503107] 9 12 3 13 2 8 11 3 7 7 ...
##  $ pP38     : num [1:2503107] 6 4 1 7 6 1 6 4 3 2 ...
##  $ pSTAT1   : num [1:2503107] 35 82 7 48 32 30 57 37 23 85 ...
##  $ pSTAT3   : num [1:2503107] 15 33 7 23 11 3 17 16 14 29 ...
##  $ prpS6    : num [1:2503107] 13 13 9 5 12 4 2 7 11 6 ...
##  $ pMAPKAPK : num [1:2503107] 20 36 8 22 30 38 12 20 25 20 ...
##  $ IkB      : num [1:2503107] 6 3 1 1 3 3 0 2 2 2 ...
##  $ pNFkB    : num [1:2503107] 13 7 2 4 16 4 16 8 9 12 ...
##  $ pERK1_2  : num [1:2503107] 2 0 0 1 0 1 0 0 0 2 ...
```

Select functional marker of interest.

```r
protein_names
```

```
##  [1] "pCREB"    "pSTAT5"   "pP38"     "pSTAT1"   "pSTAT3"   "prpS6"
##  [7] "pMAPKAPK" "IkB"      "pNFkB"    "pERK1_2"
```

Transform counts.

```r
df %<>% dplyr::mutate_at(protein_names, function(x) asinh(x/5))
```

Subset to one celltype.

```r
df %<>% dplyr::filter(celltype == "NK")
```

Fit the cytoglm model with 1000 bootstrap samples.

```r
glm_fit = CytoGLMM::cytoglm(df,
                            protein_names = protein_names,
                            condition = "condition",
                            group = "donor",
                            num_boot = 1000)
```

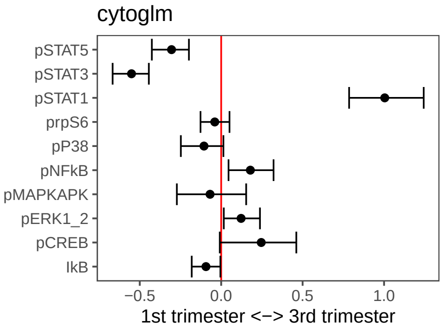
Fit the cytoglmm model.

```r
glmm_fit = CytoGLMM::cytoglmm(df,
                              protein_names = protein_names,
                              condition = "condition",
                              group = "donor")
```

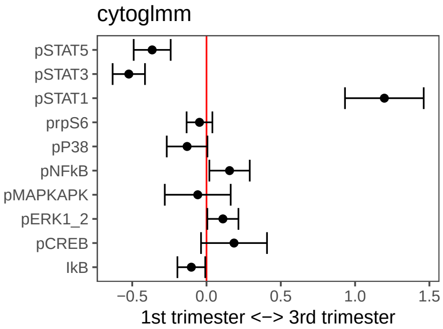
Use print(glm_fit) or print(glmm_fit) on the fitted object to obtain some additional details of the model that we just fitted.

Plot differential analysis results.

```r
plot(glm_fit)
```



```r
plot(glmm_fit)
```



Extract p-values.

```r
summary(glm_fit)
```

```
## # A tibble: 10 x 3
##    protein_name pvalues_unadj pvalues_adj
##    <chr>                <dbl>       <dbl>
##  1 pSTAT1               0.002     0.00667
##  2 pSTAT3               0.002     0.00667
##  3 pSTAT5               0.002     0.00667
##  4 pNFkB                0.028     0.07
##  5 pERK1_2              0.068     0.136
##  6 IkB                  0.088     0.147
##  7 pCREB                0.106     0.151
##  8 pP38                 0.142     0.178
##  9 prpS6                0.49      0.544
## 10 pMAPKAPK             0.592     0.592
```

```r
summary(glmm_fit)
```

```
## # A tibble: 10 x 3
##    protein_name pvalues_unadj pvalues_adj
##    <chr>                <dbl>       <dbl>
##  1 pSTAT3            4.73e-21    4.73e-20
##  2 pSTAT1            8.21e-19    4.11e-18
##  3 pSTAT5            8.46e- 9    2.82e- 8
##  4 pNFkB             2.53e- 2    6.31e- 2
##  5 IkB               3.26e- 2    6.50e- 2
##  6 pERK1_2           3.90e- 2    6.50e- 2
##  7 pP38              6.15e- 2    8.78e- 2
##  8 pCREB             1.02e- 1    1.28e- 1
##  9 prpS6             2.86e- 1    3.18e- 1
## 10 pMAPKAPK          6.01e- 1    6.01e- 1
```

Filter adjusted p-values at some threshold.

```
summary(glm_fit) %>%
  dplyr::filter(pvalues_adj < 0.05)
```

```
## # A tibble: 3 x 3
##   protein_name pvalues_unadj pvalues_adj
##   <chr>                <dbl>       <dbl>
## 1 pSTAT1               0.002     0.00667
## 2 pSTAT3               0.002     0.00667
## 3 pSTAT5               0.002     0.00667
```

```
summary(glmm_fit) %>%
  dplyr::filter(pvalues_adj < 0.05)
```

```
## # A tibble: 3 x 3
##   protein_name pvalues_unadj pvalues_adj
##   <chr>                <dbl>       <dbl>
## 1 pSTAT3            4.73e-21    4.73e-20
## 2 pSTAT1            8.21e-19    4.11e-18
## 3 pSTAT5            8.46e- 9    2.82e- 8
```

References

Abdelaal, Tamim, Vincent van Unen, Thomas Höllt, Frits Koning, Marcel JT Reinders, and Ahmed Mahfouz. 2019. "Predicting Cell Populations in Single Cell Mass Cytometry Data." *Cytometry Part A* 95 (7): 769–81.

Aghaeepour, Nima, Edward A. Ganio, David Mcilwain, Amy S. Tsai, Martha Tingle, Van GassenSofie, Dyani K. Gaudilliere, et al. 2017. "An Immune Clock of Human Pregnancy." *Science Immunology* 2 (15): eaan2946.

Aghaeepour, Nima, Radina Nikolic, Holger H Hoos, and Ryan R Brinkman. 2011. "Rapid Cell Population Identification in Flow Cytometry Data." *Cytometry Part A* 79 (1): 6–13.

Arvaniti, Eirini, and Manfred Claassen. 2017. "Sensitive Detection of Rare Disease-Associated Cell Subsets via Representation Learning." *Nature Communications* 8: 14825.

Barber, Rina Foygel, Emmanuel J Candès, and others. 2015. "Controlling the False Discovery Rate via Knockoffs." *The Annals of Statistics* 43 (5): 2055–85.

Becher, Burkhard, Andreas Schlitzer, Jinmiao Chen, Florian Mair, Hermi R Sumatoh, Karen Wei Weng Teng, Donovan Low, et al. 2014. "High-Dimensional Analysis of the Murine Myeloid Cell System." *Nature Immunology* 15 (12): 1181.

Bendall, Sean C, Erin F Simonds, Peng Qiu, D Amir El-ad, Peter O Krutzik, Rachel Finck, Robert V Bruggner, et al. 2011. "Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum." *Science* 332 (6030): 687–96.

Benjamini, Yoav, and Yosef Hochberg. 1995. "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.

Benjamini, Yoav, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing Under Dependency." *Annals of Statistics*, 1165–88.

Brodin, Petter, Vladimir Jojic, Tianxiang Gao, Sanchita Bhattacharya, Cesar J Lopez Angel, David Furman, Shai Shen-Orr, et al. 2015. "Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences." *Cell* 160 (1-2): 37–47.

Bruggner, Robert V, Bernd Bodenmiller, David L Dill, Robert J Tibshirani, and Garry P Nolan. 2014. “Automated Identification of Stratifying Signatures in Cellular Subpopulations.” *Proceedings of the National Academy of Sciences* 111 (26): E2770–E2777.

Candès, Emmanuel, Yingying Fan, Lucas Janson, and Jinchi Lv. 2018. “Panning for Gold: ‘Model-X’ Knockoffs for High Dimensional Controlled Variable Selection.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80 (3): 551–77.

Carroll, Raymond J, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective.* CRC Press.

Chen, Hao, Mai Chan Lau, Michael Thomas Wong, Evan W Newell, Michael Poidinger, and Jinmiao Chen. 2016. “Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline.” *PLOS Computational Biology* 12 (9): e1005112.

Chevrier, Stéphane, Helena L Crowell, Vito RT Zanotelli, Stefanie Engler, Mark D Robinson, and Bernd Bodenmiller. 2018. “Compensation of Signal Spillover in Suspension and Imaging Mass Cytometry.” *Cell Systems* 6 (5): 612–20.

Efron, Bradley, and Robert J Tibshirani. 1994. *An Introduction to the Bootstrap.* CRC press.

Finak, Greg, Ali Bashashati, Ryan Brinkman, and Raphaël Gottardo. 2009. “Merging Mixture Components for Cell Population Identification in Flow Cytometry.” *Advances in Bioinformatics* 2009.

Finak, Greg, Jacob Frelinger, Wenxin Jiang, Evan W Newell, John Ramey, Mark M Davis, Spyros A Kalams, De RosaStephen C, and Raphael Gottardo. 2014. “OpenCyto: An Open Source Infrastructure for Scalable, Robust, Reproducible, and Automated, End-to-End Flow Cytometry Data Analysis.” *PLOS Computational Biology* 10 (8): e1003806.

Finck, Rachel, Erin F Simonds, Astraea Jager, Smita Krishnaswamy, Karen Sachs, Wendy Fantl, Pe’erDana, Garry P Nolan, and Sean C Bendall. 2013. “Normalization of Mass Cytometry Data with Bead Standards.” *Cytometry Part A* 83 (5): 483–94.

Fithian, William, and Lihua Lei. 2020. “Conditional Calibration for False Discovery Rate Control Under Dependence.” *arXiv Preprint arXiv:2007.10438.*

Fuller, Wayne A. 1987. *Measurement Error Models.* John Wiley & Sons.

Ge, Yongchao, and Stuart C Sealfon. 2012. “flowPeaks: A Fast Unsupervised Clustering for Flow Cytometry Data via K-Means and Density Peak Finding.” *Bioinformatics* 28 (15): 2052–8.

Holmes, Susan, and Wolfgang Huber. 2019. *Modern Statistics for Modern Biology.* Cambridge University Press.

Huber, Wolfgang, Anja von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. 2003. “Parameter Estimation for the Calibration and Variance Stabilization of Microarray Data.” *Statistical Applications in Genetics and Molecular Biology* 2 (1).

Kronstad, Lisa M, Christof Seiler, Rosemary Vergara, Susan P Holmes, and Catherine A Blish. 2018. “Differential Induction of IFN-α and Modulation of CD112 and CD54 Expression Govern the Magnitude of NK Cell IFN-γ Response to Influenza A Viruses.” *The Journal of Immunology* 201 (7): 2117–31.

Lee, Hao-Chih, Roman Kosoy, Christine E Becker, Joel T Dudley, and Brian A Kidd. 2017. “Automated Cell Type Discovery and Classification Through Knowledge Transfer.” *Bioinformatics* 33 (11): 1689–95.

Le Gars, Mathieu, Christof Seiler, Alexander W. Kay, Nicholas L. Bayless, Elina Starosvetsky, Lindsay Moore, Shai S. Shen-Orr, et al. 2019. “Pregnancy-Induced Alterations in NK Cell Phenotype and Function.” *Frontiers in Immunology* 10 (2469): 1–13.

Levine, Jacob H, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, et al. 2015. “Data-Driven Phenotypic Dissection of AML Reveals Progenitor-Like Cells That Correlate with Prognosis.” *Cell* 162 (1): 184–97.

Li, Huamin, Uri Shaham, Kelly P Stanton, Yi Yao, Ruth R Montgomery, and Yuval Kluger. 2017. "Gating Mass Cytometry Data by Deep Learning." *Bioinformatics* 33 (21): 3423–30.

Lo, Kenneth, Florian Hahne, Ryan R Brinkman, and Raphael Gottardo. 2009. "FlowClust: A Bioconductor Package for Automated Gating of Flow Cytometry Data." *BMC Bioinformatics* 10 (1): 145.

Lun, Aaron TL, Arianne C Richard, and John C Marioni. 2017. "Testing for Differential Abundance in Mass Cytometry Data." *Nature Methods* 14 (7): 707.

Meehan, Stephen, Guenther Walther, Wayne Moore, Darya Orlova, Connor Meehan, David Parks, Eliver Ghosn, et al. 2014. "AutoGate: Automating Analysis of Flow Cytometry Data." *Immunologic Research* 58 (2-3): 218–23.

Naim, Iftekhar, Suprakash Datta, Jonathan Rebhahn, James S Cavenaugh, Tim R Mosmann, and Gaurav Sharma. 2014. "SWIFT – Scalable Clustering for Automated Identification of Rare Cell Populations in Large, High-Dimensional Flow Cytometry Datasets, Part 1: Algorithm Design." *Cytometry Part A* 85 (5): 408–21.

Nowicka, M, C Krieg, LM Weber, FJ Hartmann, S Guglietta, B Becher, MP Levesque, and MD Robinson. 2017. "CyTOF Workflow: Differential Discovery in High-Throughput High-Dimensional Cytometry Datasets [Version 2; Referees: 2 Approved]." *F1000Research* 6 (748).

Perry, Patrick O. 2017. "Fast Moment-Based Estimation for Hierarchical Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 (1): 267–91.

Qian, Yu, Chungwen Wei, Eun-Hyung LeeF, John Campbell, Jessica Halliley, Jamie A Lee, Jennifer Cai, et al. 2010. "Elucidation of Seventeen Human Peripheral Blood B-Cell Subsets and Quantification of the Tetanus Response Using a Density-Based Method for the Automated Identification of Cell Populations in Multidimensional Flow Cytometry Data." *Cytometry Part B: Clinical Cytometry* 78 (S1): S69–S82.

Qiu, Peng, Erin F Simonds, Sean C Bendall, Gibbs JrKenneth D, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. 2011. "Extracting a Cellular Hierarchy from High-Dimensional Cytometry Data with SPADE." *Nature Biotechnology* 29 (10): 886.

Ranganath, Thanmayi, Laura Jane Simpson, Anne-Maud Ferreira, Christof Seiler, Elena Vendrame, Nancy Q Zhao, Jason D Fontenot, Susan P Holmes, and Catherine A Blish. 2020. "Characterization of the Impact of Daclizumab Beta on Circulating Natural Killer Cells by Mass Cytometry." *Frontiers in Immunology* 11 (714): 1–13.

Rocke, David M, and Stefan Lorenzato. 1995. "A Two-Component Model for Measurement Error in Analytical Chemistry." *Technometrics* 37 (2): 176–84.

Saeys, Yvan, Van GassenSofie, and Bart N Lambrecht. 2016. "Computational Flow Cytometry: Helping to Make Sense of High-Dimensional Immunology Data." *Nature Reviews Immunology* 16 (7): 449.

Samusik, Nikolay, Zinaida Good, Matthew H Spitzer, Kara L Davis, and Garry P Nolan. 2016. "Automated Mapping of Phenotype Space with Single-Cell Data." *Nature Methods* 13 (6): 493.

Schuyler, Ronald P., Conner Jackson, Josselyn E. Garcia-Perez, Ryan M. Baxter, Sidney Ogolla, Rosemary Rochford, Debashis Ghosh, Pratyaydipta Rudra, and Elena W. Y. Hsieh. 2019. "Minimizing Batch Effects in Mass Cytometry Data." *Frontiers in Immunology* 10: 2367.

Shekhar, Karthik, Petter Brodin, Mark M Davis, and Arup K Chakraborty. 2014. "Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE)." *Proceedings of the National Academy of Sciences* 111 (1): 202–7.

Sörensen, Till, Sabine Baumgart, Pawel Durek, Andreas Grützkau, and Thomas Häupl. 2015. "ImmunoClust – An Automated Analysis Pipeline for the Identification of Immunophenotypic Signatures in High-Dimensional Cytometric Datasets." *Cytometry Part A* 87 (7): 603–15.

Theorell, Axel, Yenan Troi Bryceson, and Jakob Theorell. 2019. "Determination of Essential Phenotypic Elements of Clusters in High-Dimensional Entities–DEPECHE." *PLOS ONE* 14 (3).

Trussart, Marie, Charis E Teh, Tania Tan, Lawrence Leong, Daniel HD Gray, and Terence P Speed. 2020. "Removing Unwanted Variation with CytofRUV to Integrate Multiple CyTOF Datasets." *eLife* 9 (September): e59630.

Van Gassen, Sofie, Britt Callebaut, Van HeldenMary J, Bart N Lambrecht, Piet Demeester, Tom Dhaene, and Yvan Saeys. 2015. "FlowSOM: Using Self-Organizing Maps for Visualization and Interpretation of Cytometry Data." *Cytometry Part A* 87 (7): 636–45.

Van Gassen, Sofie, Brice Gaudilliere, Martin S Angst, Yvan Saeys, and Nima Aghaeepour. 2020. "CytoNorm: A Normalization Algorithm for Cytometry Data." *Cytometry Part A* 97 (3): 268–78.

Vendrame, Elena, Christof Seiler, Thanmayi Ranganath, Nancy Q Zhao, Rosemary Vergara, Michel Alary, Annie Claude Labbé, et al. 2020. "TIGIT Is Upregulated by HIV-1 Infection and Marks a Highly Functional Adaptive and Mature Subset of Natural Killer Cells." *AIDS* 34 (6): 801–13.

Wakefield, Jon. 2013. *Bayesian and Frequentist Regression Methods*. Springer Series in Statistics. Springer, New York.

Weber, Lukas M, Malgorzata Nowicka, Charlotte Soneson, and Mark D Robinson. 2019. "Diffcyt: Differential Discovery in High-Dimensional Cytometry via High-Resolution Clustering." *Communications Biology* 2 (1): 1–11.

Weber, Lukas M, and Mark D Robinson. 2016. "Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data." *Cytometry Part A* 89 (12): 1084–96.

Zare, Habil, Parisa Shooshtari, Arvind Gupta, and Ryan R Brinkman. 2010. "Data Reduction for Spectral Clustering to Analyze High Throughput Flow Cytometry Data." *BMC Bioinformatics* 11 (1): 403.

Zhao, Nancy Q, Elena Vendrame, Anne-Maud Ferreira, Christof Seiler, Thanmayi Ranganath, Michel Alary, Annie-Claude Labbé, et al. 2020. "Natural Killer Cell Phenotype Is Altered in HIV-Exposed Seronegative Women." *PLoS ONE* 15 (9): e0238347.