FastMix: A Versatile Multi-Omics Data Integration Pipeline for Cell Type-Specific Biomarker Inference

Yun Zhang[1,*], Hao Sun[2,*], Aishwarya Mandava[1], Brian D. Aevermann[1], Tobias R. Kollmann[3], Richard H. Scheuermann[1,4], Xing Qiu[2,#], Yu Qian[1,#]

[*]Contributed equally

[#]Correspondence should be sent to: xing_qiu@urmc.rochester.edu and mqian@jcvi.org

Affiliations:

[1]Informatics, J. Craig Venter Institute, La Jolla, CA, United States,
[2]Biostatistics and Computational Biology, University of Rochester, Rochester, NY, United States,
[3]Telethon Kids Institute, Perth Children's Hospital, University of Western Australia; Nedlands, Australia,
[4]Pathology, University of California, San Diego, La Jolla, CA, United States

Abstract

We developed a novel analytic pipeline - `FastMix` - to integrate flow cytometry, bulk transcriptomics, and clinical covariates for statistical inference of cell type-specific gene expression signatures. `FastMix` addresses the "large p, small n" problem via a carefully designed linear mixed effects model (LMER), which is applicable for both cross-sectional and longitudinal studies. With a novel moment-based estimator, `FastMix` runs and converges much faster than competing methods for big data analytics. The pipeline also includes a cutting-edge flow cytometry data analysis method for identifying cell population proportions. Simulation studies showed that `FastMix` produced smaller type I/II errors with more accurate parameter estimation than competing methods. When applied to real transcriptomics and flow cytometry data in two vaccine studies, `FastMix`-identified cell type-specific signatures were largely consistent with those obtained from the single cell RNA-seq data, with some unique interesting findings.

Introduction

High throughput multi-omics technologies are becoming popular. In a multi-omics study, different types of sample characteristics of the same subject, e.g., genomics, transcriptomics, epigenomics, proteomics and metabolomics, are measured using a variety of bioassays. Recent publications [1-7] have shown that systems biology approaches based on multi-omics integrative data analysis can effectively identify important patterns that otherwise would be missed using a single assay. One key challenge for multi-omics data integration is the "large p, small n" problem. Each type of assay can measure many analytes. As a result, the total number of experiment variables (p) involved in a multi-omics study is usually large. The number of experimental subjects and their samples (n), however, is usually more limited due to cost and enrollment capability. When $p >> n$, it is unreliable to infer or interpret the relationship among the variables using standard regression models.

Dimensionality reduction and regularization are two common approaches to address this issue. Common data dimensionality reduction techniques include linear projection methods such as principal component analysis (PCA) [8], canonical correlation analysis (CCA) [9] and partial least squares (PLS) [10], as well as non-linear embedding methods such as t-distributed stochastic neighbor embedding (t-SNE) [11] and uniform manifold approximation projection (UMAP) [12]. Well-established regularization methods include ridge [13], LASSO [14], and elastic-net [15]. Currently, major efforts in the field of multi-omics integrative analysis focus on using one or both types of approaches to address the "large p" problem. For example, DIABLO [16] uses sparse generalized canonical correlation analysis (sGCCA) with LASSO penalty to integrate data

from multiple omics assays and predict patient's disease type. LUCID [17] uses a joint probabilistic model with latent variables for integrated clustering regularized by LASSO. In the emerging single cell genomics field, UMAP and other embedding techniques are frequently used for the dimensionality reduction purpose for multi-modality data integration [18, 19]. However, the fundamental question of biomarker detection, which requires the integration of both assay data and clinical covariates with differential analysis, has not been solved.

Among the applicable statistical models, linear-mixed effects regression (LMER) is a powerful and generalizable framework that can be used to address the "large *p*, small *n*" problem for multi-omics data integration. Regularized fixed effects regression models have been shown to solve the "large *p*, small *n*" issue and reduce the variability in the estimation procedure by shrinking the estimates toward zero. On the other hand, LMER shrinks the estimates toward the *fixed effects* instead of zero, so that they have less variability and are less biased [20, 21]. However, LMER is not widely applied to high-throughput multi-omics data because of its high computational cost and numerical instability. Conventionally, a LMER model is solved by an expectation-maximization (EM) algorithm, which iteratively finds maximum likelihood estimate of regression parameters [22]. This iterative process is slow and prone to convergence issues, which makes the LMER almost impossible to be applied to analyze data from high-throughput studies.

To reduce the high computational cost of the iterative EM algorithm, we designed a non-iterative, moment-based covariance estimator, which is not only more robust than the iterative EM process but also more efficient, requiring a small fraction of EM's run time. To both

demonstrate the utility and evaluate the performance of the proposed approach, we combined the moment-based estimation of LMER together with downstream differential expression (DE) analysis into a computational pipeline for inferring cell type-specific differentially expressed genes (DEGs) from bulk gene expressions and flow cytometry (FCM) data, a problem that is commonly encountered in immunology studies but not specifically addressed by the existing multi-omics data integration methods. As shown in Figure 1a, the proposed model – `FastMix` – takes in three sets of input data: (i) bulk gene expressions measured by microarray or RNA sequencing; (ii) proportions of cell populations identified from FCM data; and (iii) experiment covariates and clinical parameters such as demographics, cohorts, and visits of the subjects. Figure 1b depicts the main steps in the `FastMix` modeling and testing framework and key techniques to obtain accurate parameter estimations. Expected improvement of parameter estimation by the `FastMix` is illustrated in Figure 1c.

`FastMix` optimizes the bias-variance tradeoff in a way that the DEGs (dots in black) can be estimated closer to the ground truth than the standard approach (Figure 1c). Figure 1d provides a schematic representation of the whole analytic pipeline. Unlike traditional unsupervised analyses that rely on predefined marker genes (which are often incomplete or totally unknown when the cell types are novel) to define cell types for estimating the cell composition data, our proposed pipeline integrates the FCM data with bulk gene expression data to supervise the cell type-specific inference. `FastMix` inference provides a baseline method for cell type-specific data analysis to complement the cutting-edge single cell transcriptomics assay which still needs time to be fully mature and widely affordable.

In addition, the pipeline depicted in Figure 1d addresses the analysis of FCM data by including a cutting-edge computational method – DAFi [23] – to identify composition/proportions of the cell populations. It makes the pipeline advantageous over the existing data integration approaches when the scientific study includes FCM assay data. The DAFi-based analysis improves not only the reproducibility of the FCM data analysis but also the accuracy of the proportions of the cell populations for improving the downstream `FastMix` inference. It is important to note that the pipeline depicted in Figure 1d can be applied to analyze other types of multimodal datasets for compositional and bulk profiling integrative analysis. For example, multimodal data from metagenomics and metabolomics assays commonly include bulk analysis and composition of microbial communities for which `FastMix` can be applied to identify the community-specific biomarkers. `FastMix` is freely accessible as an open source package at https://github.com/terrysun0302/FastMix.

isResults

Fast unfolding of cell type mixture by integrating multimodal omics data

`FastMix` is primarily designed to integrate two popular assays – flow cytometry and transcriptome profiling – in multi-omics studies. In this scenario, `FastMix` takes in three sets of input data: (i) clinical covariates (denoted as $\mathbf{Clin}$), such as age, sex, treatment group, (ii) cell type proportions measured by flow cytometry assays on heterogeneous populations (denoted as $\mathbf{Cell}$), and (iii) bulk gene expression measured by microarray or RNA sequencing (denoted as Y; Y is a sample-by-gene matrix following the regression model convention). Without `FastMix`, separate regression models can be used to quantify the linear associations between

two out of the three data inputs, i.e., associations between Y and $\mathbf{Clin}$, associations between Y and $\mathbf{Cell}$, or associations between $\mathbf{Cell}$ and $\mathbf{Clin}$. `FastMix` simultaneously studies the associations between all three sets of variables in the following unified linear regression model

$$Y = XW + E,$$

where X is a three-component design matrix, $X \coloneqq (\mathbf{Cell} \quad \mathbf{Clin} \quad \mathbf{Cell} \times \mathbf{Clin})$, W is a matrix of regression coefficients (weights) to be estimated, and E is a matrix of errors. By including the interaction term $\mathbf{Cell} \times \mathbf{Clin}$, `FastMix` makes cell type-specific inferences beyond the bulk level analysis. Typically, the above model is an under-determined system. `FastMix` reduces the model complexity by using the linear mixed effects regression (LMER) techniques. It introduces the gene-specific mixed effects, such that $\beta_{li} = \beta_l + \gamma_{li}$, where β_l is the *fixed effect* of the lth covariate to the entire transcriptome, and γ_{li} is the gene-specific *random effect* of the ith gene. `FastMix` also provides a computational algorithm that is much faster than the traditional expectation-maximization (EM) algorithm to solve large-scale LMER model with high-throughput data. The `FastMix` algorithm is a non-iterative procedure that uses a novel moment-based estimator for the covariance matrix of random effects (denoted as $\hat{B}_T$). Of note, `FastMix` estimates the covariance matrix with outlier trimming and bias correction, therefore it is robust to numerical aberrations induced by outliers and DEGs in the data. Estimates of the fixed effects ($\hat{\beta}_l$) and random effects ($\hat{\gamma}_{li}$) can then be computed using the weighted least squares (WLS) and empirical best linear unbiased predictor (EBLUP) techniques. (See Figure 1b and the Methods section for more details.)

The overarching goal of the FastMix model is to perform DE analyses with respect to the components in the design matrix. Though no classical hypothesis test can be applied to the random effects for theoretical reasons; in practice, FastMix introduces a novel competitive test with quasi-*p*-value to identify DEGs that have significantly larger or smaller predicted random effects $\hat{\gamma}_{li}$ (i.e., cell type-specific effects) to practically rank the importance of genes in the whole transcriptome. To this end, FastMix incorporates a DEG indicator, and assigns the random effects a mixture distribution conditional on the DEG indicator based on empirical Bayes method [24, 25]. For each component of the design matrix, FastMix inference on *random effects* can be interpreted as:

- **Cell** – detection of cell type signature genes that distinguish cell types from each other,
- **Clin** – bulk-level gene expression differential analysis,
- **Cell** × **Clin** – cell type-specific differential analysis, i.e., cell type-specific DEGs.

Note that FastMix is able to incorporate arbitrary weight matrices at the sample level, which can be used to account for the serial correlation in longitudinal studies. We will provide some practical guidance on how to construct such a weight matrix in the Methods section. We refer to this model as the weighted FastMix model. FastMix with known weights is shown to perform better than FastMix without weight if the data are known to fail the independent and identically distributed assumption (please see Supplementary Material, Section 2.3 and 3.5).

In the rest of this section, we illustrated the properties of FastMix in extensive simulation studies. In two real data studies, we applied (weighted) FastMix to carry out cell type-specific

153 inference with a focus on neutrophils in a hepatitis B virus (HBV) vaccine study and a focus on
154 lymphocytes in an influenza infection study. We chose these two cell populations because
155 neutrophils play important roles in pathogenesis of liver diseases and immune responses to
156 HBV vaccines [26, 27]. Also, influenza infection is known to be associated with a relative
157 lymphopenia/neutrophia ratio [28].

158 ***Simulation I: the effect of trimming and robustness estimation of covariance matrix***

159 We designed simulation I to illustrate the advantage of the moment-based covariance matrix
160 estimator. For illustration purpose, we considered two cell types, namely Cell1 and Cell2, whose
161 random effects are γ_{1i} and γ_{2i}, respectively. Figure 1c shows the effect of the proposed
162 embedded robust covariance estimator $\widehat{B}_T$ (green ellipse) for estimating the true covariance
163 matrix between the random effects B (black ellipse). In the simulation, we generated 5000
164 genes (dots); among them, 250 genes were true DEGs (black dots) in the direction of Cell1, i.e.,
165 Cell1-specific DEGs. Note that the existence of true DEGs can be considered as outliers under
166 the null hypothesis [29, 30] that may lead to over-estimation of the covariance matrix. In the
167 estimation procedure of `FastMix`, an initial covariance estimator (red ellipse) was constructed
168 first using the ordinary least squares (OLS) regression technique (see the Methods section),
169 which is non-robust to bias caused by DEGs. Trimming and bias correction techniques were
170 then applied to re-estimate the initial estimated covariance matrix. Simulation I showed that
171 the final estimator $\widehat{B}_T$ was robust to the existence of outliers (DEGs), and accurately
172 recapitulated the true covariance matrix, i.e., the overlay of the green ellipse and the black
173 ellipse.

174 ***Simulation II: comparing performance of*** `FastMix` ***with other regression models***

175 Simulation II was conducted to verify the statistical and computational properties of the
176 proposed method. We generated synthetic gene expression values for 5000 genes and 50
177 subjects. For each subject, we generated three cell proportions (Cell1, Cell2, and Cell3), one
178 continues clinical covariate (Severity), and one categorical clinical covariate (Sex). In the

simulation design, four scenarios were considered: with or without true DEGs, and with or without correlation between random effects. Simulation details are described in the Methods Section.

Here, we systematically evaluated the computational efficiency and accuracy for estimating B, the covariance matrix of random effects. The `FastMix` model is a special case of LMER, with robust estimation of B and bias-correction procedure for fixed effect. The standard implementation of LMER in R is the lme4 package, which uses an iterative expectation-maximization (EM) algorithm to obtain the maximum likelihood estimator for B. The lme4 implementation is very time-consuming for estimating the full covariance matrix B. In practice, users may specify an independent correlation structure, i.e., assuming no correlation between random effects thus B is diagonal. The lme4 implementation with independent assumption (lme4_ind) is more efficient than the default lme4 implementation since much simpler covariance structure is assumed in the model. Similarly, we also implemented the independent assumption for the random effects in the `FastMix` algorithm (`FastMix_ind`). For the completeness of comparison, we reported the time consumed in seconds and mean square error (MSE) for estimating B using lme4_ind, lme4, `FastMix`, and `FastMix_ind` under the four simulated scenarios in Table 1a.

Table 1a showed that, when the random effects were independent and without DEGs, lme4_ind was the theoretically best approach and had the lowest MSE. While the accuracy of lme4_ind and `FastMix_ind` were both at the minimal level (MSE = 0.02 and MSE = 0.04, respectively), `FastMix_ind` used only 2% of the computational time of lme4_ind. When the

random effects were correlated and without DEGs, FastMix had the smallest MSE (0.21) and was more than 300 times faster than the full-pledged lme4 algorithm, which had the second best MSE (0.32). For the simulation scenarios with DEGs, the FastMix implementations were always (with or without correlation) the best performers with the smallest MSEs and used tiny amount of computational time. The lme4-based approaches were not able to obtain accurate estimates, because the maximum likelihood estimator used in lme4 was not robust to effects introduced by DEGs (outliers). On the other hand, FastMix was robust to these effects due to the use of trimming. In Supplementary Material, sections 3.2 and 3.3, we also showed that FastMix greatly reduced the bias in the fixed effect estimation compared to the lme4 approach and other robust covariance estimators [31-33]. Among all the methods compared, FastMix had the most robust performance (Supplementary Material, Tables 1 and 3).

Next, we compared FastMix with ordinary least square (OLS) and Ridge regression for regression coefficient estimation. One primary reason to use LMER instead of the standard OLS regression is to reduce the variability of the estimated regression coefficients. For the same purpose, Ridge regression is also well-known for stabilizing the regression coefficients using regularization. In the second simulation, we compared the accuracy of estimating the gene-specific linear coefficients β_{li}, using FastMix, OLS, and Ridge regressions. We considered the most real simulation scenario, i.e., with correlation and DEGs, for this evaluation. The regularization parameter in ridge regression was selected by the generalized cross-validation (GCV) criterion. Table 1b showed that, FastMix had the smallest total MSE among the three compared methods. In this simulation, the standard deviations of FastMix and Ridge (0.017 and 0.046, respectively) were both much smaller than that of OLS (0.2), suggesting that both

methods could achieve the shrinkage effect, i.e., stabilizing the regression coefficients. A closer look at the biases of individual coefficients suggested that the `FastMix` and OLS estimates could be regarded as practically unbiased. On the other hand, estimates of Ridge regression had large bias, because the L^2 regularization in Ridge regression shrank the estimates toward zero [13], not the fixed effects.

Simulation III: Comparing `FastMix` ***with existing cell type-specific differential analysis method***

Shen-Orr et al. proposed csSAM [34] – a cell type-specific differential analysis method for heterogeneous biological samples using gene expression data and relative cell type frequencies. csSAM is also a regression-based model solved by the standard OLS technique; and the differential analysis is conducted by the established SAM [35] pipeline for bulk gene expression. A major limitation of csSAM is that it only performs two-group comparison, i.e., one binary covariate. For a fair comparison, we redesigned a simpler simulation study for the same number of genes and subjects, which had three cell proportions (Cell1, Cell2, and Cell3) and one binary covariate (Group), to compare the type I error rate, power, and computational time for cell type-specific DEG detection using `FastMix` and csSAM. By design, there were true cell type-specific DEGs for Cell1 and Cell2, but not Cell3; two scenarios with and without correlation between random effects were also considered. Simulation details are described in the Methods Section.

Table 1c-e showed the simulation performance of the two cell type-specific methods. In both correlation and no correlation scenarios, `FastMix` had acceptable type-I error rate (5%~7%),

while csSAM had much higher type-I error rate than `FastMix` in all cases (Table 1c); `FastMix` also had better statistical power (on average 65%) for detecting Cell1-specific and Cell2-specific true DEGs than csSAM (on average 50%) in the simulation (Table 1d). Overall, `FastMix` not only showed superior performance in the simulation results, but also used just 1/10 of the computational time of csSAM (Table 1e).

In Supplementary Material, Section 3.4, we showed that when there were more up-regulated DEGs and fewer down-regulated DEGs in a slightly different simulation, similar patterns of the type-I error, statistical power, and computational efficiency performance were observed for both methods as those shown in Table 1c-e.

`FastMix` *multi-omics integration reveals consistent cell type-specific signature genes with scRNA-seq technology*

We applied `FastMix` to a multimodal study that investigates immune responses to the licensed hepatitis B vaccine – Engerix-B – for the Human Vaccine Project (HVP) [36]. The HVP01 study [37] contains well used assays on whole blood or peripheral blood mononuclear cell (PBMC) samples from adults with wide age range (40-80 years old), including flow cytometry for immunophenotyping, RNA-seq for bulk transcriptomics, and virus neutralization assay for serum antibody titers (anti-HBs). In addition, this study also has scRNA-seq data for immune cells using the Smart-Seq2 [38] protocol, which we would use as the ground truth to validate our `FastMix` results.

Engerix-B requires three doses to reach clinically proven immune protection [39]. In the HVP01 study, there are 15 subjects. After Dose 3, all subjects responded to vaccination, but some had much higher immune protection measured by the anti-HBs titer than others (Supplementary Figure S1). We grouped subjects who had anti-HBs titer >5000 mUI/mL after Dose 3 as high responders (5 subjects), and otherwise low responders (10 subjects). Immunophenotyping by flow cytometry and gene expression by RNA-seq of whole blood and single immune cells were collected at 5 time points (Day 0,1,3,7, and 14). Based on the markers used in the flow cytometry panels, we identified the abundant neutrophils (CD45+ CD66+), non-neutrophils (CD45+ CD66-), and rest populations (Figure 2a) following the DAFi gating hierarchy [23] (see the Methods section). Using all time points, we fitted a weighted `FastMix` model for the bulk RNA-seq gene expression with a design matrix of cell proportions, clinical covariates including response group and age (Supplementary Figure S1), and their interactions.

First, we looked at the DEG list of the main terms, which can be interpreted as the signature genes of each cell population. Broadly speaking, these signature genes are differentially expressed in the specific cell population when compared with the rest of the cell populations. Out of the 13,157 genes available in the processed bulk RNA-seq data, `FastMix` identified 851 signature genes for the neutrophil population, 520 signature genes for the non-neutrophil population, and 30 signature genes for the rest population. Because the rest population is the most heterogenous population, we would expect that not many signature genes could be identified for the rest population that contained a mixture of cell types.

To validate the FastMix signature genes for the well-defined cell population (i.e., neutrophils), we used a completely independent assay data from the unbiased scRNA-seq whole transcriptome expression profiling. We followed a standard scRNA-seq analysis pipeline including low-dimensional embedding of cells on UMAP [40] with cell clusters color labeled by the ground truth cell types based on cell surface markers in FCM sorting, and scRNA-seq DE analysis by a nonparametric hypothesis testing approach [41] for cell type DE gene detection for scRNA-seq data, which were conceptually comparable with the signature genes detected by FastMix. UMAP visualization of the ground truth cell types (Figure 2b) showed a good separation of the neutrophil population from other cell populations. We applied the scRNA-seq DE analysis to detect DEGs between the neutrophils and all other cells, which identified 2,744 neutrophil cell type DE (a.k.a. signature) genes from 58,036 annotated genes in total.

We compared the FastMix and scRNA-seq results of neutrophil signature genes in Figure 2c. The majority (>50%) of the FastMix signature genes overlapped with the scRNA-seq signature genes. Specifically, 72% of the top 100 FastMix signature genes were consistent with the scRNA-seq signature genes. The overlapping rate gradually decreased as we included more top genes in the comparison, meaning that FastMix ranked more “ground truth” (scRNA-seq) signature genes at the top in its DEG list. For pragmatic use, we further selected 365 scRNA-seq signature genes that have substantial fold change (FC), i.e., $|\log FC| > 1$. The Venn diagram (Figure 2d) showed that 39 of the top 100 FastMix signature genes were overlapped with the selected scRNA-seq signature gene list, suggesting that the FastMix signature genes were not only close to the scRNA-seq ground truth, but also contained more practically useful genes in the top ranked genes (16% of the scRNA-seq signature genes passed the logFC threshold vs. 39%

of the top 100 FastMix signature genes passed the logFC threshold). Among the 39 common genes, many of them are highly relevant to both neutrophils and Hepatitis B, e.g., *CXCR1/2* plays an important role [26] in hepatic inflammatory response [42]. The same can be seen for the interferon-induced proteins from *IFIT* and *IFITM* family genes (the 39 genes include *IFIT2, IFITM2, IFITM3,* etc.) Furthermore, we plotted the scRNA-seq expression values of the 39 common signature genes across all cell types (Figure 3a), compared with the bottom genes (Figure 3b) and top (Figure 3c) genes identified by FastMix in violin plots. It is clear to see abundant gene expression in the neutrophils for common and top FastMix signature genes, but almost no expression in the bottom genes.

Identifying cell type-specific interferon signaling pathway genes after Hepatitis B vaccination using FastMix

Next, we compared FastMix and csSAM for identifying the cell type-specific DEGs. With the above FastMix model of HVP01 study, we focused our comparison on the neutrophil-specific DEGs with respect to the response group (since csSAM only performs two-group DE analysis for the cell type-specific SAM model). FastMix identified 495 neutrophil-specific DEGs at 5% false discovery rate (FDR); however, csSAM identified 0 DEG at the same 5% FDR level (the default significance level used in csSAM).

Further, we performed pathway enrichment analysis with top cell type-specific genes ranked by both FastMix and csSAM. Using the step-by-step csSAM, we obtained the 100 top ranked genes based on csSAM estimated FDR. For fair comparison, we also extracted the top 100 FastMix cell type-specific DEGs, and fed both FastMix and csSAM top 100 genes to the

ReactomePA [43] R package for pathway enrichment analysis. FastMix identified 45, 8, and 1 significant cell type-specific pathways for the neutrophils, non-neutrophils, and rest population, respectively (Supplementary Table S1-S3). Figure 4a showed the enriched pathways identified by the top 100 FastMix neutrophil-specific DEGs for high responders. The interferon (IFN) immune signaling pathways were substantially presented in the high responder group, including Interferon Signaling, Interferon alpha/beta signaling, Antiviral mechanism by IFN-stimulated genes, Interferon gamma signaling (the top 4). In comparison, using the top 100 csSAM gene list for each cell type, no enriched pathway was identified for the neutrophil population, only one pathway – neutrophil degranulation – was identified for the non-neutrophil population, and five pathways for the rest population (Supplementary Table S4).

Besides the significant pathways, we also extracted the unique genes that contributed to the enriched pathways from the top 100 FastMix neutrophil-specific DEG list with respect to response group (Figure 4b). In particular, we identified *BST2* (Tetherin/CD317), which is a key host cell defense molecule in response to stimuli from IFN pathway [44, 45]. Traditional understanding of *BST2* expression is with mature B cells and plasmacytoid dendritic cells while it has cell type-dependent variation [46]. Our analysis showed that *BST2* was also expressed in neutrophils, whose increased expression level (estimated linear coefficient = 1.016; please see Methods section for coefficient estimation) was correlated with the high anti-HB levels after Dose 3 of Engerix B.

Inferring cell type-specific temporal pattern from longitudinal data using FastMix

The NIH-funded ImmPort [47] Shared Data portal (www.immport.org/shared/home) shares various immunology studies with the research community. For systems immunology, the activation of immune cell response is a dynamic process; therefore, longitudinal data are commonly collected to investigate how the immune system responds to a certain vaccine or treatment at multiple time points. It is particularly challenging to systematically integrate multi-omics data over a set of time points. We downloaded SDY180 from ImmPort, which employes the systems immunology approaches to investigate immune responses to Influenza (Fluzone® 2009–2010 seasonal influenza vaccine) and Pneumococcal (Pneumovax23® 23-valent pneumococcal vaccine) vaccines [48]. For the Influenza arm, we identified 102 samples that have paired flow cytometry and microarray gene expression data for 12 subjects over 8 or 9 time points; for the Pneumococcal arm, we have 100 samples of 12 subjects over 8 or 9 time points (Supplementary Table S5). The subjects' age range from 20-50 years old; and the nine time points span from 7 days before vaccination to 28 days after vaccination. In the weighted `FastMix` model that we fitted, we included both age and time point as model covariates.

Following the DAFi gating hierarchy (see the Methods section), we identified lymphocytes, granulocytes, monocytes, and rest population from multiple flow cytometry panels for SDY180 (Figure 5a and Supplementary Figure S2). The temporal pattern of the cell proportion changed over time for the Influenza arm are shown in Figure 5b. With only the flow cytometry data, we noticed that the proportion of lymphocytes had a substantial drop on Day 1 after vaccination and was recovered by Day 3.

Using a simple pre-post (i.e. between two days) comparison, the original SDY180 study [48] curated an interferon module, namely M1.2, that includes genes showing significant *global* changes in blood transcript abundance between the baseline Day 0 and Influenza Vaccine Day 1. In bulk level, representative genes (*CXCL10*, *IFIT1*, and *LAMP3*) in the M1.2 module showed consistently a peak in gene expression on Day 1 after vaccination (Figure 5c and Supplementary Figure S3), confirming the global finding; the temporal plots also show that the bulk expression of these genes fell back to around the baseline level on and after Day 3.

Applying `FastMix` to the Influenza arm, we could further designate the *specific* cell population that are associated with the temporal activation of these interferon genes. Among the 24 gene in M1.2 module, `FastMix` identified 22 genes with highly significant p-values (< 0.05) for lymphocyte-specific differential expression (Figure 5d and Supplementary Table S6); the top 9 M1.2 genes ranked in the top 1% (out of 10732 genes) of the lymphocyte-specific DE list. However, the majority of the M1.2 genes showed no significance in granulocytes and monocytes (Supplementary Table S6). These results strongly indicate that the activation of the interferon module is lymphocyte-specific: the differential expression of interferon signaling genes are driven by the up-regulation of the lymphocyte-specific expression. Though the proportion of lymphocytes decreased on Day 1 after vaccination (Figure 4b), the bulk gene expression of M1.2 genes increased on Day 1 (Figure 4c). The `FastMix` statistical inference precisely linked the temporal changes in Figure 4b for lymphocytes and Figure 4c for those interferon-stimulated genes. Furthermore, `FastMix` produced positive estimated coefficients for lymphocytes for all M1.2 genes (Supplementary Table S7), confirming the up-regulation of the cell type-specific gene expression.

We also looked at the cell type and age interaction terms. The lymphocyte-specific *p*-values w.r.t. age for the M1.2 interferon module genes showed very strong significance (23 out of 24 significant *p*-values) (Figure 4e and Supplementary Table S8), whose coefficient estimates showed negative association between the subject age and lymphocyte-specific expression (Supplementary Table S7 and Supplementary Figure S4).

For completeness of method comparison, a plausible csSAM analysis would be only to compare the pre- and post-vaccination groups for cell type-specific DEGs due to technical limitations. Even for the simple two-group test, the csSAM approach is suboptimal, because there is no appropriate way to handle the within-subject correlation structure in the multiple time points. Therefore, no significant cell type-specific DEGs w.r.t. the pre- and post-vaccination groups were identified at the 5% FDR level. Using the top 100 csSAM gene list for each cell type, no enriched pathway was identified for any cell type, suggesting that csSAM is inadequate for performing cell type-specific DE analysis with complex study design.

Lastly, applying `FastMix` to the Pneumococcal arm, only one significant *p*-value was obtained in Figure 4c-d (Supplementary Table S6 and S8). Clearly, the lymphocyte-specific interferon activation was only observed in the Influenza arm, but not the Pneumococcal arm, agreeing with the existing knowledge in PBMC samples [48]. The Pneumococcal arm may serve as the "true negative" for our method validation; and the non-significant results showed the "specificity" of the `FastMix` method.

Discriminant Analysis after `FastMix`

Because `FastMix` is designed to take multiple types of input variables including clinical parameters, it can be used to identify the relationship between the independent (e.g., subject demographics) and dependent variables (e.g., response to a vaccination). For example, in our experiment using the HBV vaccination data in the HVP01 study, when the subjects were grouped into responding and non-responding, `FastMix` could calculate four scores for discriminating purpose: (a) single_score, an 1-dimensional score based on all input genes; (b) single_sparse_score, a 1-dimensional score based on genes with significant interactions with the response; (c) multi_score, an n-dimensional score based on all genes; and (d) multi_sparse_score, a multivariate score based on genes with significant interactions with the response (see Supplementary Material, Section Discriminant Analysis after `FastMix`, for technical details). Figure 4c-d showed that the discriminative scores (for straightforward illustration, single_sparse_score was used in the Figure) can be plotted to identify whether age is an informative factor in the discriminative analysis (i.e., classification of responding vs non-responding subjects). We can clearly see the significant (Wilcoxon *p*-value = 1.9e -16) difference between the responding (yellow) and non-responding (grey) groups when age was included in the analysis (Figure 4c), while the difference was unclear without the age (Figure 4d). This tells us that age is an important variable that is highly relevant in host immune response to the HBV vaccine.

Discussion

Using extensive simulation studies, we showed that: (i) regression coefficients estimated by `FastMix` had comparable mean squared errors (MSE) as those computed from lme4 - the

reference implementation of LMER model based on EM algorithm, (ii) depending on settings, `FastMix` was at least 25 times, and in some cases more than 300 times faster than lme4, and (iii) `FastMix` was substantially more accurate (measured by MSE) than the ordinary least squares (OLS) and Ridge regression. See Table 1 for more details.

In addition, we compared the type-I error rate and statistical power of `FastMix` for cell type-specific differential expression (DE) analysis with an existing pipeline, csSAM [34], using both simulations and real data. `FastMix` achieved slightly better statistical power with much lower type-I error than csSAM, using about 10% of csSAM's run time (see Table 1c-e). We applied the `FastMix` pipeline to analyze the multimodal data from two clinical studies [37, 48] that measured host responses to three different vaccines (influenza, pneumococcal, and hepatitis B). Input data included bulk gene expressions, FCM, as well as clinical covariates including vaccine responding groups defined by serum antibody titers as well as time points for multiple vaccine doses. A common bottleneck in evaluating multimodal data integration methods using real data is the lack of ground truth. Performance assessment of many preexisting methods relies on subjective interpretation of their data integration results using existing knowledge. In contrast, we addressed this issue by using single cell RNA-seq (scRNA-seq) data available in one of the studies as an objective gold standard. Excitingly but not surprisingly, DEGs selected by `FastMix` overlapped significantly with those selected by the cutting-edge analysis of the scRNA-seq data. On the other hand, `FastMix` seemed to be able to select biologically important genes for neutrophils that were missed by the scRNA-seq analysis.

The general contribution of `FastMix`, from a statistical perspective, is the extension from the traditional pairwise linear associations between multi-omics data types into a multiple regression model with both fixed and random effects (LMER) that can take multiple types of inputs simultaneously and infer cell type-specific biomarkers as well as signature genes based on cohorts defined by experiment variates or clinical parameters. One roadblock for realizing the LMER analysis in practice is the complex and slow iterative EM algorithm for estimating the regression parameters. We solved this issue by an efficient moment-based method that achieved similar accuracy as EM but using only a fraction of its run time (Table 1a). Note that this method includes both trimming and the corresponding bias-correction, so that the estimated covariance structure (used in the LMER) is robust to outliers and practically unbiased. It is important to note that the LMER construction in `FastMix` also addressed the collinearity issue in an interpretable way, without needing a black-box non-linear transformation used in many of the existing multi-omics data integration approaches. Inspired by competitive tests used in gene set enrichment analyses [49-51], we designed a quasi-*p*-value to rank and select genes with significantly larger/smaller random effects (cell-type-specific effects) than most other genes. We believe this approach may be applicable in other situations.

`FastMix` provides an end-to-end solution for integrative analysis of flow cytometry (FCM) data and bulk transcriptomics data. FCM and transcriptomics are commonly used in immunology studies. Among the 1924 experiments in the 495 studies collected by US NIAID's ImmPort database (https://immport.org/shared/home) as of June 2021, the top two assay types are FCM (706; 36.7%) and transcription profiling (213; 11.1%). However, existing solutions for integrating data from transcriptomics and FCM assays for cell type-specific immune profiling

are suboptimal. FCM data analysis mainly relies on subjective manual gating analysis, which is difficult to be integrated with other computational modules. Identification of cell type-specific signature genes and DEGs relies on predefined marker genes in the transcriptomics data, without utilizing the FCM data that provide canonical phenotypic definitions of the cell types. We previously developed a computational method – DAFi [23] – to identify cell populations from FCM data in an objective way, which produces more accurate proportions of cell populations in the biological sample than the subjective manual gating analysis [23, 52]. Combining DAFi and `FastMix` (Figure 1d) produces a novel unbiased solution for immunologists to identify cell-based biomarkers, including DEGs and cell populations with significantly different abundances between cohorts, from the FCM and transcriptomics data.

Besides synthetic data, using scRNA-seq data provides ground truth for assessing the performance of `FastMix`. The consistency between FastMix and scRNA-seq from our experiment (Figure 3) showed that `FastMix` can be used to infer cell type-specific knowledge from bulk transcriptomics and FCM data. However, FastMix-identified biomarker genes are also complementary to results of the scRNA-seq data analysis. For example, `FastMix` identified the neutrophil-specific genes *MMP9* and *RSAD2*/Viperin (Figure 3c), which were not found in the scRNA-seq data analysis (Figure 3a). MMP9 is a regulatory factor in neutrophil migration [53] and Viperin is an important anti-viral protein induced in neutrophils [54] (Figure 3c). Also, `FastMix` identified the IFIT gene family members (*IFIT1, IFIT2*, and *IFIT3*) that can limit the HBV replication [55]. Basically, `FastMix` provides an *in-silico* alternative when scRNA-seq data is unavailable or unreliable. Besides inferring the cell type-specific signature genes, `FastMix` can produce discriminative scores of model variables, which quantify the contributions of

model variables to the sample classification. This is a unique and novel feature that previous models have not provided.

The main limitation of `FastMix` is that it does not solve the well-known problem for inferring characteristics of rare cell populations from bulk assay data. When the proportion of a cell population is small, its contribution to the bulk gene expressions is easily overwhelmed by the abundant cell populations. Even a minor change in gene regulation of the major cell types can dominate the variation of the bulk gene expressions. This challenge can potentially be solved if there are replicates of the same measurement, which are unfortunately usually unavailable in most biomedical studies. Ideally, scRNA-seq, bulk transcriptomics, as well as FCM, when available, can be integrated together for achieving the optimal performance for identifying cell-based DEGs and other biomarkers for both abundant and rare cell types in the whole cell type hierarchy. The estimation can also benefit from longitudinal (and repeated) measurements when they are available, which will be investigated in our future work

Applications of `FastMix` can be easily extended to include metagenomics and metabolomics data. For example, a straightforward application of `FastMix` is to identify genetic factors across species to explain variation of metabolomic profiles based on microbial community composition data. `FastMix` allows us to do “reverse engineering” from observed metabolite abundances in diverse microbiome communities to infer species-specific contributions.

Figure Legend

Figure 1. `FastMix` **schematics and analytical pipeline. (a)** `FastMix` takes three input data matrices: a bulk gene expression matrix, a matrix of cell type proportions, and a matrix of clinical covariates (both continues and categorical). **(b)** Flow chart of key steps of `FastMix`. (Details please refer to complementary material.) i. The `FastMix` model utilizes linear mixed-effects regression (LMER) model and mixture distribution to construct a unified regression model for the three data inputs. ii. Reparametrize the `FastMix` model by vectorization and Kronecker product so the data can be analyzed in a unified LMER model. iii. The `FastMix` algorithm gains computational efficiency through using a novel moment-based estimator of the covariance matrix $\hat{B}^{(0)}$, followed by solving for the fixed effects estimate $\widehat{\boldsymbol{\beta}}^{(1)}$ and the random effects estimate $\widehat{\boldsymbol{\gamma}}_i^{(1)}$, both of which depend on $\hat{B}^{(0)}$. iv. In `FastMix`, DEG identification is viewed as an outlier detection problem. It uses a trimming technique to improve the robustness due to the existence of DEGs (outliers). v. After trimming, re-estimate the variance-covariance matrix using the robust estimator $\hat{B}_{\mathrm{T}}$ with bias correction, followed by re-estimating $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}_i$ using $\hat{B}_{\mathrm{T}}$. vi. `FastMix` performs hypothesis test and constructs quasi-*p*-values that indicate the significance of cell type-specific DEGs. **(c)** Using trimming improves the estimation of the covariance matrix. Axes are random effect signals of two cell populations (Cell1 and Cell2); dots (grey and black) are simulated data of 5000 genes, among which, 250 genes are true DEGs in the Cell1 direction (black dots). Three ellipses are the density contour curves that represent the 95% confidence region of the centered data distribution with covariance matrices of: B that is the true covariance matrix shown in black, $\hat{B}^{(0)}$ that is the initial non-robust covariance

estimator shown in red, and $\hat{B}_T$ that is the robust covariance estimator based on trimming shown in green. Due to the existence of the true DEGs (outliers), $\hat{B}^{(0)}$ overestimated the true covariance matrix. The trimming-based estimator $\hat{B}_T$ is very close to the true covariance matrix. **(d)** Sample analytical pipeline for cell type-specific differential analysis between disease and control groups by integrating flow cytometry data and bulk RNA-seq data using two newly developed computational algorithms DAFi and `FastMix`.

Table 1. Simulation performance. (a) Comparison of `FastMix` implementations (`FastMix` with independence assumption, i.e., `FastMix_ind`, and default `FastMix` with no assumption on the covariance matrix) and lme4 implementations (lme4 with independence assumption, i.e. lme4_ind, and default lme4 with no assumption on the covariance matrix) for estimating B, the covariance matrix of random effects, in linear mixed effects regression (LMER). Four simulation scenarios are considered: with or without true DEGs, and with or without correlation between random effects. Mean computational time and mean MSE are reported. Computational time is reported in seconds, and estimation accuracy is reported in mean squared error (MES). MSE is defined as $\sum_{i=1}^{p}\sum_{j=1}^{p} 1/p^2\left(\widehat{B}_{ij} - B_{ij}\right)^2$. Simulations are repeated 200 times. **(b)** Comparison of `FastMix` with ordinary least squares (OLS) and Ridge regression for regression coefficient, β_{ij}, estimation. The first row is the mean MSE (standard deviation in brackets) defined as $1/(mp)\sum_{i=1}^{m}\sum_{j=1}^{p}\left(\widehat{\beta}_{ij} - \beta_{ij}\right)^2$. The other rows are the mean bias (standard deviation in brackets) of each fix effect coefficient estimation. Simulations are repeated 200 times. All results are reported after multiplying by 100 for better readability. **(c-e)**

552 Mean (standard deviation in brackets) of type-I error rate **(c)**, statistical power **(d)**, and
553 computational time reported in seconds **(e)** of csSAM and `FastMix` for cell type-specific DEG
554 detection, in the same simulation scheme repeated 200 times. The simulation design includes
555 independent random effects (i.e., cor = 0) and correlated random effects (i.e., cor = 0.5). True
556 cell type-specific DEGs are only assigned in cell1 and cell2 in the simulations. Type-I error rate
557 and statistical power are reported in percentage (%).

558 **Figure 2.** `FastMix` **and scRNA-seq results for HVP01 study. (a)** DAFi gating strategy to identify
559 singlets, leukocytes, live leukocytes, CD66- CD45+ population (parent: live leukocytes), and
560 CD66+ CD45- population (parent: live leukocytes). **(b)** UMAP visualization of scRNA-seq cell type
561 clusters. Cells are colored by cluster labels derived by flow cytometry panels. **(c)** Overlapping of
562 the 851 (out of 13157 total genes) `FastMix` neutrophil-specific signature genes and the 2744
563 scRNA-seq neutrophil signature genes available in the bulk RNA-seq data. **(d)** Venn diagram of
564 the overlapping between the top 100 `FastMix` neutrophil signature genes and the scRNA-seq
565 neutrophil signature genes with |logFC| > 1. The 39 common genes are shown in the text box.

566 **Figure 3. Expression of neutrophil-specific signature genes in the scRNA-seq experiment. (a)**
567 The 39 common signature genes identified by `FastMix` and scRNA-seq analysis (same in
568 Figure 2c). **(b)** The bottom 39 genes ranked by `FastMix`. **(c)** The top 39 genes ranked by
569 `FastMix`.

570 **Figure 4. Pathway enrichment analysis for HVP01 study. (a)** Enriched pathways identified by
571 the top 100 `FastMix` neutrophil-specific DEGs for high responders. **(b)** Unique genes from the

572 CD45pCD66p.Response (i.e., neutrophil and high response) interaction DEG list that are
573 identified in the enriched pathways in (d).

574 **Figure 5.** `FastMix` **analysis for SDY180. (a)** DAFi gating strategy to identify lymphocytes,
575 granulocytes, and monocytes, CD45+ CD14- (parent: granulocytes and monocytes), CD45+
576 CD14+ (parent: granulocytes and monocytes), granulocytes (parent: CD45+ CD14-), and
577 monocytes (parent: CD45+ CD14+). **(b)** Boxplots of cell proportions (lymphocytes, granulocytes,
578 monocytes) over time in the Influenza vaccine study. **(c)** Boxplots of bulk expression levels of
579 interferon-stimulated genes (e.g., *CXCL10, IFIT1, LAMP3*) over time in the Influenza vaccine
580 study. Red box: matching temporal pattern change of lymphocytes proportion and bulk gene
581 expression. **(d)** Heatmap of -log10-transformed *p*-values for lymphocyte-specific differential
582 expression for the interferon module genes in both Influenza and Pneumococcal study arms. **(e)**
583 Heatmap of -log10-transformed *p*-values for lymphocyte-specific differential expression w.r.t.
584 age for the interferon module genes in both Influenza and Pneumococcal study arms.

Methods

Cell type-specific inference based on bulk tissue modeling

In the regression model framework, composite tissue data can be modeled as

$$Y_{ji} = \sum_{k=1}^{K} \mathbf{Cell}_{jk} \cdot b_{kij} + \epsilon_{ij}, \;\; i = 1, \dots, m, \; j = 1, \dots, n. \#(1)$$

Specifically, Y_{ji} is the observed bulk expression of the ith gene and jth sample; $\mathbf{Cell}_{jk}$ is the observed proportion of the kth cell type (or cell population) in the jth sample, and b_{kij} is the cell type-specific expression level of the ith gene and jth sample contributed solely by the kth cell type, and ϵ_{ij} is the uncertainty in measuring $\mathbf{Cell}_{jk}$ and Y_{ji}. Many downstream analyses are focused to associate b_{kij} (cell type-specific gene expression) instead of Y_{ji} (bulk gene expression) to the clinical metadata, which can be modeled as

$$b_{kij} = \beta_{ki} + \sum_{p=1}^{P} \mathbf{Clin}_{jp} \cdot a_{ipk} + e_{kij}. \#(2)$$

Here β_{ki} is the baseline expression level of the ith gene in the kth cell type, $\mathbf{Clin}_{jp}$ is the pth clinical covariate associated with the jth sample, a_{ipk} quantifies the linear association between the pth clinical covariate and the ith gene specific to the kth cell type. In this context, cell type-specific differential analysis can be conducted by testing the following hypotheses

$$H_{0,ipk}: a_{ipk} = 0, \text{ v.s. } H_{1,ipk}: a_{ipk} \neq 0. \#(3)$$

598 For the kth cell type, the ith gene is a cell type-specific DEG with respect to the pth clinical
599 covariate if the *p*-value from the Equation (3) hypothesis test is statistically significant.

600 Based on the above framework, one straightforward approach to perform cell type-specific
601 analysis would consist of two stages: (i) apply *in silico* algorithm, such as deconvolution, to
602 estimate $\hat{b}_{kij}$; and (ii) apply a suitable DE analysis to associate $\hat{b}_{kij}$ with the clinical data.
603 However, there is a major challenge of this approach, i.e., Equation (1) is a typical "large *p*,
604 small *n*" problem because there are approximately Knm unknown parameters (b_{kij}) to be
605 estimated from only nm observations (Y_{ji}). While many computational methods such as
606 nonnegative matrix factorization [56-59], regularization [60], and Bayesian methods [61-64],
607 are used to obtain approximate solutions an under-determined system for deconvolution [65],
608 the bias and variance of the estimated $\hat{b}_{kij}$ are inevitably large, which will consequently impact
609 the accuracy of the downstream DE analysis.

610 `FastMix` ***model***

611 We propose to jointly model the two-stage analysis in one unified regression model by
612 combining Equation (1) and Equation (2)

$$\begin{aligned} Y_{ji} &= \sum_{k=1}^{K} \mathbf{Cell}_{jk} \cdot \left(\beta_{ki} + \sum_{p=1}^{P} \mathbf{Clin}_{jp} \cdot a_{ipk} + e_{kij} \right) + \epsilon_{ij} \\ &= \sum_{k=1}^{K} \mathbf{Cell}_{jk}\, \beta_{ki} + \sum_{k=1}^{K} \sum_{p=1}^{P} \mathbf{Cell}_{jk}\, \mathbf{Clin}_{jp} \cdot a_{ipk} + \tilde{\epsilon}_{ij} \,. \end{aligned} \qquad \#(4)$$

Here $\tilde{\epsilon}_{ij}$ is the combined error term, $\tilde{\epsilon}_{ij} = \epsilon_{ij} + \sum_{k=1}^{K} \mathbf{Cell}_{jk}\, e_{kij}$, a_{ipk} that quantifies the interaction between the kth cell type and the pth clinical covariate. To model the direct association between the bulk gene expression and clinical covariates, we further add a main term $\mathbf{Clin}_{jp}$ to Equation (4), which is commonly used in the traditional bulk DE analysis. Therefore, the unified model includes main terms $\mathbf{Cell}_{jk}$ and $\mathbf{Clin}_{jp}$, and their interaction term $\mathbf{Cell}_{jk}\mathbf{Clin}_{jp}$, which can be stated in the standard multivariate linear regression model $Y = XW + E$, or explicitly

$$Y_{ji} = \sum_{l=1}^{L} X_{jl}\,\beta_{li} + \epsilon_{ij}. \#(5)$$

Here X_{jl} is an element in matrix $X := (\mathbf{Cell} \quad \mathbf{Clin} \quad \mathbf{Cell} \times \mathbf{Clin})$, which has n rows and $L = K + P + KP$ columns. We call each column of matrix X a linear predictor, and β_{li}'s the linear coefficients. Model (5) is a combination of both bulk and cell type-specific DE analyses. This unified modeling approach allows us to bypass the error-prone and computationally intensive parameter estimation stage (Equation (1)), and only focus on the more biological interpretable DE stage for the associations between the three types of variables (cell types, clinical covariates, and their interactions) with the bulk/cell type-level gene expression.

One important advantage of the unified model approach is that with reasonably large sample size ($n > L$), Model (5) no longer has the "large p, small n" problem because there are only mL unknown parameters (β_{li}) to be estimated with nm observations (Y_{ji}). `FastMix` does not explicitly estimate the deconvoluted cell type-specific expression values; rather, it uses joint modeling and techniques to be introduced in the following sections to implement a fast-

algorithmic solution for the large-scale unified model (Equation (5)) for bulk and cell type-specific DE analyses.

Common strategies to solve a large-scale model such as Equation (5) is to apply regularizations, a.k.a. penalized regressions such as ridge [13], LASSO [14], and elastic-net [15], to increase the stability and prediction accuracy of the original regression model. However, these techniques have two drawbacks: (i) they shrink the estimated linear coefficients toward zero and create nontrivial bias; and (ii) the best penalty parameter(s) are typically trained by time-consuming cross-validation (CV) procedures, which may not always be computationally feasible for high-throughput data analysis. As an alternative, we propose to use linear mixed effects regression (LMER) to reduce model complexity. Specifically, we assume a two-component decomposition of the unknown linear coefficient such that

$$\beta_{li} = \beta_l + \gamma_{li}, \#(6)$$

where β_l is the *fixed effect* of the lth linear predictor to the entire transcriptome, and γ_{li} is the gene-specific *random effect* associated with the lth linear predictor. By combining Equations (5) and (6), the `FastMix` model with mixed effects is

$$Y_{ji} = \sum_{l=1}^{L} X_{jl}\,(\beta_l + \gamma_{li}) + \epsilon_{ij}. \#(7)$$

Using the LMER model fitting approach does not need to train hyperparameters with intensive CV procedures. Also, the LMER model can shrink the estimated gene-specific linear coefficients

toward the fixed effects (i.e. average of the entire transcriptome) instead of zero [20], thereby achieving comparable variance-reduction effects with less bias.

The next step is to model DEGs and non-DEGs (NDEGs) based on Equation (7). In most practical cases, the majority of the genes are NDEGs; and only a small fraction of the genes are truly DEGs that may be used as biomarkers for specific biological conditions. In this regard, we propose to approach the DEG identification problem as an outlier detection problem in more general setting. Specifically, we propose to model the gene-specific random effects, γ_{li}'s, using a mixture distribution and adapt a nonparametric empirical Bayes method [24, 25] to conduct per-gene statistical inference.

Let ι be a binary indicator for DEG ($\iota = 1$) and NDEG ($\iota = 0$). The prior probability of a gene being NDEG or DEG is $P(\iota = 0) = \pi_0$ or $P(\iota = 1) = 1 - \pi_0$, respectively. The mixture distribution of the multivariate vector $\boldsymbol{\gamma}_i = (\gamma_{li}, l = 1, \cdots, L)'$ is

$$\boldsymbol{\gamma}_i \sim f(\mathbf{x}), \qquad f(\mathbf{x}) = \pi_0 f_0(\mathbf{x}) + (1 - \pi_0) f_1(\mathbf{x}), \#(8)$$

where $\mathbf{x}$ is a dummy variable, $f_0(\cdot)$ is the component distribution for NDEGs and $f_1(\cdot)$ is the component distribution for DEGs. Furthermore, it is reasonable to assume that: (i) $\pi_0 \gg 1 - \pi_0$, i.e. most of the genes are NDEGs; (ii) the conditional distribution of the multivariate vector $\boldsymbol{\gamma}_i$ given $\iota = 0$ is a L-dimensional normal random vector centered at the origin with covariance matrix B; and (iii) let $D_\alpha \subset \mathbb{R}^L$ be the confidence region of $f_0(\cdot)$ centered at the origin with probability $1 - \alpha$ with a relatively large α, then

$$P(\boldsymbol{\gamma}_i \in D_\alpha | \iota = 1) \ll P(\boldsymbol{\gamma}_i \in D_\alpha | \iota = 0). \#(9)$$

Intuitively, Equation (9) implies that compared with NDEGs, the DEGs can be viewed as "outliers" (Figure 1c). No parametric assumptions are applied to $f_1(\cdot)$. From the above assumptions, the marginal distribution for the nonparametric empirical Bayes method is

$$f(\mathbf{x}|\iota) = \begin{cases} f_0(\mathbf{x}) \coloneqq \phi(\mathbf{x}|\mathbf{0}, B), & \iota = 0 \\ f_1(\mathbf{x}), & \iota = 1 \end{cases} \#(10)$$

where $\phi(\cdot\,|\mathbf{0}, B)$ is the density function of a multivariate normal random vector defined on $\mathbb{R}^L$ with zero mean and covariance matrix B.

In summary, Equations (7), (8) and (10) specify the full `FastMix` model of the unified pipeline for cell type-specific DE analysis:

$$Y_{ji} = \sum_{l=1}^{L} X_{jl}\,(\beta_l + \gamma_{li}) + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma_\epsilon^2) \text{ for the LMER model;}$$

$$\boldsymbol{\gamma}_i \sim f(\mathbf{x}),\ f(\mathbf{x}) = \pi_0 f_0(\mathbf{x}) + (1-\pi_0) f_1(\mathbf{x}) \text{ for mixture distribution; and}$$

$$f(\mathbf{x}|\iota) = \begin{cases} f_0(\mathbf{x}) \coloneqq \phi(\mathbf{x}|\mathbf{0}, B), & \iota = 0 \\ f_1(\mathbf{x}), & \iota = 1 \end{cases} \text{ for nonparametric empirical Bayes.}$$

Computationally efficient `FastMix` ***algorithm***

The `FastMix` model has many theoretical advantages by using a LMER model; however, fitting such a large LMER model with high-throughput data is still computationally challenging. To reduce the high computational cost of fitting large LMER models by conventional methods, such as the iterative expectation-maximization (EM) algorithm [22], we design a highly efficient

algorithm with a novel robust moment-based covariance estimator, which avoids the iterations and convergence process, thus largely saves the computational time.

In the following subsections, high-level descriptions of the key steps are provided here. All technical details, including the derivations, proofs, and step-by-step procedures are provided in Supplementary Material.

Vectorization and Kronecker product

The `FastMix` LMER model can be concisely represented in vectorization form using Kronecker product [66]

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

$$\mathbf{X} := 1_m \otimes X = \begin{pmatrix} X \\ \vdots \\ X \end{pmatrix}, \qquad \boldsymbol{\gamma} := \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_m \end{pmatrix}, \qquad \mathbf{Z} := I_m \otimes X = \begin{pmatrix} X & & \\ & \ddots & \\ & & X \end{pmatrix}. \#(10)$$

Note that $\mathbf{X}$ is $N \times L$-dimensional, $\mathbf{Z}$ is $N \times mL$-dimensional, and $\boldsymbol{\gamma}$ is $mL \times 1$-dimensional, where $N = mn$ is the total number of observations. In this form, $\mathbf{Y}$ is a long vector of length N, by column-wise stacking of the bulk gene expression matrix; $\boldsymbol{\beta}$ is the long vector of linear coefficients to be estimated of the same length; and $\boldsymbol{\epsilon}$ is the corresponding error vector. Now, it is clear that all three types of high- and low-dimensional data are neatly combined in the form of a standard LMER. The vectorized notions (in bold face and non-italic) will be used in the subsequent estimation derivations; it also helps to speed up the implementation of the algorithm.

Moment-based estimation

An initial estimation of the linear coefficients $\widehat{\boldsymbol{\beta}}_i^{(0)} = (\hat{\beta}_{li}, \mathrm{l} = 1, \cdots, \mathrm{L})'$ can be obtained through fitting the multivariate linear regression in Equation (5) using the ordinary least squares (OLS) criterion. $\widehat{\boldsymbol{\beta}}_i^{(0)}$ can be considered as a crude approximation of $\boldsymbol{\gamma}_i$, which contains information about the covariance matrix of $\boldsymbol{\gamma}$. Denote the sample covariance matrix of $\widehat{\boldsymbol{\beta}}_i^{(0)}$ as $\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}}^{(0)}} \in M_{L\times L}$. Even for NDEGs ($\iota = 0$), $\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}}^{(0)}}$ is not an unbiased estimator of B. Its conditional expectation can be derived as follows

$$E\left(\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}}^{(0)}} | \iota = 0\right) = B + \sigma_\epsilon^2 (X'X)^{-1}. \#(11)$$

Based on Equation (11), and the assumption that most genes are NDEGs ($\iota = 0$), we propose the following moment-based estimator for an initial estimation of B

$$\widehat{B}^{(0)} := \widehat{\Sigma}_{\widehat{\boldsymbol{\beta}}^{(0)}} - \widehat{\sigma}_\epsilon^2 (X'X)^{-1}. \#(12)$$

Based on the initial OLS-based estimates, the standard closed-form solutions of solving LMER model are used to obtain the first set of fixed effects and random effects estimates for Equation (6), denoted in $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\gamma}}_i^{(1)}$, respectively. The weighted least squares (WLS) estimator is used to compute $\widehat{\boldsymbol{\beta}}^{(1)}$; and the empirical best linear unbiased predictor (EBLUP) is used to compute $\widehat{\boldsymbol{\gamma}}_i^{(1)}$. Both estimates depend on $\widehat{B}^{(0)}$, the initial moment-base estimate of B.

Trimming for potential DEGs

712 One of the assumptions of the `FastMix` model is that there is a small subset of genes that are
713 DEGs. The existence of these potential DEGs may affect the accuracy of the initial estimates of
714 B, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}_i$ in the previous step. We designed a three-step procedure to detect and remove
715 those potential DEGs.

716 Briefly speaking, when the DEGs are present, $\widehat{\boldsymbol{\gamma}}_i^{(1)}$ no longer follows a multivariate normal
717 distribution. A DEG for one covariate may very likely be an NDEG for another covariate; also, it
718 is possible that a subset of covariates is not associated with any DEG (we call them
719 uninformative covariates). The three-step procedure includes: (i) use a standard normality test,
720 the Shapiro-Wilk test, to separate the informative and uninformative covariates; (ii) select the
721 subset of the standardized linear coefficient estimates pertain to the informative covariates and
722 calculate the Mahalanobis distance between the sub-vector and the origin (its theoretical
723 mean), i.e., this quantity quantifies how likely a gene is an outlier among all genes. Under the
724 assumption that most genes are NDEGs, the distance metrics of all genes form approximately a
725 chi-squared distribution; and (iii) use a pre-defined trim level (denoted as α, which is a user-
726 defined tuning parameter with default value $\alpha = 0.5$) to select potential DEGs based on the
727 distance metric following the chi-squared distribution.

728 After trimming of the potential DEGs that break the normality assumption, the remaining genes,
729 denoted as $S_0 \subseteq \{1, \dots, m\}$, will be used to refine the estimation in previous steps.

730 ***Re-estimation and bias correction***

Now, re-estimate B based on the trimmed gene list S_0. However, the trimming procedure inevitably introduces bias to the sample covariance matrix, because removing genes with large Mahalanobis distance artificially reduces the sample covariance matrix computed from the remaining genes. To correct for this bias, we consider a truncated chi-squared distribution and constructed a moment-based bias-correction for B as follows

$$\widehat{B}_T := \Lambda^{-1/2}\,\widetilde{\Lambda}^{1/2}\,\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}}^{(0)}}\,\widetilde{\Lambda}^{1/2}\,\Lambda^{-1/2} - \widehat{\sigma}_\epsilon^2\,(X'X)^{-1}. \#(13)$$

For the fixed effect, we utilize the similar idea of trimming to de-bias the fix effect estimation if there are any potential unbalanced DEGs (see Supplementary Material Section 2.2 for more details). The final fixed effects estimate $\widehat{\boldsymbol{\beta}}$ and random effects estimate $\widehat{\boldsymbol{\gamma}}_i$ are re-computed using $\widehat{B}_T$. Figure 1c illustrates the advantage of the trimming and re-estimation procedures.

<u>Hypothesis test and quasi-p-value</u>

Traditionally, DE analysis can be performed through hypothesis testing strategies on the linear coefficients (such as Equation (3)). There are mature regression F- and t-tests for the fixed effects in LMER models, but not for the random effects because they are considered as realizations of random variables (i.e. not unknown parameters) [67]. To overcome this theoretical challenge, we developed a practical *p*-value-like quantity (called "quasi-*p*-value") through analogy, to identify genes that have significantly larger or smaller predicted random effect with a given covariate. The quasi-*p*-value is defined as

$$\widehat{p}_{li} := 1 - \Phi\left(\frac{|\widehat{\gamma}_{li}|}{\widehat{\sigma}_{\widehat{\gamma}_l}}\right), \#(14)$$

where $\Phi(\cdot)$ is the standard normal distribution function. Note that $\widehat{p}_{li}$ is not a "true" *p*-value because ι is a random variable, not a parameter, in the LMER model; so, we cannot test hypotheses $H_0: \iota = 0$ (i.e., NDEG) versus $H_1: \iota = 1$ (i.e., DEG) in the classical sense. In practice, the quasi-*p*-value for the random effects can be used as a practical criterion to rank and select genes with strong association with the lth covariate, which are the central inference output from the `FastMix` model. The random effects results can be interpreted as the cell type marker, bulk-level, and cell type-specific DE analyses as introduced in the Results section.

For the completeness of the model output, the hypotheses for the fixed effects are

$$H_{0,l}: \beta_l = 0, \text{ versus } H_{1,l}: \beta_l \neq 0. \#(15)$$

The test statistic is $t_{\beta_l} := \frac{\widehat{\beta}_l^{(1)}}{\widehat{\sigma}_{\widehat{\beta}_l^{(1)}}}$, which follows a t-distribution with degrees of freedom approximated by the Satterthwaite's method [68]. The fixed effects tests are not gene-specific; instead, these results can be interpreted as whether a clinical covariate has a statistically significant impact on the whole transcriptome.

Weighted `FastMix` ***model***

So far, the `FastMix` model assumes independent and identically distributed (i.i.d.) samples. Sometimes, *a priori* knowledge may be available to weigh some samples over others; or in a longitudinal study, repeated measurements are not i.i.d. samples and they tend to have block interchangeable covariance structure. Such information can improve the estimation accuracy of regression-type models [69]; they can be easily incorporate in the weighted `FastMix` model

766 by constructing an appropriate weighted covariance matrix. We use techniques introduced in
767 Zhang et al. [69] (`getSigma()` function from the `PBtest` R package) to estimate the
768 weighted covariance matrix if unknown. In the simplest case, if weights are known, the
769 weighted covariance matrix is a diagonal matrix with weights in the diagonal. For the weighted
770 `FastMix` model, a data transformation step equivalent to the weighted least squares (WLS)
771 approach is adopted with the given weighted covariance matrix before running the `FastMix`
772 algorithm (see Supplementary Material).

773 ***Simulation details***

774 ***Simulation I***

775 Simulation I is one iteration of a comprehensive simulation scheme described in Simulation II
776 with correlation $\rho = 0.5$ and balanced DEG design. Cell1 and Cell2 dimensions are visualized in
777 Figure 1c.

778 ***Simulation II***

779 The simulated bulk gene expression levels are associated with $L = 11$ covariates: three cell
780 proportions (Cell1, Cell2, and Cell3), two clinical covariates (Severity and Sex), and six
781 interaction terms between cell proportions and clinical covariates. The simulation design is as
782 follows.

783 1. Specifications of the fixed effects (β_l) and the random effects (γ_{li}) of NDEGs are:

1.1. Cell1 has an overall association with all gene expressions; Cell2 and Cell3 does not. Specifically, $\beta_1 = 1.5$ and $\beta_2 = \beta_3 = 0$. For NDEGs ($\iota = 0$), the random effects are γ_{1i}, γ_{2i}, and γ_{3i}, which have marginal distribution $N(0, \sigma_u^2)$ with $\sigma_u = 1$.

1.2. Neither Severity nor Sex has overall association with the whole transcriptome ($\beta_4 = \beta_5 = 0$). For NDEGs, the corresponding random effects γ_{4i} and γ_{5i}, have marginal distribution $N(0, \sigma_e^2)$ with $\sigma_e = 0.8$.

1.3. Only the interaction term between Cell1 and Severity has an overall impact on the whole transcriptome ($\beta_6 = 0.75$). For NDEGs, the Cell1-specific random effect with respect to Severity, γ_{6i}, has marginal distribution $N(0, \sigma_\alpha^2)$ with $\sigma_\alpha = 1.2$.

1.4. All other interaction terms have no overall association with gene expression ($\beta_l = 0$ for $l = 7, \cdots, 11$).

2. 20% of all genes are true DEGs (i.e., 1000 true DEGs). Specifications of the random effects (γ_{li}) of DEGs ($\iota = 1$) are:

2.1. Genes 1-250 are associated with Cell1 (a.k.a. Cell1 signature genes), i.e., $\gamma_{1i} \sim N(b_{1i}, \sigma_u)$. The true differential expression size is $|b_{1i}| = 3 \times \sigma_u$; the signs of b_{1i} follow a Bernoulli random variable with equal probability of being positive or negative (i.e., a balanced DEG design).

2.2. Genes 251 - 500 are DEGs with respect to Severity, i.e., $\gamma_{4i} \sim N(b_{4i}, \sigma_e)$. The true differential expression size is $|b_{4i}| = 3 \times \sigma_e$ with balanced DEG design.

2.3. Genes 501 - 750 are DEGs for the interaction term between Cell2 and Severity (a.k.a. Cell2-specific DEGs with respect to Severity), i.e., $\gamma_{7i} \sim N(b_{7i}, \sigma_\alpha)$. The true differential expression sizes are $|b_{7i}| = 3 \times \sigma_\alpha$ with balanced DEG design.

2.4. Genes 751 - 1000 are DEGs for the interaction term between Cell2 and Sex (a.k.a. Cell2-specific DEGs with respect to Sex), i.e., $\gamma_{10i} \sim N(b_{10i}, \sigma_\alpha)$. The true differential expression sizes are $|b_{10i}| = 3 \times \sigma_\alpha$ with balanced DEG design.

3. Consider two correlation structures of the random effects (i.e., true B matrix): either all random effects are independent (i.e., the identity matrix), or all random effects share an interchangeable correlation structure with $\rho = 0.5$.

4. The noise term is independent and identically distributed (i.i.d.) and follows $N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 0.25^2$.

Simulation III

Because csSAM is limited to one binary clinical covariate design, simulated data are generated as follows. The simulated bulk gene expression levels are associated with $L = 7$ covariates: three cell proportions (Cell1, Cell2, and Cell3), one clinical covariates (Group), and three interaction terms between cell proportions and the clinical covariate.

1. Specifications of the fixed effects (β_l) and the random effects (γ_{li}) of NDEGs are:

1.1. Cell1 has an overall association with all gene expressions; Cell2 and Cell3 does not. Specifically, $\beta_1 = 1.5$ and $\beta_2 = \beta_3 = 0$. For NDEGs ($\iota = 0$), the random effects are γ_{1i}, γ_{2i}, and γ_{3i}, which have marginal distribution $N(0, \sigma_u^2)$ with $\sigma_u = 1$.

1.2. Group is a binary variable without fix effect ($\beta_4 = 0$). For NDEGs, the random effects γ_{4i} has marginal distribution $N(0, \sigma_e^2)$ with $\sigma_e = 0.8$.

1.3. The interaction terms between Cell1/Cell2/Cell3 and Group have fixed effects on the whole transcriptome for $\beta_5 = 0.5, \beta_6 = 0.75$, and $\beta_7 = 0$, respectively. For NDEGs, the random effects γ_{5i}, γ_{6i}, and γ_{7i} have marginal distribution $N(0, \sigma_\alpha^2)$ with $\sigma_\alpha = 1.2$.

2. There are 500 true DEGs. Specifications of the random effects (γ_{li}) of DEGs ($\iota = 1$) are:

2.1. Genes 1-250 are DEGs for the interaction term Cell1 and Group (a.k.a. Cell1-specific DEGs with respect to Group), i.e., $\gamma_{5i} \sim N(b_{5i}, \sigma_\alpha)$. The true differential expression size is $|b_{5i}| = 3 \times \sigma_\alpha$ with balanced DEG design.

2.2. Genes 251 - 500 are DEGs for the interaction term Cell2 and Group (a.k.a. Cell2-specific DEGs with respect to Group), i.e., $\gamma_{6i} \sim N(b_{6i}, \sigma_\alpha)$. The true differential expression size is $|b_{6i}| = 3 \times \sigma_\alpha$ with balanced DEG design.

3. Consider two correlation structures of the random effects (i.e., true B matrix): either all random effects are independent (i.e., the identity matrix), or all random effects share an interchangeable correlation structure with $\rho = 0.5$.

4. The noise term is independent and identically distributed (i.i.d.) and follows $N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 0.25^2$.

Flow cytometry data and automated gating by DAFi

The FCM dataset in the HVP01 Study (https://clinicaltrials.gov/ct2/show/NCT03083158) are provided by Kollmann lab, which has 75 FCS files (15 subjects across 5 visits – Day 0, 1, 3, 7 and 14). The markers included in the reagent panel can be found in Supplementary Table S9. DAFi [23] (https://github.com/JCVenterInstitute/DAFi-gating) was applied to identify the neutrophil cell population following a predefined gating sequence (Figure 2a): Singlets (FSC-A vs FSC-H) -> Leukocytes (FSC-A vs SSC-A) -> Live Leukocytes (Viability vs SSC-A) -> Neutrophils (CD66 vs CD45). Proportions of neutrophils and their 2D dot plots for all 75 FCS files can be found in Supplementary File and Supplementary Figures S5-6 (in two batches May and August). A single set of DAFi-gating boundaries was used to identify the natural shapes of neutrophils in each batch to avoid using abrupt cutoffs in the manual gating analysis and to provide straightforward cross-sample comparison.

The second study we analyzed is the SDY180 on ImmPort (https://www.immport.org/shared/study/SDY180), which is focused on immune responses to influenza and pneumococcal vaccines [48]. Among all the reagent panels used in SDY180, two of them contain CD45 and CD14 for us to define the granulocytes and monocytes (Supplementary Table S9). The 302 corresponding FCS files of the two panels are from 36 subjects across 8 visits. DAFi was applied to identify three major types of cells from the FCM data (Figure 5a): Lymphocytes (FSC-A vs SSC-A), Granulocytes (CD45 vs CD14 followed by back-gating on FSC-A vs SSC-A), and Monocytes (CD45 vs CD14 followed by back-gating on FSC-A vs SSC-A).

Data Availability

The HVP01 dataset (https://clinicaltrials.gov/ct2/show/NCT03083158) is a clinical study conducted by University of British Columbia focused on Hepatitis B vaccine Engerix-B. This study has 16 healthy subjects from two cohorts: young adults (aged 40-60) and old adults (aged 61-80). Both RNA-seq gene expressions and flow cytometry data are available across multiple visits before and after the vaccination from the same whole blood samples. Primary outcome of this study is the antibody response to the first dose of Hepatitis B vaccine.

The SDY180 dataset is downloaded from the ImmPort Shared Data portal (http://www.immport.org). This study has 18 young and healthy adult volunteers (aged 18-64) randomly assigned to three study groups (n = 6 subjects/group) receiving a single intramuscular dose of 2009–2010 seasonal influenza (Fluzone, Sanofi Pasteur, PA), pneumococcal vaccine (Pneumovax23, Merck, NJ), or placebo (saline). Blood samples were collected at multiple time visits, from 7 days before vaccination to 28 days after vaccination, for microarray, whole-blood flow cytometry, and serum analysis of neutralizing antibodies.

Data preprocessing details are in Supplementary Material.

Acknowledgements

This work is partially funded by NIH/NIAID UH2AI132342, the Human Vaccines Project, the Respiratory Pathogens Research Center (NIAID contract number HHSN272201200005C) and the University of Rochester CTSA award number UL1 TR002001 from the National Center for Advancing Translational Sciences of the National Institutes of Health. The content is solely the

881 responsibility of the authors and does not necessarily represent the official views of the

882 National Institutes of Health.

883 **Author Contributions**

884 XQ and YQ conceived the project. HS and XQ designed and implemented the `FastMix` model,

885 including simulations. TRK and RHS provided HVP data and guidance on method applications.

886 AM processed the flow cytometry data. BDA processed the scRNA-seq data. YZ led the data

887 analytical experiments and assessed the model performance. YZ and HS drafted the manuscript.

888 YQ and XQ revised the manuscript. All authors read and agreed on the manuscript.

References

1. Pinu, F.R., et al., *Systems biology and multi-omics integration: viewpoints from the metabolomics research community.* Metabolites, 2019. **9**(4): p. 76.
2. Aevermann, B.D., et al., *Machine learning-based single cell and integrative analysis reveals that baseline mDC predisposition predicts protective Hepatitis B vaccine response.* medRxiv, 2021.
3. Li, Y., et al., *Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine.* Briefings in Bioinformatics, 2021.
4. Consortium, H.-I., *Multicohort analysis reveals baseline transcriptional predictors of influenza vaccination responses.* Science immunology, 2017. **2**(14).
5. Tomic, A., et al., *SIMON, an automated machine learning system, reveals immune signatures of influenza vaccine responses.* The Journal of Immunology, 2019. **203**(3): p. 749-759.
6. Noecker, C., et al., *Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation.* MSystems, 2016. **1**(1): p. e00013-15.
7. McCall, M.N., et al., *A systems genomics approach uncovers molecular associates of RSV severity*. bioRxiv, 2020.
8. Abdi, H. and L.J. Williams, *Principal component analysis.* Wiley interdisciplinary reviews: computational statistics, 2010. **2**(4): p. 433-459.
9. Hardoon, D.R., S. Szedmak, and J. Shawe-Taylor, *Canonical correlation analysis: An overview with application to learning methods.* Neural computation, 2004. **16**(12): p. 2639-2664.
10. Abdi, H., *Partial least square regression (PLS regression).* Encyclopedia for research methods for the social sciences, 2003. **6**(4): p. 792-795.
11. Van der Maaten, L. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(11).
12. McInnes, L., J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction*. arXiv preprint arXiv:1802.03426, 2018.
13. Hoerl, A.E. and R.W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems.* Technometrics, 1970. **12**(1): p. 55-67.
14. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 1996. **58**(1): p. 267-288.
15. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the royal statistical society: series B (statistical methodology), 2005. **67**(2): p. 301-320.
16. Singh, A., et al., *DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays.* Bioinformatics, 2019. **35**(17): p. 3055-3062.
17. Peng, C., et al., *A latent unknown clustering integrating multi-omics data (LUCID) with phenotypic traits.* Bioinformatics, 2020. **36**(3): p. 842-850.
18. Jin, S., L. Zhang, and Q. Nie, *scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles.* Genome biology, 2020. **21**(1): p. 1-19.

19. Cao, K., et al., *Unsupervised topological alignment for single-cell multi-omics integration.* Bioinformatics, 2020. **36**(Supplement_1): p. i48-i56.
20. Maldonado, Y.M., *Mixed models, posterior means and penalized least-squares.* Lecture Notes-Monograph Series, 2009: p. 216-236.
21. Zhang, S., et al., *MatchMixeR: a cross-platform normalization method for gene expression data integration.* Bioinformatics, 2020. **36**(8): p. 2486-2491.
22. Bates, D., et al., *Fitting linear mixed-effects models using lme4.* arXiv preprint arXiv:1406.5823, 2014.
23. Lee, A.J., et al., *DAFi: A directed recursive data filtering and clustering approach for improving and interpreting data clustering identification of cell populations from polychromatic flow cytometry data.* Cytometry A, 2018. **93**(6): p. 597-610.
24. Efron, B., et al., *Empirical Bayes analysis of a microarray experiment.* Journal of the American statistical association, 2001. **96**(456): p. 1151-1160.
25. Qiu, X., L. Klebanov, and A. Yakovlev, *Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes.* Statistical applications in genetics and molecular biology, 2005. **4**(1).
26. Khanam, A., et al., *Blockade of neutrophil's chemokine receptors CXCR1/2 abrogate liver damage in acute-on-chronic liver failure.* Frontiers in immunology, 2017. **8**: p. 464.
27. Le, P.-H., et al., *Clinical Predictors for Neutrophil–to–Lymphocyte Ratio Changes in Patients with Chronic Hepatitis B Receiving Peginterferon Treatment.* in vivo, 2017. **31**(4): p. 723-729.
28. Tang, B.M., et al., *Neutrophils-related host factors associated with severe disease and fatality in patients with influenza infection.* Nature communications, 2019. **10**(1): p. 1-13.
29. Cui, Z., et al., *Super-delta2: An Enhanced Differential Expression Analysis Procedure for Multi-Group Comparisons of RNA-seq Data.* Bioinformatics, 2021.
30. Liu, Y., J. Zhang, and X. Qiu, *Super-delta: a new differential gene expression analysis procedure with robust data normalization.* BMC Bioinformatics, 2017. **18**(1): p. 582.
31. Maronna, R.A. and V.J. Yohai, *The behavior of the Stahel-Donoho robust multivariate estimator.* Journal of the American Statistical Association, 1995. **90**(429): p. 330-341.
32. Maronna, R.A. and R.H. Zamar, *Robust estimates of location and dispersion for high-dimensional datasets.* Technometrics, 2002. **44**(4): p. 307-317.
33. Rousseeuw, P.J. and K.V. Driessen, *A fast algorithm for the minimum covariance determinant estimator.* Technometrics, 1999. **41**(3): p. 212-223.
34. Shen-Orr, S.S., et al., *Cell type–specific gene expression differences in complex tissues.* Nature methods, 2010. **7**(4): p. 287-289.
35. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proceedings of the National Academy of Sciences, 2001. **98**(9): p. 5116-5121.
36. Wooden, S.L. and W.C. Koff, *The Human Vaccines Project: Towards a comprehensive understanding of the human immune response to immunization.* Human vaccines & immunotherapeutics, 2018. **14**(9): p. 2214-2216.
37. Shannon, C.P., et al., *Multi-omic data integration allows baseline immune signatures to predict hepatitis B vaccine response in a small cohort.* Frontiers in immunology, 2020. **11**.

974 38. Picelli, S., et al., *Full-length RNA-seq from single cells using Smart-seq2.* Nature protocols,
975 2014. **9**(1): p. 171-181.
976 39. Keating, G.M. and S. Noble, *Recombinant hepatitis B vaccine (Engerix-B®).* Drugs, 2003.
977 **63**(10): p. 1021-1051.
978 40. Blondel, V.D., et al., *Fast unfolding of communities in large networks.* Journal of
979 statistical mechanics: theory and experiment, 2008. **2008**(10): p. P10008.
980 41. Kiselev, V.Y., et al., *SC3: consensus clustering of single-cell RNA-seq data.* Nature
981 methods, 2017. **14**(5): p. 483-486.
982 42. Xu, R., et al., *Low expression of CXCR1/2 on neutrophils predicts poor survival in patients*
983 *with hepatitis B virus-related acute-on-chronic liver failure.* Scientific reports, 2016. **6**(1):
984 p. 1-9.
985 43. Yu, G. and Q.-Y. He, *ReactomePA: an R/Bioconductor package for reactome pathway*
986 *analysis and visualization.* Molecular BioSystems, 2016. **12**(2): p. 477-479.
987 44. Blasius, A.L., et al., *Bone marrow stromal cell antigen 2 is a specific marker of type I IFN-*
988 *producing cells in the naive mouse, but a promiscuous cell surface antigen following IFN*
989 *stimulation.* The Journal of Immunology, 2006. **177**(5): p. 3260-3265.
990 45. Sarojini, S., T. Theofanis, and C.S. Reiss, *Interferon-induced tetherin restricts vesicular*
991 *stomatitis virus release in neurons.* DNA and cell biology, 2011. **30**(12): p. 965-974.
992 46. Miyagi, E., et al., *Vpu enhances HIV-1 virus release in the absence of Bst-2 cell surface*
993 *down-modulation and intracellular depletion.* Proceedings of the National Academy of
994 Sciences, 2009. **106**(8): p. 2868-2873.
995 47. Bhattacharya, S., et al., *ImmPort, toward repurposing of open access immunological*
996 *assay data for translational and clinical research.* Scientific data, 2018. **5**: p. 180015.
997 48. Obermoser, G., et al., *Systems scale interactive exploration reveals quantitative and*
998 *qualitative differences in response to influenza and pneumococcal vaccines.* Immunity,
999 2013. **38**(4): p. 831-844.
1000 49. Gatti, D.M., et al., *Heading down the wrong pathway: on the influence of correlation*
1001 *within gene sets.* BMC genomics, 2010. **11**(1): p. 1-10.
1002 50. Wu, D. and G.K. Smyth, *Camera: a competitive gene set test accounting for inter-gene*
1003 *correlation.* Nucleic acids research, 2012. **40**(17): p. e133-e133.
1004 51. Zhang, Y., et al., *FUNNEL-GSEA: FUNctioNal ELastic-net regression in time-course gene*
1005 *set enrichment analysis.* Bioinformatics, 2017. **33**(13): p. 1944-1952.
1006 52. Burel, J.G., et al., *An integrated workflow to assess technical and biological variability of*
1007 *cell population frequencies in human peripheral blood by flow cytometry.* The Journal of
1008 Immunology, 2017. **198**(4): p. 1748-1758.
1009 53. Kolaczkowska, E., et al., *Neutrophil elastase activity compensates for a genetic lack of*
1010 *matrix metalloproteinase ‐9 (MMP ‐9) in leukocyte infiltration in a model of*
1011 *experimental peritonitis.* Journal of leukocyte biology, 2009. **85**(3): p. 374-381.
1012 54. Hinson, E.R., et al., *Viperin is highly induced in neutrophils and macrophages during*
1013 *acute and chronic lymphocytic choriomeningitis virus infection.* The Journal of
1014 Immunology, 2010. **184**(10): p. 5723-5731.
1015 55. Pei, R., et al., *Interferon-induced proteins with tetratricopeptide repeats 1 and 2 are*
1016 *cellular factors that limit hepatitis B virus replication.* Journal of innate immunity, 2014.
1017 **6**(2): p. 182-191.

56. Gaujoux, R. and C. Seoighe, *Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study.* Infection, Genetics and Evolution, 2012. **12**(5): p. 913-921.
57. Venet, D., et al., *Separation of samples into their constituents using gene expression data.* Bioinformatics, 2001. **17**(suppl_1): p. S279-S287.
58. Lähdesmäki, H., et al., *In silico microdissection of microarray data from heterogeneous cell populations.* BMC bioinformatics, 2005. **6**(1): p. 54.
59. Repsilber, D., et al., *Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach.* BMC bioinformatics, 2010. **11**(1): p. 1-15.
60. Newman, A.M., et al., *Robust enumeration of cell subsets from tissue expression profiles.* Nature methods, 2015. **12**(5): p. 453-457.
61. Quon, G., et al., *Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction.* Genome medicine, 2013. **5**(3): p. 29.
62. Zhang, Y., et al., *The effect of tissue composition on gene co-expression.* Briefings in Bioinformatics, 2019.
63. Qiao, W., et al., *PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions.* PLoS Comput Biol, 2012. **8**(12): p. e1002838.
64. Quon, G. and Q. Morris, *ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing.* Bioinformatics, 2009. **25**(21): p. 2882-2889.
65. Mohammadi, S., et al., *A critical survey of deconvolution methods for separating cell types in complex tissues.* Proceedings of the IEEE, 2016. **105**(2): p. 340-366.
66. Horn, R.A., R.A. Horn, and C.R. Johnson, *Topics in matrix analysis*. 1994: Cambridge university press.
67. Robinson, G.K., *That BLUP is a good thing: the estimation of random effects.* Statistical science, 1991. **6**(1): p. 15-32.
68. Satterthwaite, F.E., *An approximate distribution of estimates of variance components.* Biometrics bulletin, 1946. **2**(6): p. 110-114.
69. Zhang, Y., et al., *Highly efficient hypothesis testing methods for regression-type tests with correlated observations and heterogeneous variance structure.* BMC Bioinformatics, 2019. **20**(1): p. 185.

a

Bulk gene expression matrix

	gene 1	...	...	...	...	...	gene m
sample 1							
sample n							

Cell type proportions

	cell 1	...	cell K
sample 1			
sample n			

Clinical covariates

	clin 1	...	clin P
sample 1			
sample n			

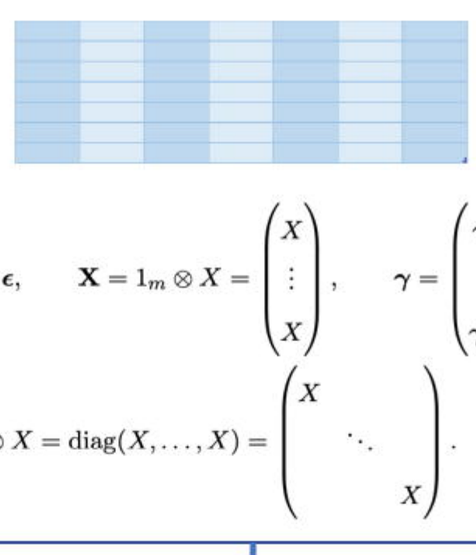
b

i. FastMix model

$$Y_{ji} = \sum_{l=1}^{L} X_{jl}\,(\beta_l + \gamma_{li}) + \epsilon_{ij}, \qquad \boldsymbol{\gamma}_i \sim f(\mathbf{x}|\iota), \quad \epsilon_{ij} \sim N(0, \sigma_\epsilon^2).$$

$$f(\mathbf{x}|\iota) = \begin{cases} f_0(\mathbf{x}) := \phi(\mathbf{x}|\mathbf{0}, B), & \iota = 0, \\ f_1(\mathbf{x}), & \iota = 1. \end{cases} \qquad f(\mathbf{x}) = \pi_0 f_0(\mathbf{x}) + (1-\pi_0) f_1(\mathbf{x}).$$

ii. Vectorization and Kronecker product



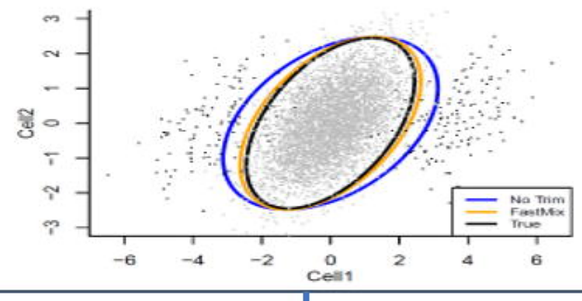
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \qquad \mathbf{X} = 1_m \otimes X = \begin{pmatrix} X \\ \vdots \\ X \end{pmatrix}, \qquad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_{\cdot 1} \\ \vdots \\ \gamma_{\cdot m} \end{pmatrix},$$

$$\mathbf{Z} := I_m \otimes X = \mathrm{diag}(X, \ldots, X) = \begin{pmatrix} X & & \\ & \ddots & \\ & & X \end{pmatrix}.$$

iii. Moment-based estimation

$$\hat{B}^{(0)}, \hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\gamma}}_i^{(1)}$$

iv. Trimming for potential DEGs



v. Re-estimation and bias correction

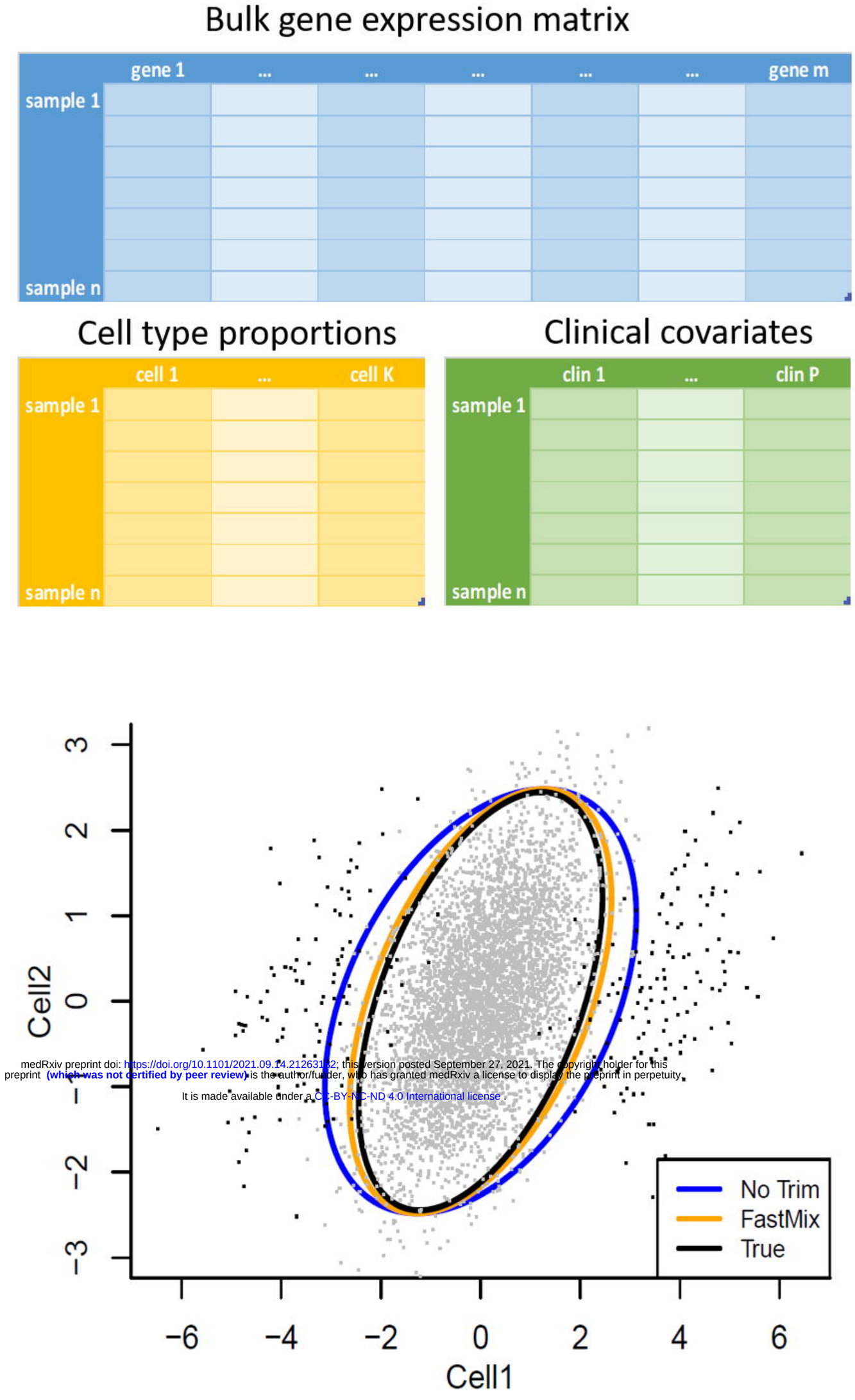
$$\hat{B}^{(0)} \rightarrow \hat{B}_{\mathrm{T}}, \hat{\boldsymbol{\beta}}^{(1)} \rightarrow \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}_i^{(1)} \rightarrow \hat{\boldsymbol{\gamma}}_i$$

vi. Hypothesis test and quasi-p-value

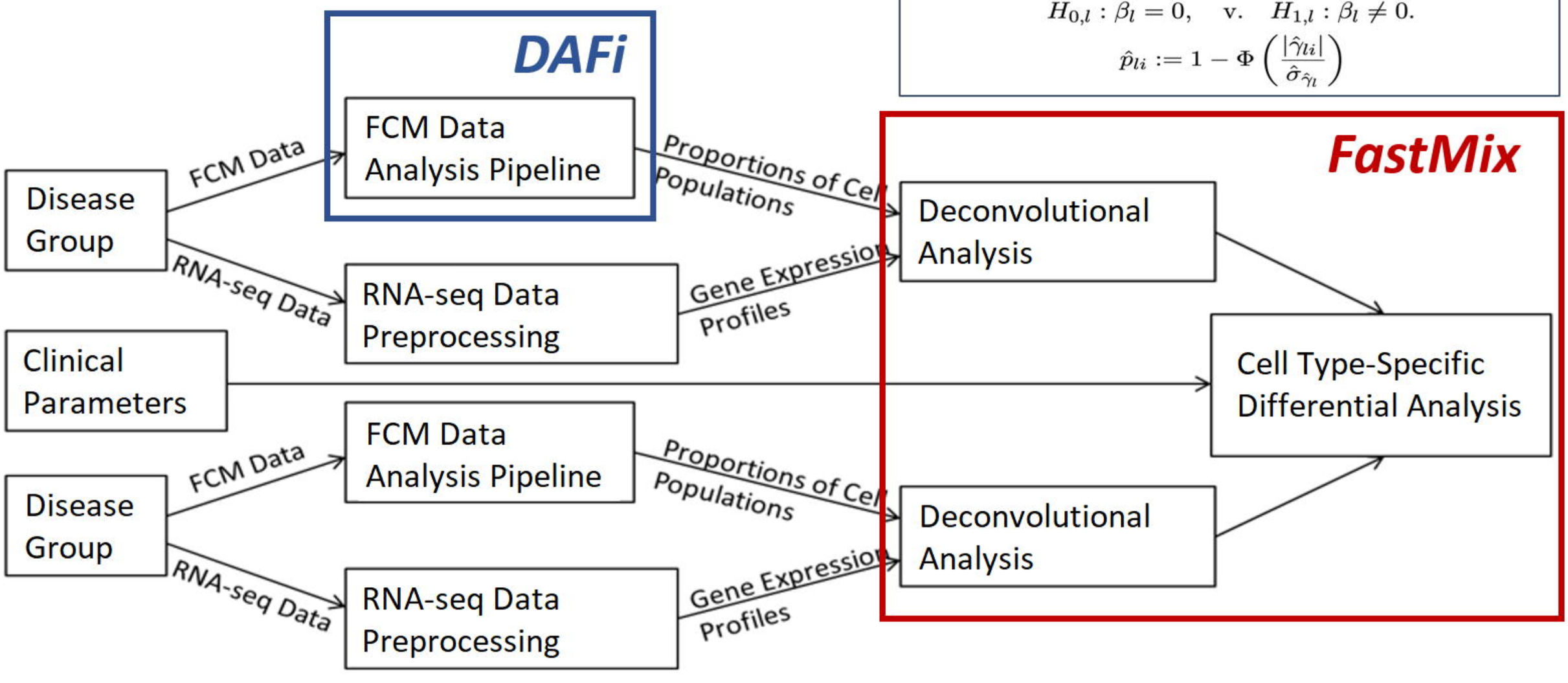
$$H_{0,l} : \beta_l = 0, \quad \text{v.} \quad H_{1,l} : \beta_l \neq 0.$$

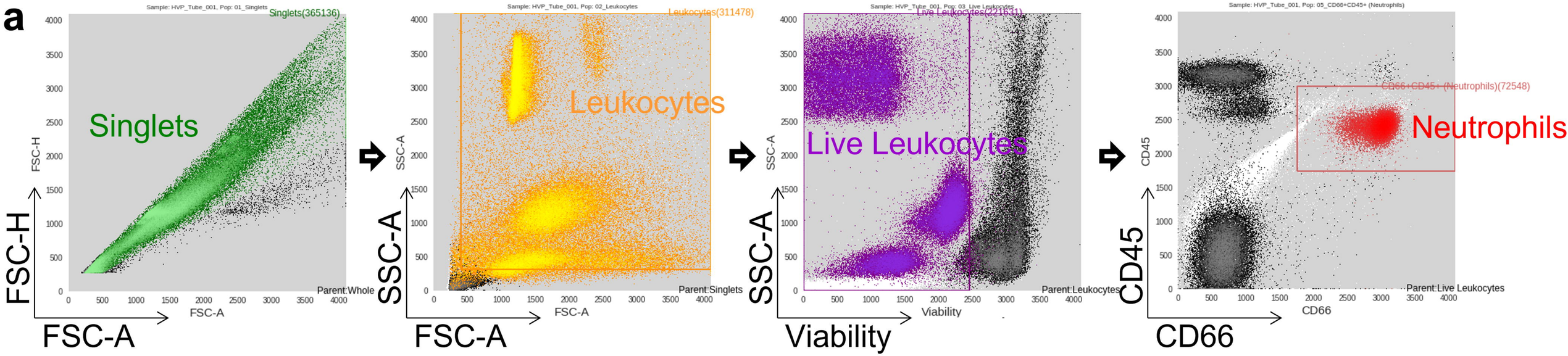
$$\hat{p}_{li} := 1 - \Phi\left(\frac{|\hat{\gamma}_{li}|}{\hat{\sigma}_{\hat{\gamma}_l}}\right)$$

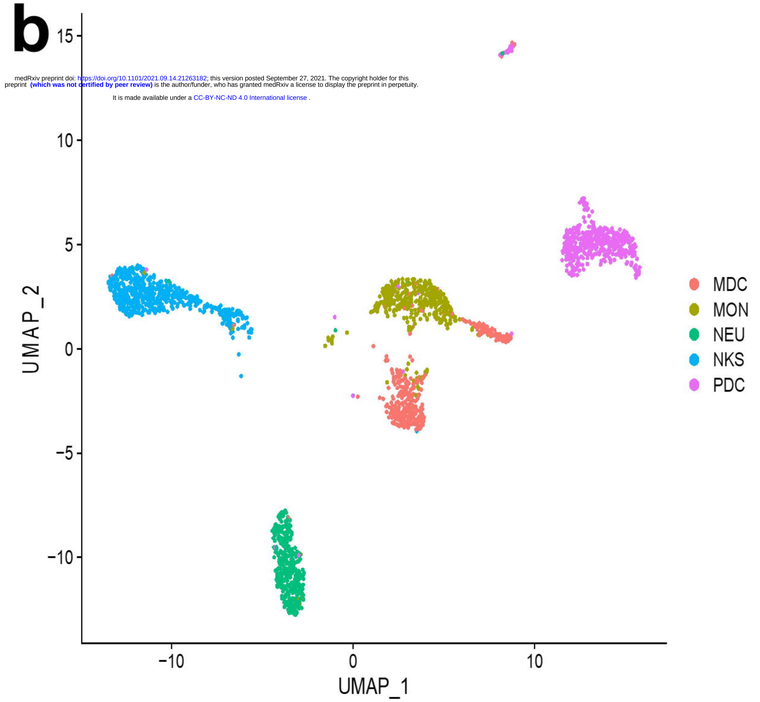
c

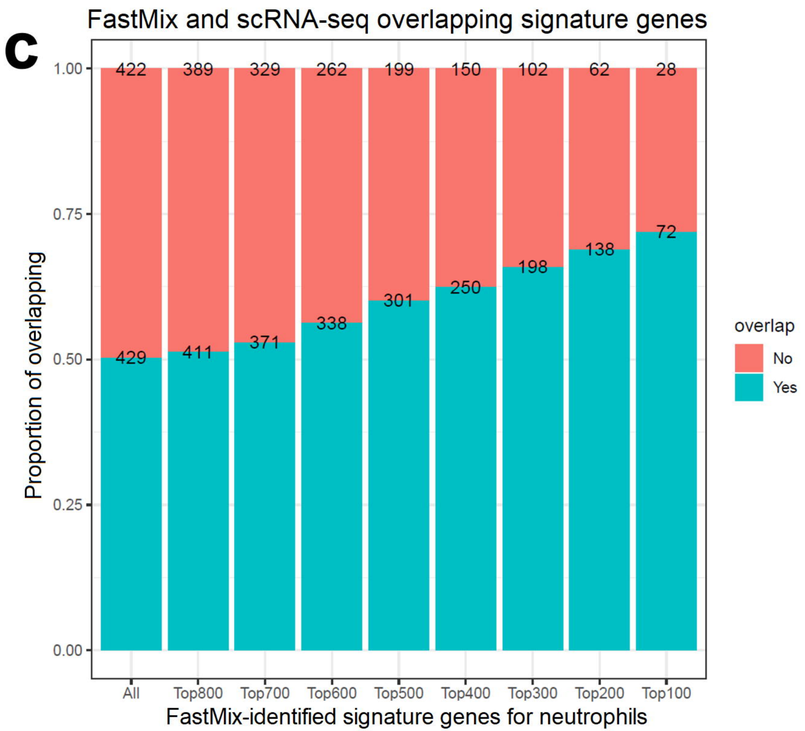


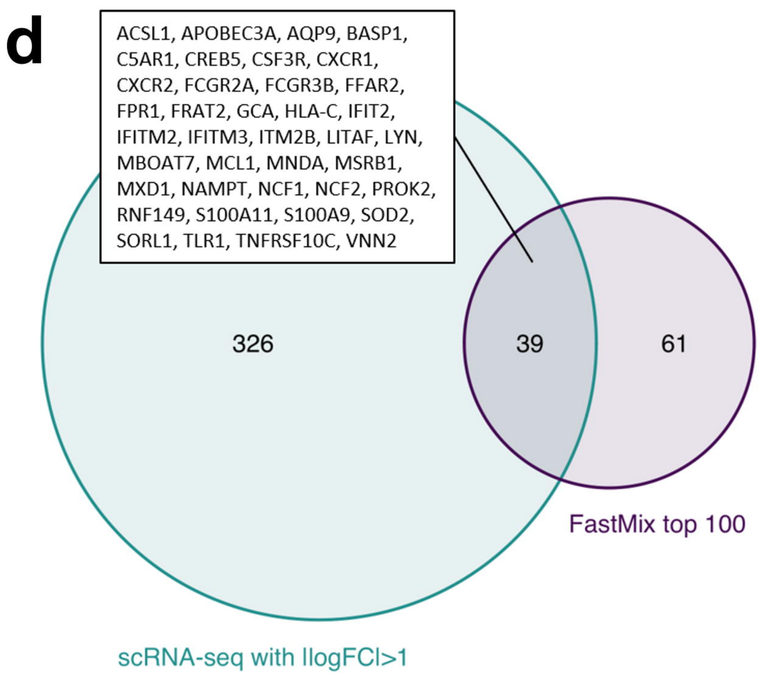
d

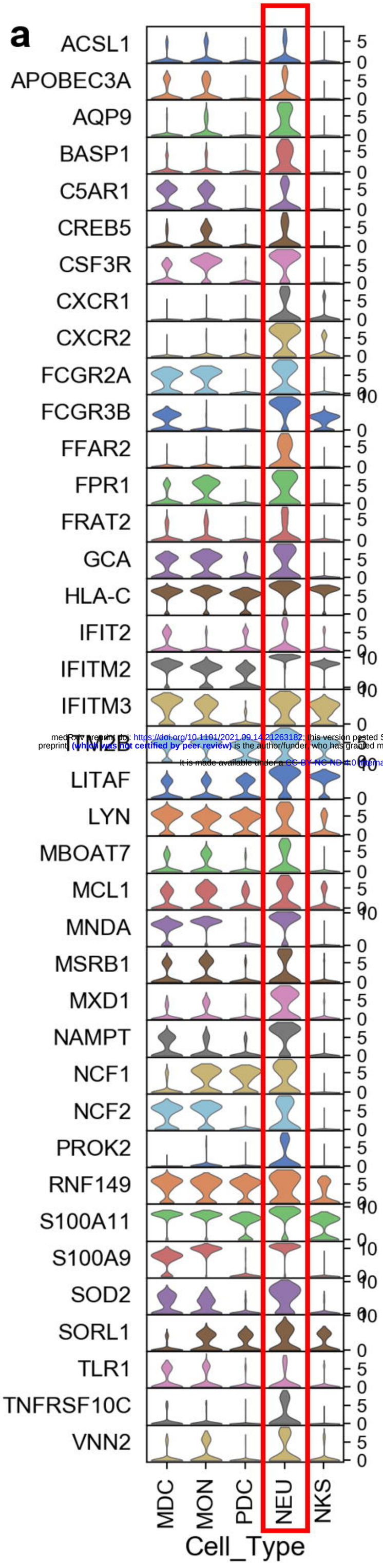


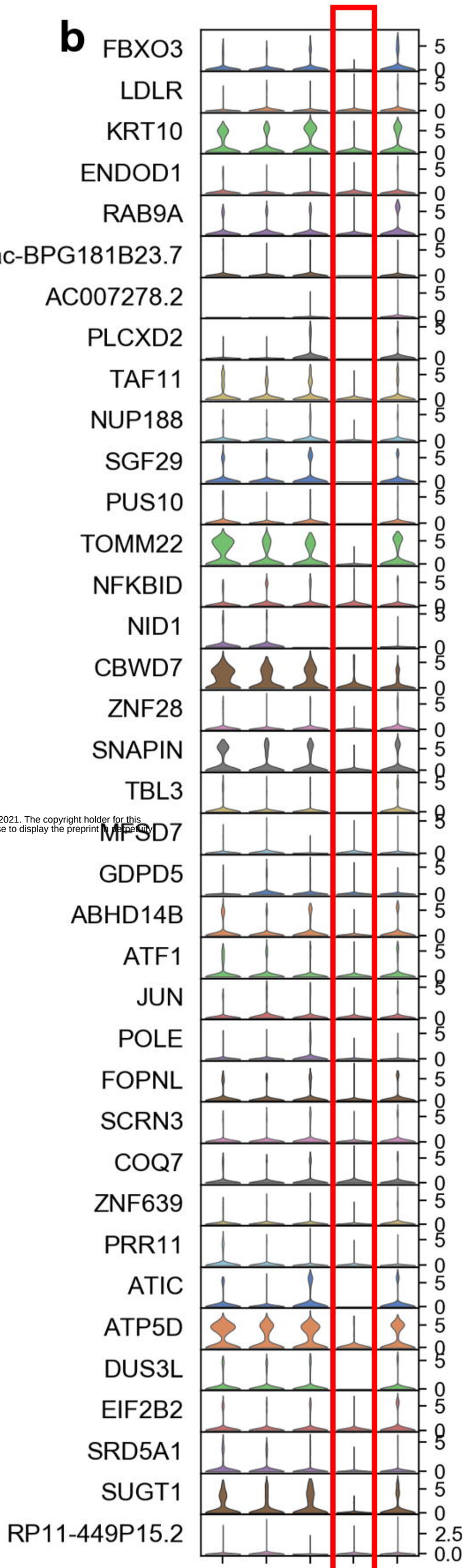

a
Sample: HVP_Tube_001, Pop: 01_Singlets
Singlets(365136)
Singlets
FSC-H
FSC-A
Parent:Whole
Sample: HVP_Tube_001, Pop: 02_Leukocytes
Leukocytes(311478)
Leukocytes
SSC-A
FSC-A
Parent:Singlets
Sample: HVP_Tube_001, Pop: 03_Live Leukocytes
Live Leukocytes(221531)
Live Leukocytes
SSC-A
Viability
Parent:Leukocytes
Sample: HVP_Tube_001, Pop: 05_CD66+CD45+ (Neutrophils)
CD66+CD45+ (Neutrophils)(72548)
Neutrophils
CD45
CD66
Parent:Live Leukocytes

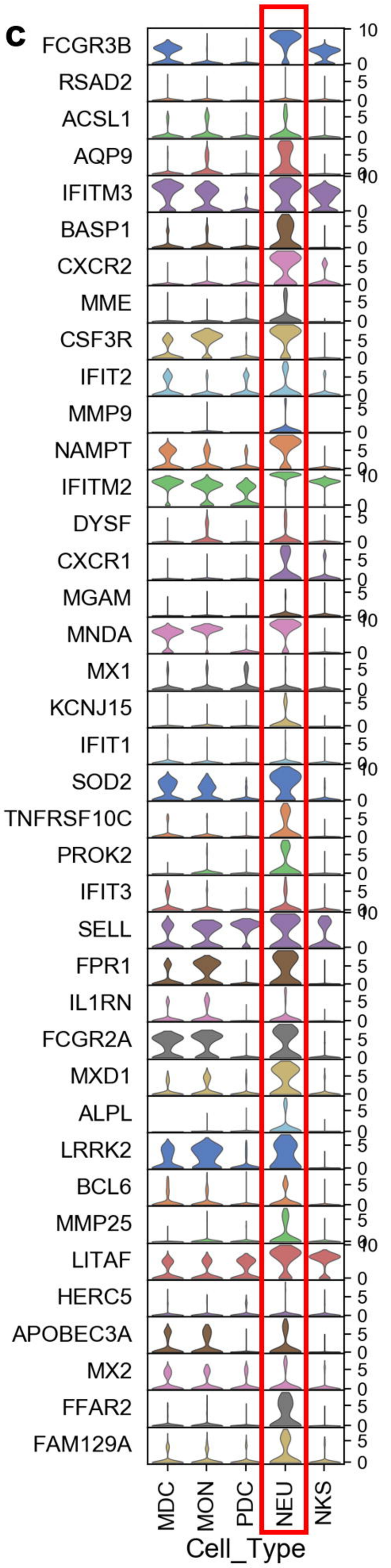

b
UMAP_2
UMAP_1
MDC
MON
NEU
NKS
PDC


c
FastMix and scRNA-seq overlapping signature genes
Proportion of overlapping
FastMix-identified signature genes for neutrophils
All
Top800
Top700
Top600
Top500
Top400
Top300
Top200
Top100
422
389
329
262
199
150
102
62
28
429
411
371
338
301
250
198
138
72
overlap
No
Yes

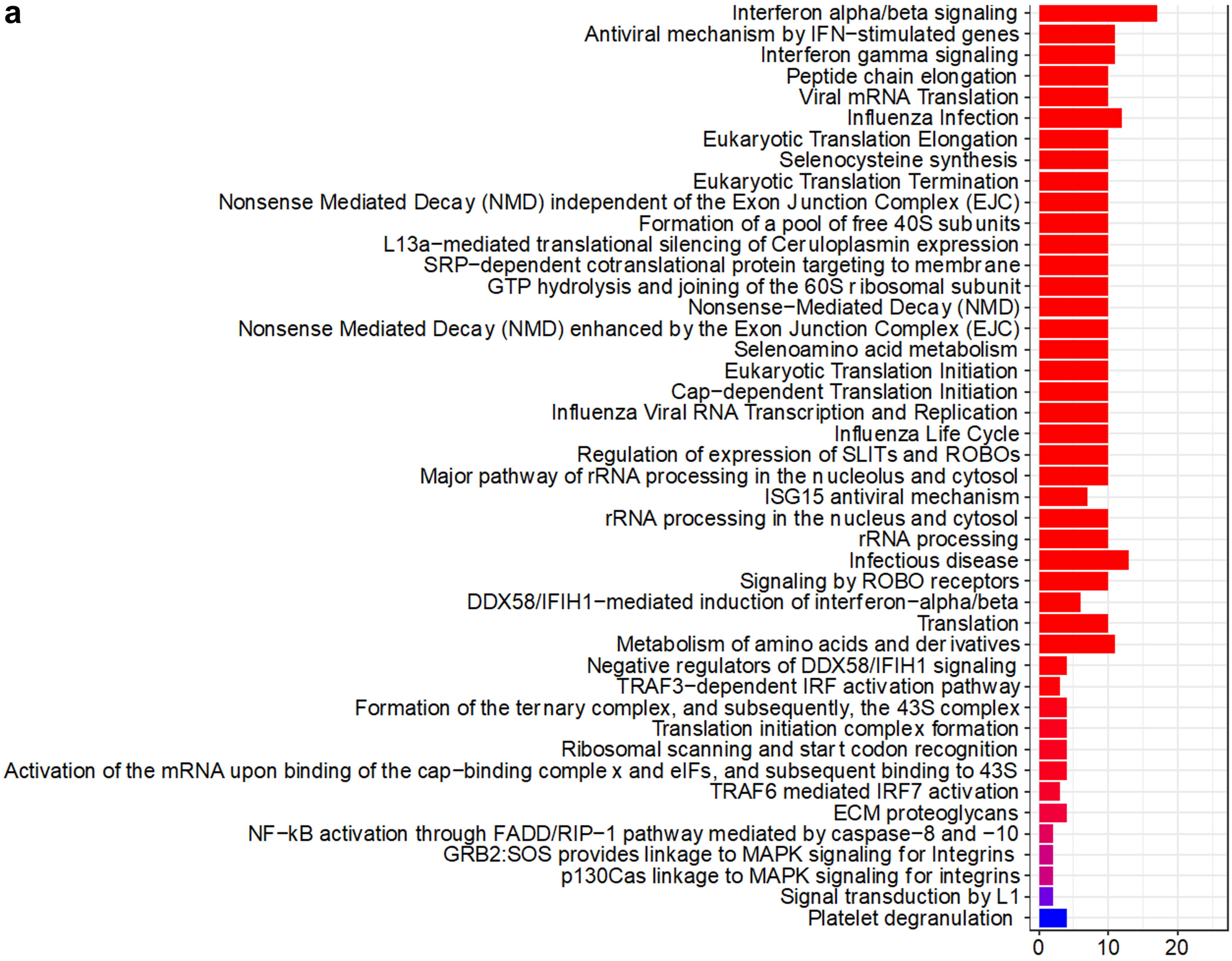

d
ACSL1, APOBEC3A, AQP9, BASP1,
C5AR1, CREB5, CSF3R, CXCR1,
CXCR2, FCGR2A, FCGR3B, FFAR2,
FPR1, FRAT2, GCA, HLA-C, IFIT2,
IFITM2, IFITM3, ITM2B, LITAF, LYN,
MBOAT7, MCL1, MNDA, MSRB1,
MXD1, NAMPT, NCF1, NCF2, PROK2,
RNF149, S100A11, S100A9, SOD2,
SORL1, TLR1, TNFRSF10C, VNN2
326
39
61
FastMix top 100
scRNA-seq with |logFC|>1







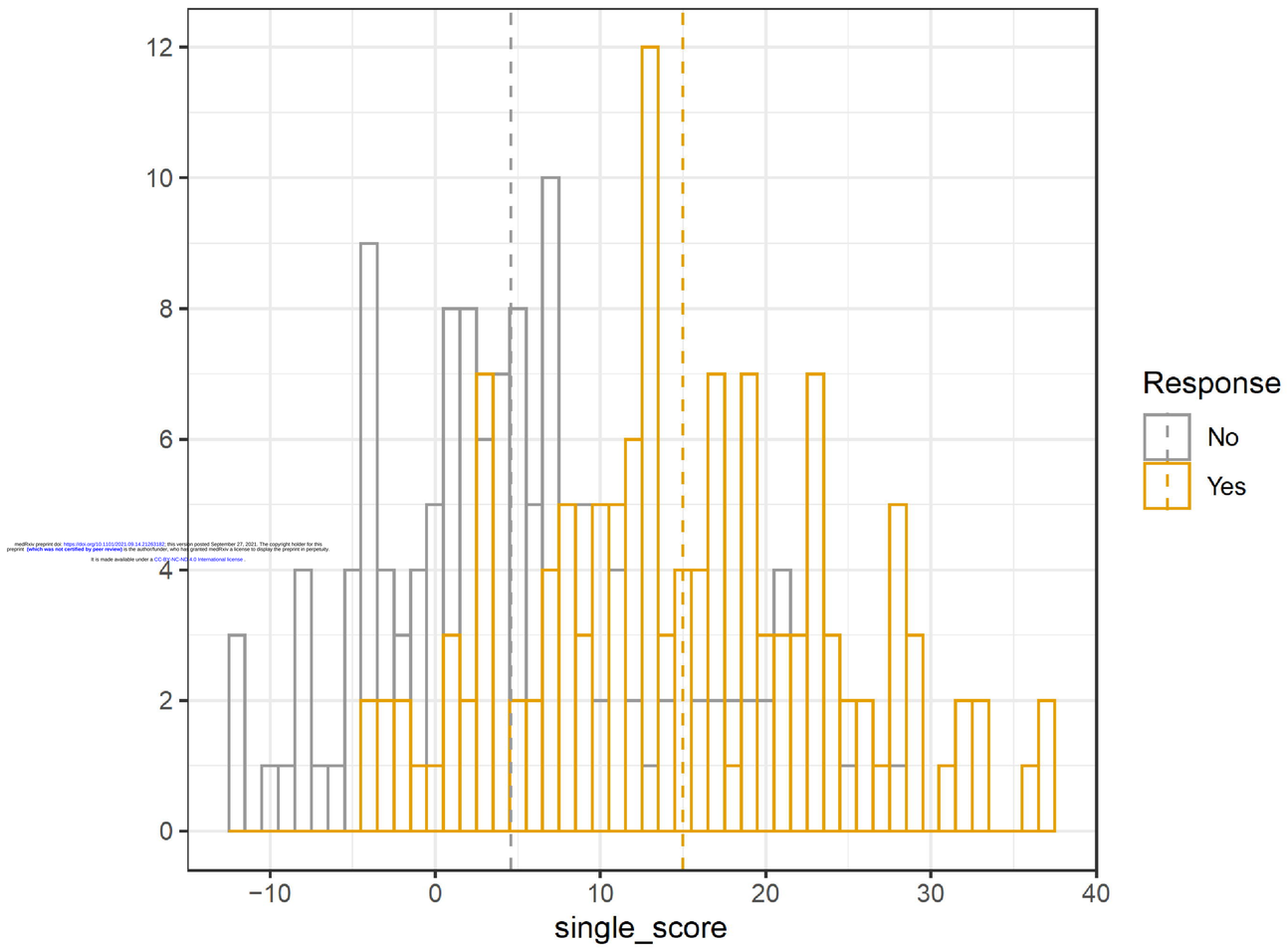
a



b

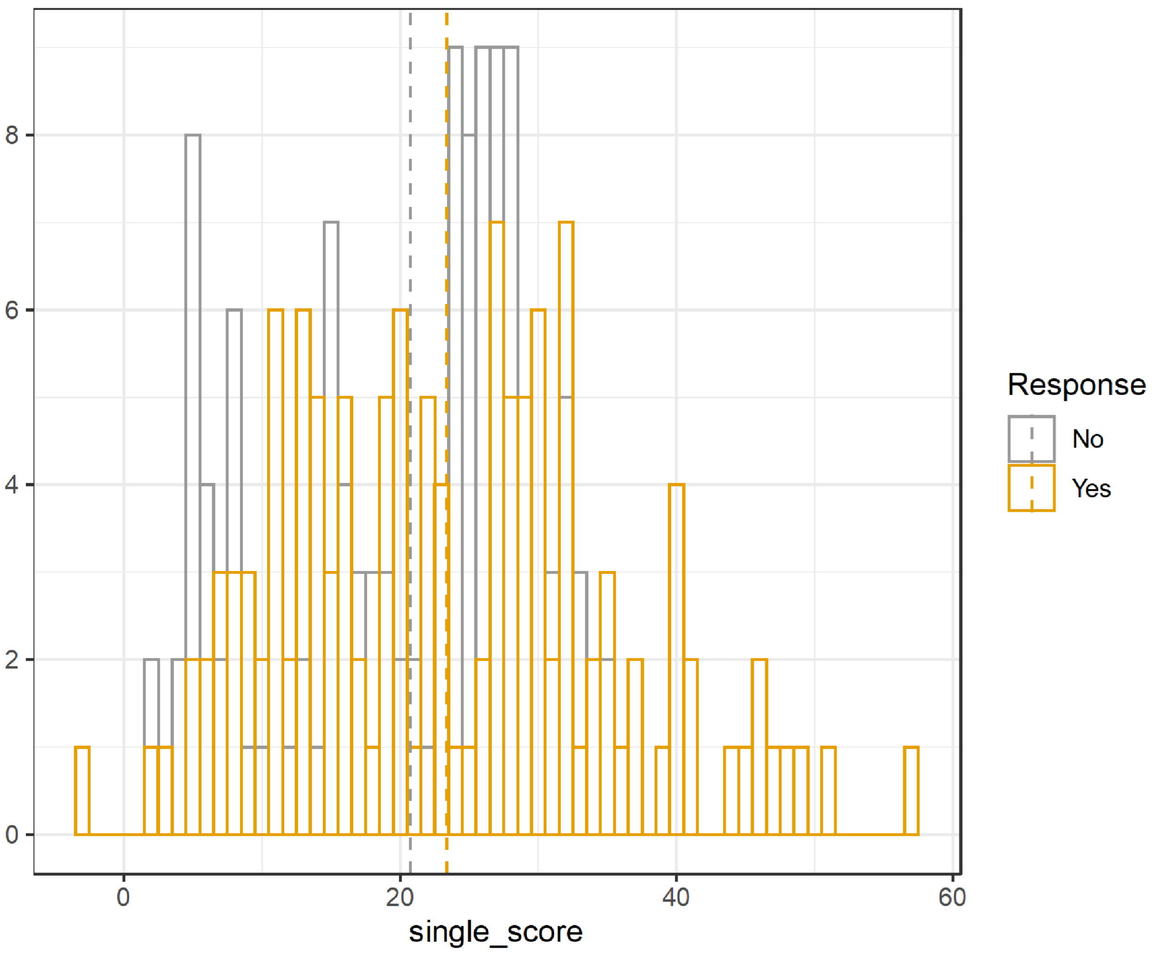
	CD45pCD66p.Response
1	AGRN
2	BST2
3	CHMP5
4	DDX58
5	DHX58
6	EIF2AK2
7	FCGR1B
8	GBP1
9	GBP3
10	GBP5
11	HERC5
12	IDO1
13	IFI35
14	IFI6
15	IFIH1
16	IFIT1
17	IFIT2
18	IFIT3
19	IFITM3
20	IRF7
21	ISG15
22	ITGA2B
23	ITGB3
24	MT2A
25	MX1
26	OAS1
27	OAS2
28	OAS3
29	OASL
30	RPL23
31	RPL27
32	RPL31
33	RPL34
34	RPL39
35	RPL9
36	RPS15A
37	RPS3A
38	RPS4Y1
39	RPS7
40	RSAD2
41	SERPING1
42	SPARC
43	TRIM22
44	USP18
45	XAF1

c

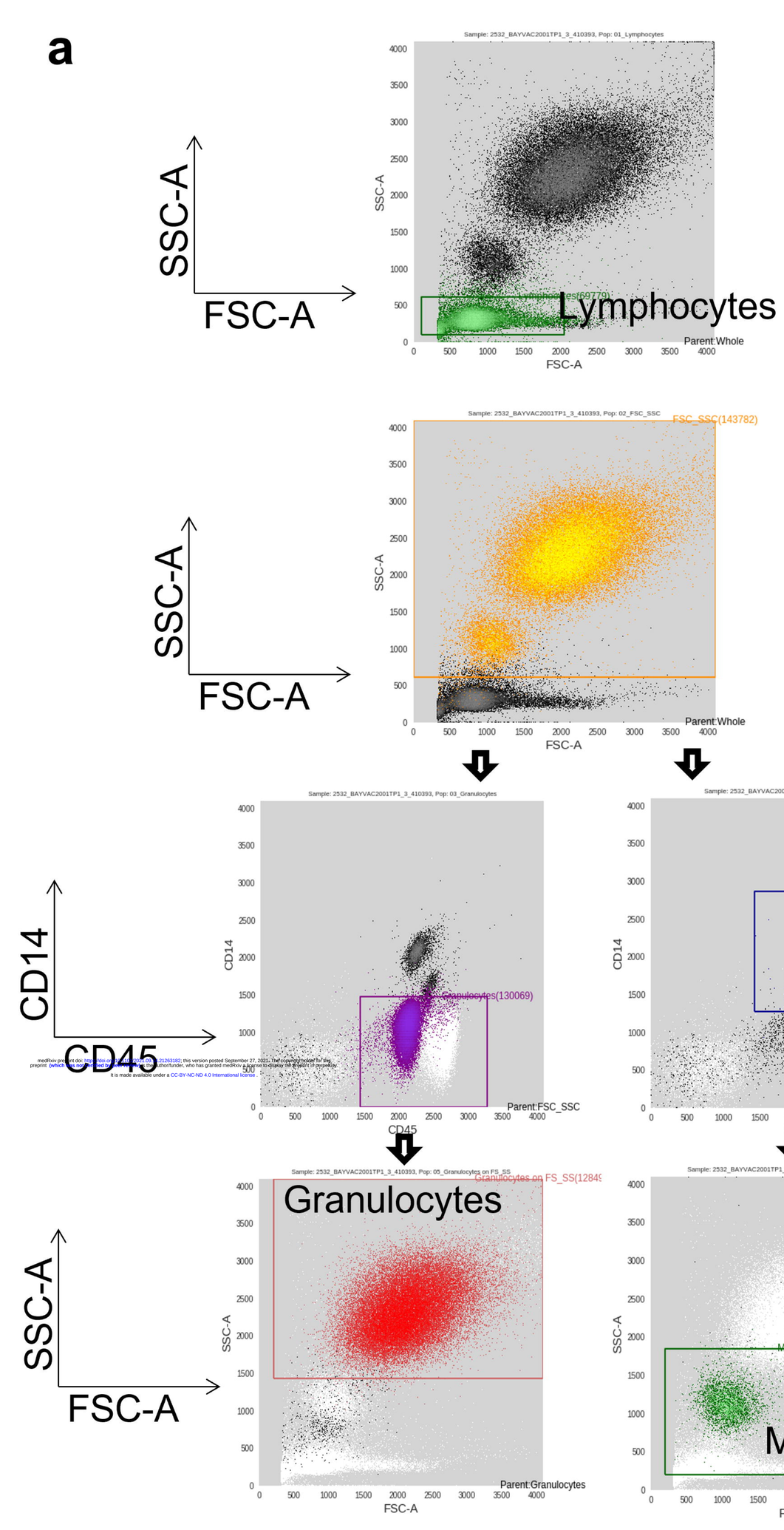


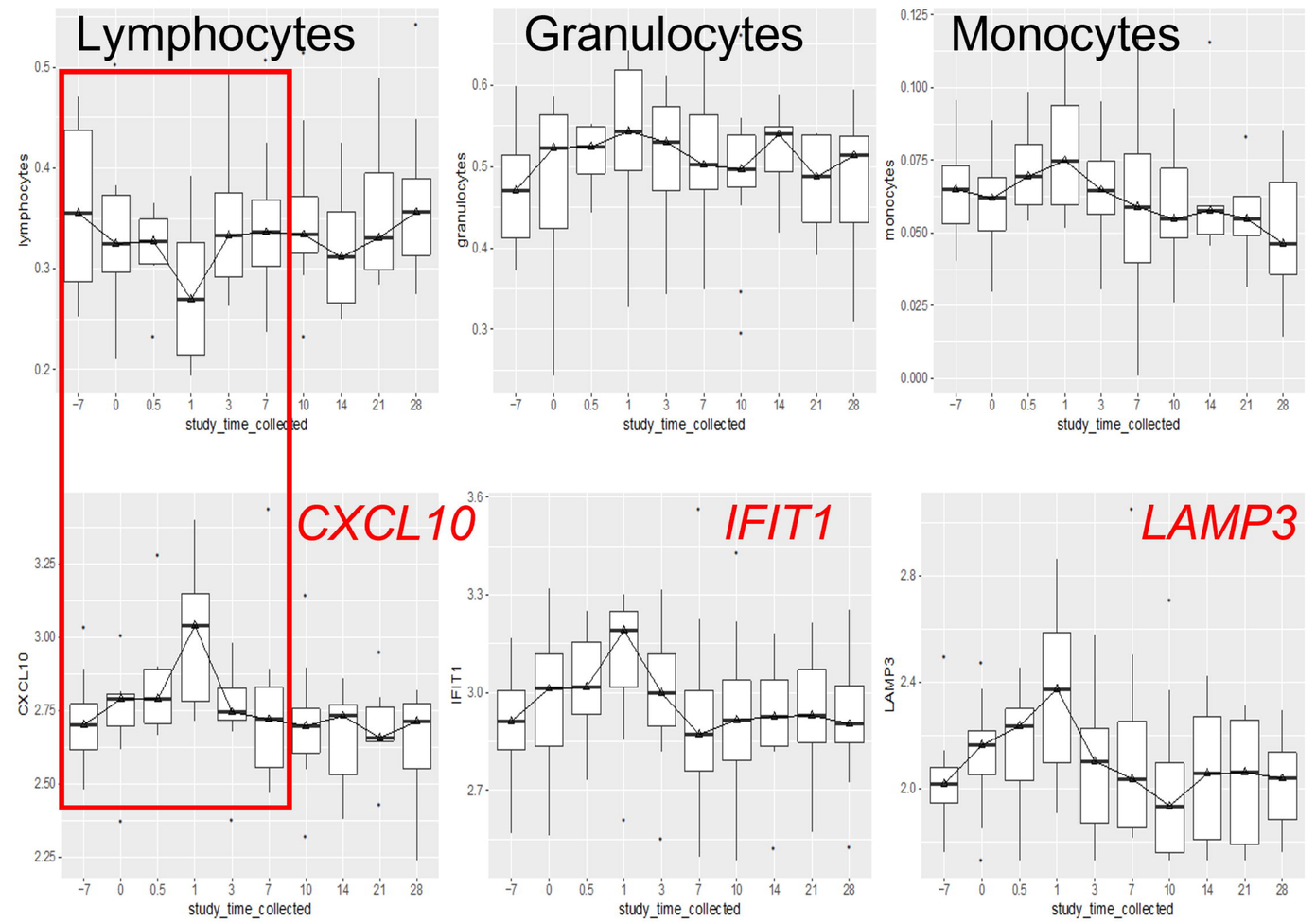
Wilcoxon p-value = 1.9e-16

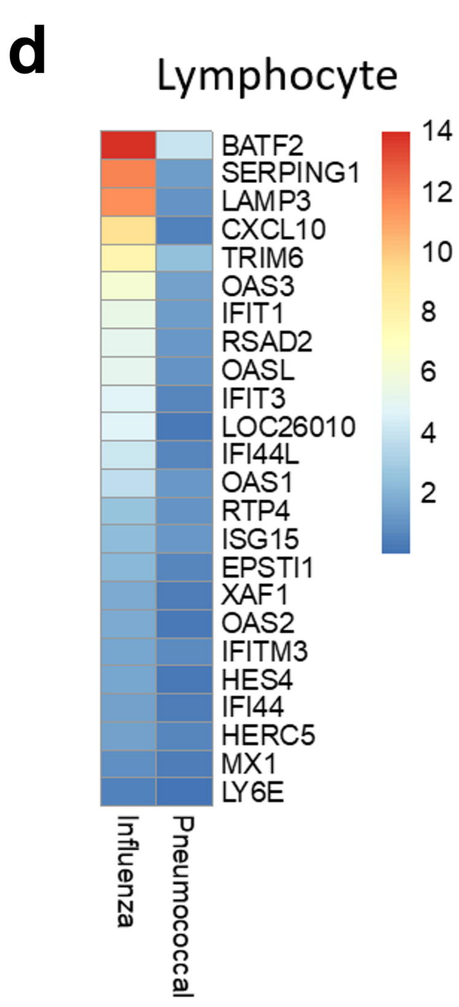
d

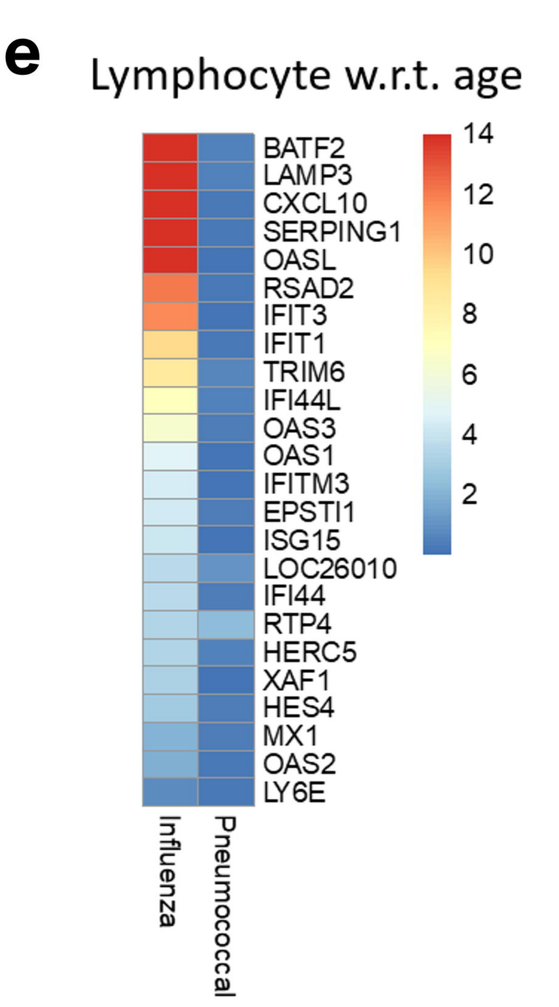


Wilcoxon p-value = 0.12









a

	No DEGs				With DEGs			
	cor=0		cor=0.5		cor=0		cor=0.5	
Method	Time	MSE	Time	MSE	Time	MSE	Time	MSE
`lme4_ind`	1137.9	0.02	854.1	30.70	765.6	1.69	749.9	34.09
`lme4`	8163.6	0.22	9798.9	0.32	8378.2	2.09	9525.6	1.96
`FastMix_ind`	27.9	0.04	27.6	34.3	29.4	0.49	28.7	34.57
`FastMix`	29.3	0.20	27.5	0.21	30.8	0.68	28.8	1.16

b

	OLS	Ridge	FastMix
MSE	2.708 (0.200)	1.765 (0.047)	0.919 (0.022)
Cell1	-0.124 (1.709)	-2.127 (1.821)	-0.060 (0.979)
Cell2	0.108 (1.517)	0.005 (1.637)	0.059 (0.876)
Cell3	-0.102 (1.520)	-0.122 (1.647)	-0.124 (0.876)
Severity	-0.089 (0.986)	3.956 (0.399)	-0.081 (0.618)
Sex	-0.135 (0.855)	-0.087 (0.360)	-0.184 (0.544)
Cell1.Severity	0.015 (1.912)	-34.210 (1.349)	0.003 (1.014)
Cell2.Severity	0.252 (2.107)	9.125 (1.438)	0.212 (1.238)
Cell3.Severity	0.047 (1.895)	8.986 (1.348)	0.068 (1.014)
Cell1.Sex	-0.015 (1.874)	-0.308 (1.291)	0.183 (0.994)
Cell2.Sex	-0.015 (2.078)	0.063 (1.390)	-0.042 (1.188)
Cell3.Sex	0.287 (1.866)	0.298 (1.293)	0.218 (0.996)

c

cor = 0	**Type-I Error**	`csSAM`	`FastMix`
	Cell1.Group	28.99 (11.19)	6.36 (1.11)
	Cell2.Group	17.23 (7.70)	6.31 (1.09)
	Cell3.Group	13.05 (7.61)	5.04 (0.21)
cor = 0.5	**Type-I Error**	`csSAM`	`FastMix`
	Cell1.Group	17.34 (9.11)	6.85 (0.42)
	Cell2.Group	9.71 (5.97)	6.85 (0.46)
	Cell3.Group	6.86 (5.54)	5.00 (0.23)

d

cor = 0	**Power**	`csSAM`	`FastMix`
	Cell1.Group	56.48 (17.75)	61.86 (4.11)
	Cell2.Group	40.54 (16.88)	62.79 (4.36)
cor = 0.5	**Power**	`csSAM`	`FastMix`
	Cell1.Group	62.82 (15.18)	64.22 (6.50)
	Cell2.Group	46.72 (14.88)	64.68 (6.31)

e

Comp. Time	`csSAM`	`FastMix`
cor = 0	209.05	20.82
cor = 0.5	206.96	19.95