Title

FastMix: a versatile data integration pipeline for cell type-specific biomarker inference.

Journal

Bioinformatics, 38(20)

ISSN

1367-4803

Authors

Zhang, Yun
Sun, Hao
Mandava, Aishwarya
et al.

Publication Date

2022-10-14

DOI

10.1093/bioinformatics/btac585

Peer reviewed

Bioinformatics, 38(20), 2022, 4735–4744
https://doi.org/10.1093/bioinformatics/btac585
Advance Access Publication Date: 26 August 2022
Original Paper

OXFORD

Gene expression

FastMix: a versatile 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,*] and Yu Qian[1,*]

[1]Department of Informatics, J. Craig Venter Institute, La Jolla, CA 92037, USA, [2]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA, [3]Systems Vaccinology, Telethon Kids Institute, Perth Children's Hospital, University of Western Australia, Nedlands, WA 6009, Australia and [4]Department of Pathology, University of California, San Diego, La Jolla, CA 92093, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Olga Vitek

Received on January 26, 2022; revised on August 18, 2022; editorial decision on August 21, 2022; accepted on August 25, 2022

Abstract

Motivation: Flow cytometry (FCM) and transcription profiling are the two widely used assays in translational immunology research. However, there is no data integration pipeline for analyzing these two types of assays together with experiment variables for biomarker inference. Current FCM data analysis mainly relies on subjective manual gating analysis, which is difficult to be directly integrated with other automated computational methods. Existing deconvolutional analysis of bulk transcriptomics relies on predefined marker genes in the transcriptomics data, which are unavailable for novel cell types and does not utilize the FCM data that provide canonical phenotypic definitions of the cell types.

Results: We developed a novel analytics pipeline—FastMix—for computational immunology, which integrates flow cytometry, bulk transcriptomics and clinical covariates for identifying *cell type-specific* gene expression signatures and biomarker genes. FastMix addresses the 'large p, small n' problem in the gene expression and flow cytometry integration analysis via a linear mixed effects model (LMER) for both cross-sectional and longitudinal studies. Its novel moment-based estimator not only reduces bias in parameter estimation but also is more efficient than iterative optimization. The FastMix pipeline also includes a cutting-edge flow cytometry data analysis method—DAFi—for identifying cell populations of interest and their characteristics. Simulation studies showed that FastMix produced smaller type I/II errors than competing methods. Validation using real data of two vaccine studies showed that FastMix identified a consistent set of signature genes as in independent single-cell RNA-seq analysis, producing additional interesting findings.

Availability and implementation: Source code of FastMix is publicly available at https://github.com/terrysun0302/FastMix.

Contact: xing_qiu@urmc.rochester.edu or mqian@jcvi.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Different characteristics of the same subject in transcriptomics, proteomics and metabolomics can now be measured using a variety of bioassays. Recent publications (Aevermann *et al.*, 2021; HIPC-I Consortium, 2017; Li *et al.*, 2021; McCall *et al.*, 2021; Noecker *et al.*, 2016; Pinu *et al.*, 2019; Tomic *et al.*, 2019) have shown that the integrative analysis of multi-omics data can identify patterns missed by individual assays. However, how to deal with the large number of experiment variables (p) involved in a multi-omics study, which is often much larger than the number of samples (n), is highly challenging.

Currently, dimensionality reduction and regularization are two major approaches to address this 'large p, small n' problem. For example, DIABLO (Singh *et al.*, 2019) uses sparse generalized canonical correlation analysis (sGCCA) with L^1 regularization [a.k.a. LASSO (Tibshirani, 1996)] to predict patient's disease type from multiple assays. LUCID (Peng *et al.*, 2020) uses latent unknown clusters model with LASSO to integrate multi-omics data. UMAP

and other embedding techniques are frequently used for dimensionality reduction in multi-modal data integration (Cao *et al.*, 2020; Jin *et al.*, 2020). Regularized fixed effects regressions solve the 'large p, small n' issue and reduce the variability in the estimation procedure by shrinking the estimates toward zero (Hoerl and Kennard, 1970; Tibshirani, 1996; Zou and Hastie, 2005). In comparison, a linear mixed effects regression (LMER) shrinks the estimates toward the *fixed effects* instead of zero, which are less biased (Maldonado *et al.*, 2009; Zhang *et al.*, 2020). Conventional LMER fitting algorithms such as lme4 (Bates *et al.*, 2015), the reference implementation of LMER in R programming language, use iterative expectation-maximization (EM) algorithm to fit the model based on the (restricted) maximum likelihood principle. This iterative process is slow and may not converge, making LMER impractical for large-scale data analysis.

We developed a non-iterative, moment-based robust estimation procedure for a wide class of LMER useful for gene expression and flow cytometry data integration. We also designed a novel statistical inference framework customized for the fitted LMER, which not only selects significant fixed effects, but also classifies *informative random effects* based on a mixture model. To exemplify the utility of the proposed method (dubbed `FastMix`), we applied it to integrate three types of data: (i) bulk gene expressions; (ii) proportions of cell populations identified from flow cytometry (FCM); and (iii) experimental and/or clinical covariates. `FastMix` is designed to identify not only cell type-specific differentially expressed genes, a common need in biological studies, but also cell type-specific signature genes, a problem not specifically addressed by the existing methods. Figure 1 depicts the overall structure of `FastMix` including automated analysis of flow cytometry data using DAFi (Lee *et al.*, 2018) for identifying proportions of the cell populations. `FastMix` provide an *in silico* solution for inference of cell type-specific expression signatures that complements the cutting-edge single-cell transcriptomics.

2 Materials and methods

2.1 Motivating example: integrating flow cytometry, transcriptomics and clinical data

In this example, we consider three sets of input data: (i) clinical covariates (denoted as **Clin**), (ii) cell type proportions (denoted as **Cell**) and (iii) bulk gene expression (denoted as Y). The composite tissue data are modeled as

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

Specifically, Y_{ji} is the bulk expression of the ith gene and jth sample; $\mathbf{Cell}_{jk}$ is the proportion of the kth cell type (population) in the jth sample, and b_{kij} is the gene expression contributed solely by the kth cell type and ϵ_{ij} is the error. We propose to associate b_{kij} (cell type-specific gene expression) instead of Y_{ji} with the clinical data as follows

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

Here, β_{ki} is the baseline expression level, $\mathbf{Clin}_{jp}$ 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. For the kth cell type, the ith gene is declared as a cell type-specific differentially expressed gene (csDEG) for $\mathbf{Clin}_{jp}$, if the following null hypothesis is rejected

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

One straightforward approach to identify csDEGs would consist of two stages: (i) apply an *in silico* deconvolution algorithm to estimate $\hat{b}_{kij}$; and (ii) apply a differential gene expression analysis (DGEA) to associate $\hat{b}_{kij}$ with the clinical data. However, Equation (1) is a 'large p, small n' problem because there are approximately Knm unknown parameters (b_{kij}) to be estimated from only nm observations (Y_{ji}). While certain computational methods such as non-negative matrix factorization (Gaujoux and Seoighe, 2012; Lähdesmäki *et al.*, 2005; Repsilber *et al.*, 2010; Venet *et al.*, 2001), regularization (Newman *et al.*, 2015) and Bayesian methods (Qiao *et al.*, 2012; Quon *et al.*, 2013; Quon and Morris, 2009; Zhang*et al*, 2019a,b), can be used to obtain approximate solutions of an underdetermined system for deconvolution (Mohammadi *et al.*, 2017), the bias and variance of the estimated $\hat{b}_{kij}$ are inevitably large, which make downstream DGEA inaccurate.

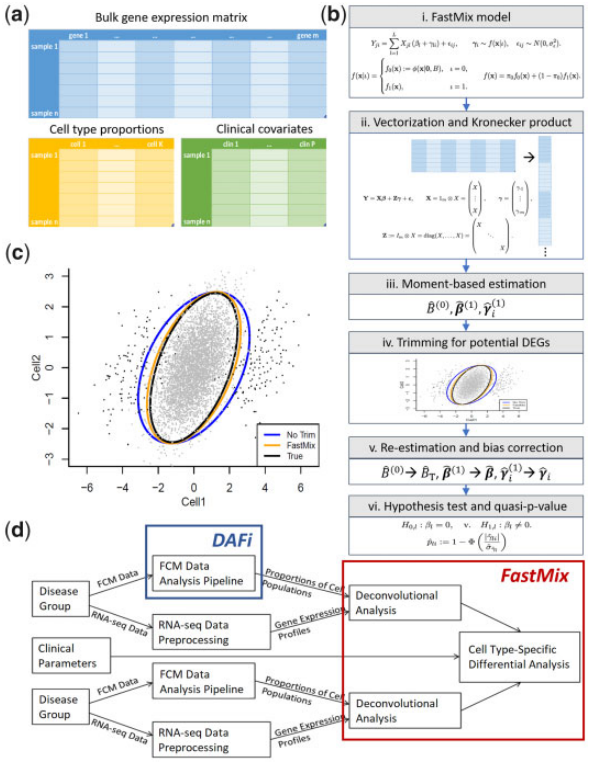

Fig. 1. `FastMix` schematics and analytical pipeline. (**a**) `FastMix` takes three input data matrices: a bulk gene expression matrix, a matrix of cell type proportions, and a matrix of clinical covariates (both continuous and categorical). (**b**) Flow chart of key steps of `FastMix`. (Details please 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 $\hat{\boldsymbol{\beta}}^{(1)}$ and the random effects estimate $\hat{\gamma}_i^{(1)}$, both of which depend on $\hat{B}^{(0)}$. (iv) In `FastMix`, csDEG identification is viewed as an outlier detection problem. It uses a trimming technique to improve the robustness due to the existence of csDEGs (outliers). (v) After trimming, re-estimate the variance-covariance matrix using the robust estimator $\hat{B}_T$ with bias correction, followed by re-estimating $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}_i$ using $\hat{B}_T$. (vi) `FastMix` performs hypothesis test and constructs quasi-P-values that indicate the significance of csDEGs. (**c**) Using trimming improves the estimation of the covariance matrix. Axes are random effect signals of two cell populations (Cell1 and Cell2); dots are simulated data of 5000 genes, among which, 250 genes are true csDEGs in the Cell1 direction (dark 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 as "True" in the legend, $\hat{B}^{(0)}$ that is the initial non-robust covariance estimator shown as "No Trim" in the legend and $\hat{B}_T$ that is the robust covariance estimator based on trimming shown as "FastMix" in the legend. Due to the existence of the true csDEGs (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`

2.2 FastMix model

We propose to jointly model the two-stage analysis in one unified regression model. First, we combine Equations (1) and (2) to obtain:

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

Here, $\tilde{\epsilon}_{ij} = \epsilon_{ij} + \sum_{k=1}^{K} \mathbf{Cell}_{jk} e_{kij}$ is the combined error term. To model the direct association between the bulk gene expression and clinical covariates (a common task in bulk DGEA), we further add a main term $\mathbf{Clin}_{jp}$ in Equation (4). 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 restated in the following standard multivariate regression model

$$Y = XW + E; \;\; Y_{ji} = \sum_{l=1}^{L} X_{jl} \beta_{li} + \epsilon_{ij}\,. \tag{5}$$

Here, X_{jl} is an element in matrix $X := \begin{pmatrix} \mathbf{Cell} & \mathbf{Clin} & \mathbf{Cell} \times \mathbf{Clin} \end{pmatrix}$, which has n rows and $L = K + P + KP$ columns (linear predictors). Both bulk and cell type-specific DGEA (csDGEA) can be made from the estimated linear coefficients (β_{li}) based on this unified model. Note that there are only about mL unknown parameters (β_{li}) in Equation (5), which is much smaller than Knm unknown parameters (b_{kij}) in Equation (1). Equation (5) is no longer a 'large p, small n' regression problem when $L < n$, which is seen in many real-world applications. However, Equation (5) is still a large-scale regression model; so it is tempting to apply regularization techniques such as ridge (Hoerl and Kennard, 1970), LASSO (Tibshirani, 1996) and elastic-net (Zou and Hastie, 2005) to increase numerical stability and improve prediction accuracy. However, these techniques have two drawbacks: (i) they shrink the estimated linear coefficients toward zero and create non-trivial bias; and (ii) the best penalty parameter(s) are typically trained by time-consuming cross-validation (CV) procedures. As an alternative, we propose to use linear mixed effects regression (LMER) to reduce model complexity. Specifically, we decompose the linear coefficients as

$$\beta_{li} = \beta_l + \gamma_{li}, \tag{6}$$

where β_l is the *fixed effect* of $X_{\cdot l}$ to the entire transcriptome, and γ_{li} is the gene-specific *random effect* associated with $X_{\cdot l}$. By combining Equations (5) and (6), we obtain the following general `FastMix` model

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

Compared to regularized regressions, Model (7) does not contain hyperparameters that need to be trained, and shrinks the gene-specific linear coefficients toward the fixed effects (β_l) instead of zero (Maldonado, 2009), thereby achieving variance-reduction with less bias.

While the inference on β_l can be made with a standard regression t-test; no classical hypothesis test is applicable to gene-specific random effects (γ_{li}) for theoretical reasons (Robinson, 1991). Note that in most practical cases, the majority of the genes are non-differentially expressed genes (NDEGs). In this regard, we reconsider the DGEA based on random effects as an *outlier detection* problem, and adapt a non-parametric empirical Bayes method (Efron *et al.*, 2001; Qiu *et al.*, 2005) to perform statistical inference. Let ι be a binary indicator for csDEG ($\iota = 1$) and NDEG ($\iota = 0$). The prior probability of a gene being NDEG or csDEG is $P(\iota = 0) = \pi_0$ or $P(\iota = 1) = 1 - \pi_0$, respectively. The mixture model (MM) of the multivariate vector $\boldsymbol{\gamma}_i = (\gamma_{li}, l = 1, \ldots, L)'$ is

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

where $\mathbf{x} \in \mathbb{R}^L$ is a dummy variable, $f_0(\cdot)$ is the component distribution for NDEGs and $f_1(\cdot)$ is the component distribution for csDEGs.

Furthermore, we 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 (no parametric assumptions are needed for $f_1(\cdot)$); 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

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

Intuitively, Equation (9) implies that, compared with NDEGs, the DEGs can be viewed as 'outliers' (Fig. 1c). From the above assumptions, the marginal distribution for the non-parametric empirical Bayes method is

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

where $\phi(\cdot|0, B)$ 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 complete `FastMix` model of the unified pipeline for the application of csDGEA:

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

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

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

2.2.1 Computationally efficient `FastMix` algorithm

Conventional algorithms to fit a large LMER model such as Equation (7) with high-throughput data are not only time consuming but also prone to convergence issues and non-uniformity (the DEGs and NDEGs do not follow the same distribution) in the data. To address these challenges, we designed a novel LMER fitting algorithm based on moment matching and trimming. The following sections provide high-level descriptions of key steps in `FastMix`. Technical details, including derivations, proofs and step-by-step procedures are provided in Supplementary Text.

2.2.2 Vectorization and Kronecker product

The `FastMix` LMER model can be concisely represented in vectorization form using Kronecker product (Horn *et al.*, 1994):

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \epsilon, \\ \mathbf{X} := 1_m \otimes X &= \begin{pmatrix} X \\ \vdots \\ X \end{pmatrix},\; \gamma := \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_m \end{pmatrix}, \\ \mathbf{Z} := I_m \otimes X &= \begin{pmatrix} X & & \\ & \ddots & \\ & & X \end{pmatrix}. \end{aligned} \tag{11}$$

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.

2.2.3 Moment-based estimation

An initial estimation of the linear coefficients, $\hat{\boldsymbol{\beta}}_i^{(0)} = (\hat{\beta}_{li},\; l = 1, \cdots, L)'$, can be obtained through fitting the multivariate linear regression in Equation (5) using the ordinary least squares (OLS) criterion for each gene. Denote the sample covariance matrix

of $\hat{\boldsymbol{\beta}}_i^{(0)}$ as $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}^{(0)}} \in M_{L\times L}$. Even for NDEGs ($\iota = 0$), $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}^{(0)}}$ is not an unbiased estimator of B because

$$E\left(\hat{\Sigma}_{\hat{\boldsymbol{\beta}}^{(0)}} | \iota = 0\right) = B + \sigma_\epsilon^2 (X\iota X)^{-1}. \quad (12)$$

Based on the assumption that most genes are NDEGs, we propose the following bias-corrected, moment-based estimator for an initial estimation of B

$$\hat{B}^{(0)} := \hat{\Sigma}_{\hat{\boldsymbol{\beta}}^{(0)}} - \hat{\sigma}_\epsilon^2 (X\iota X)^{-1}. \quad (13)$$

Based on $\hat{B}^{(0)}$, We apply the weighted least squares (WLS) method to compute $\hat{\boldsymbol{\beta}}^{(1)}$, and use the *empirical best linear unbiased predictor* (EBLUP) to compute $\hat{\gamma}_i^{(1)}$.

2.2.4 Trimming

Recall that there is a small but important subset of csDEGs presented in the data. When the csDEGs are present, $\hat{\gamma}_i^{(1)}$ no longer follows a multivariate normal distribution. Of note, a csDEG for one covariate may be an NDEG for another covariate. It is also possible that a covariate is not associated with any gene (we call it an *uninformative covariate*). We designed a trimming procedure to remove the negative impact of csDEGs to the parameter estimation for `FastMix`: (i) use an information-based criterion to identify informative covariates; (ii) for informative covariates, remove genes that are 'outliers' based on a Mahalanobis distance. The remaining genes, denoted as $S_0 \subseteq \{1, \ldots, m\}$, will be used to refine the initial estimation.

2.2.5 Re-estimation and bias correction

Based on our previous work on gene expression normalization (Cui *et al.*, 2021; Liu *et al.*, 2017; Qiu *et al.*, 2013, 2014), we know that trimming can introduce noticeable bias in statistical inference. We derive the trimmed and bias-corrected covariance estimates for B as follows

$$\hat{B}_T := \Lambda^{-1/2} \tilde{\Lambda}^{1/2} \hat{\Sigma}_{\hat{\boldsymbol{\beta}}^{(0)}} \tilde{\Lambda}^{1/2} \Lambda^{-1/2} - \hat{\sigma}_\epsilon^2 (X'X)^{-1}. \quad (14)$$

After obtaining $\hat{B}_T$, we adopt a method in Appendix 1 of Cui *et al.* (2021) to correct bias in $\hat{\boldsymbol{\beta}}$ induced by trimming. Finally, random effects are re-computed with EBLUP using bias-corrected covariance and fixed effects. Figure 1c illustrates the advantage of the trimming and re-estimation procedures.

2.3 Hypothesis test and quasi-*P*-value

Hypotheses about β_l can be tested with standard regression *F*- and *t*-tests. These results can be interpreted as whether $X_{\cdot l}$ has significant association with the whole transcriptome. On the other hand, we cannot test $H_0 : \iota = 0$ (i.e. NDEG) versus $H_1 : \iota = 1$ (i.e. csDEG) because ι is a random variable, not a parameter. Instead, we develop the following *P*-value-like quantity (called the 'quasi-*P*-value') to identify genes that have extremely large/small random effects based on their distribution:

$$\hat{p}_{li} := 1 - \Phi\left(\frac{|\hat{\gamma}_{li}|}{\hat{\sigma}_{\hat{\gamma}_l}}\right). \quad (15)$$

Here, $\Phi(\cdot)$ is the standard normal distribution function. Although $\hat{p}_{li}$ is not a classical *P*-value, in practice, it is a pragmatic and efficient way to rank and select genes with strong association with the lth covariate, which are the central inference output from the `FastMix` model. 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 DGEA,
- **Cell × Clin**—cell type-specific DGEA (csDGEA).

2.4 Weighted `FastMix` model

We implemented weights (for covariance matrix) in the `FastMix` model so that users can: (a) incorporate quality scores to weigh samples, and (b) account for serial correlation in longitudinal studies. If unknown, the weighted covariance matrix can be estimated by techniques we presented in (Zhang *et al.*, 2019a,b) [`getSigma()` function from the `PBtest` R package]. See Supplementary Material, Supplementary Section S2.3, for implementation details.

2.5 Discriminant analysis

By combining LMER and MM, we can define four types of discriminant scores in `FastMix` that predict a binary response (e.g. response to a vaccination) from the input data (e.g. bulk gene expressions, cell proportions and subject demographics): (i) `single_score`, an 1-dimensional score based on all input genes; (ii) `single_sparse_score`, a 1-dimensional score based on genes with significant interactions with the response; (iii) `multi_score`, an n-dimensional score based on all genes; and (iv) `multi_sparse_score`, a multivariate score based on genes with significant interactions with the response. See Supplementary Text, Supplementary Section S5 for technical details.

The analysis of flow cytometry data of the two vaccine studies using DAFi (Lee *et al.*, 2018) for identifying proportions of the immune cell populations can be found in Supplementary Materials.

3 Results

3.1 Simulation studies

We outline three simulation studies designed to demonstrate the utility and advantages of `FastMix`. Technical details are described in Supplementary Text, Supplementary Section S3.

3.1.1 Simulation I: robustness in estimating covariance matrices

We simulated two cell populations with proportions Cell1 and Cell2, with random effects γ_{1i} and γ_{2i}, whose covariance structure is shown with the black ellipse in Figure 1c. We simulated expressions of 5000 genes (dots); among them, 250 genes were true DEGs (black dots) of Cell1. Figure 1c compares the true covariance with those estimated from the standard method without trimming (large grey ellipse) and the proposed robust and bias-corrected covariance estimator $\hat{B}_T$ (small grey ellipse).

Simulation I showed that the proposed covariance estimator $\hat{B}_T$ was robust to the existence of outliers (DEGs), and accurately recapitulated the true covariance matrix, i.e. the overlay of the grey ellipse and the black ellipse.

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

We generated synthetic gene expression values for 5000 genes and 50 samples. For each sample, we simulated three cell population proportions (Cell1, Cell2 and Cell3), one continuous clinical covariate (Severity) and one categorical clinical covariate (Sex). Table 1 reports computational cost and mean square error (MSE) for estimating B using lme4 and `FastMix`, with independent (denoted as lme4_ind and `FastMix_ind`) and dependent covariance structure.

When random effects are independent and without csDEGs, lme4_ind achieved the best MSE (0.04). Using only 2% of the computational time of lme4_ind, `FastMix_ind` had a slightly larger but comparable MSE (0.04). When random effects were correlated and without csDEGs, `FastMix` (with general covariance structure) had the smallest MSE (0.21) and was more than 300 times faster than lme4, which had the second best MSE (0.32). With csDEGs, lme4-based approaches had large MSEs, because they were not robust to the presence of outliers (csDEGs). In comparison, `FastMix` approaches had much smaller MSEs and used tiny amount of computational time. In Supplementary Material, Supplementary Sections S3.2 and S3.3, we showed that `FastMix` also greatly reduced the bias in regression coefficients compared to the lme4

Table 1. Simulation performance

	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	30.7	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	0.2	27.5	0.21	30.8	0.68	28.8	1.16

Note: 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 csDEGs, and with or without correlation between random effects. Mean computational time (in seconds) and mean MSE are reported. MSE is defined as $\sum_{i=1}^{p}\sum_{j=1}^{p} 1/p^2(\hat{B}_{ij} - B_{ij})^2$. Simulations are repeated 200 times.

approach and other robust covariance estimators (Maronna and Yohai, 1995; Maronna and Zamar, 2002; Rousseeuw and Driessen, 1999). Among all the methods compared, `FastMix` had the most robust performance (Supplementary Material, Supplementary Tables S1 and S3).

Next, we compared `FastMix` with ordinary least square (OLS) and Ridge regression for regression coefficient estimation, using the most realistic scenario, i.e. with correlation and csDEGs. The regularization parameter in ridge regression was selected by the generalized cross-validation (GCV) criterion. Table 2 showed that: (i) `FastMix` and Ridge estimates were more accurate than OLS due to the shrinkage effect; and (ii) Ridge regression had much larger bias than `FastMix` and OLS (both are practically unbiased), because Ridge regression shrank the estimates toward zero, not the fixed effects (`FastMix`).

3.1.3 Simulation III: comparing `FastMix` with existing csDGEA method

We compare `FastMix` with csSAM, a popular csDGEA method for heterogeneous biological samples using gene expression data and relative cell type frequencies based on deconvolution techniques. Because csSAM can only performs two-group comparison (binary covariate), we designed Simulation III based on simulation II but contained only one binary covariate (Group) for this comparison. The results are summarized in Table 3a–c. Overall, `FastMix` had acceptable type-I error rates (5–7%) which were much lower than those of csSAM (Table 3a). `FastMix` also had better power for detecting csDEGs than csSAM (Table 3b). Notably, `FastMix` used just 1/10 of the computational time of csSAM (Table 3c).

In Supplementary Section S3.5, we tried a variant of Simulation III with unbalanced csDEGs (more up-regulated csDEGs than down-regulated csDEGs). The results were similar to those shown in Table 3a–c. In order to evaluate the model performance for varying DAFi outputs, we conducted another variant of Simulation III, in which proportions of cell populations ($\mathbf{Cell}_{ik}$) are simulated as a mixture of the original DAFi-output values and noise. We found that the performance of `FastMix` was robust to noisy DAFi output ($\mathbf{Cell}_{jk}$). These results can be found in Supplementary Tables SB–SE in Supplementary Data S6.

3.2 Real-data analyses

3.2.1 `FastMix` integration reveals consistent cell type-specific signature genes with scRNA-seq

We applied `FastMix` to a multi-modal study [HVP01, the Human Vaccine Project (Shannon *et al.*, 2020)] that measures human immune responses to a licensed hepatitis B vaccine—Engerix-B. The HVP01 study provides us data from flow cytometry, bulk RNA-seq and virus neutralization (anti-HBs) on blood samples from adults. In

Table 2. Simulation performance

	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)

Note: 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 $1/(mp)\sum_{i=1}^{m}\sum_{j=1}^{p}(\hat{\beta}_{ij} - \beta_{ij})^2$. 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.

Table 3. Simulation performance

a

cor = 0	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)
cor = 0.5	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)

b

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)

c

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

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

addition, it also has single cell RNA-sequence (scRNA-seq) data for immune cells using the Smart-Seq2 (Picelli *et al.*, 2014) protocol, which we used as the ground truth to validate FastMix results.

After Dose 3 of Engerix-B, all 15 subjects responded to vaccination, but some had much higher level of anti-HBs titer than others (Supplementary Fig. S1). Using anti-HBs titer > 5000 mUI/mL as cutoff, we separated the subjects into two groups: high responders (5 subjects) and low responders (10 subjects). Then we identified neutrophils ($CD45^+CD66^+$) and non-neutrophils ($CD45^+CD66^-$) following the DAFi gating hierarchy (Fig. 2a). Using all 5 time points (Day 0,1,3,7,14), a weighted FastMix model is fitted to integrate the bulk RNA-seq gene expressions, proportions of cell populations and clinical covariates including response groups and ages (Supplementary Fig. S1) for identification of neutrophil-specific genes that may regulate the anti-HBs response levels.

Out of the 13 157 genes in the processed bulk RNA-seq data, 851 of them are identified by FastMix as neutrophil specific and 520 as non-neutrophil specific. To validate these 851 neutrophil-specific genes, we used the independent scRNA-seq data. We followed a standard scRNA-seq analysis pipeline including low-dimensional embedding of cells on UMAP (Blondel *et al.*, 2008) and a non-parametric hypothesis testing approach for single cell DEG detection (Aevermann *et al.*, 2021). The UMAP visualization of the ground truth cell types (Fig. 2b) showed a good separation of the neutrophil population from other cell populations. 2744 neutrophil-specific DEGs (a.k.a. signature) were identified by the scRNA-seq data analysis from the total 58 036 annotated genes.

Figure 2c compared the FastMix and scRNA-seq results. The majority (>50%) of the FastMix identified genes were also found by the scRNA-seq analysis. In fact, 72% of the top 100 FastMix genes were the same as identified by the scRNA-seq. The overlapping rate gradually decreased as we included more top genes in the comparison, meaning that FastMix ranked 'ground truth' (scRNA-seq) signature genes at the top in its DEG list. We further selected 365 scRNA-seq signature genes that have substantial fold change (FC), i.e. |logFC| > 1. The Venn diagram (Fig. 2d) showed that 39 of the top 100 FastMix signature genes were found in the list. That is, only 16% of the scRNA-seq signature genes versus 39% of the top 100 FastMix signature genes could pass the logFC threshold. These 39 genes include the *IFIT* and *IFITM* family genes (*IFIT2, IFITM2, IFITM3)* for interferon-induced proteins. Many of them are highly relevant to neutrophils and Hepatitis B in literature review, e.g. *CXCR1/2* plays an important role (Khanam *et al.*, 2017) in hepatic inflammatory response (Xu *et al.*, 2016). Furthermore, we plotted in violin plots the scRNA-seq expression values of the 39 overlapping genes across multiple cell types (Fig. 3a), compared with the bottom genes (Fig. 3b) and the top genes (Fig. 3c) identified by FastMix. We clearly see in the neutrophils for the high expressions of the overlapping and the top genes, and low expressions of the bottom genes.

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

Next, we compared FastMix and csSAM (Shen-Orr *et al.*, 2010), the only pre-existing method that can process data from both flow cytometry and bulk transcriptomics together, for identifying the neutrophil-specific DEGs. We chose to focus on neutrophils here because neutrophils play important roles in pathogenesis of liver diseases and immune responses to HBV vaccines (Khanam *et al.*, 2017; Le *et al.*, 2017). Using the default 5% FDR cutoff, csSAM performs two-group DE analysis by a cell type-specific SAM model, which identified no DEG for neutrophils. In contrast, FastMix identified 495 neutrophil-specific DEGs using the same 5% FDR.

We performed enrichment analysis to interpret the genes identified by FastMix and csSAM. For a fair comparison, we extracted the same number of csDEGs from results of FastMix and csSAM. The top 100 genes from each method were fed into ReactomePA (Yu and He, 2016) 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 Tables S13–S15). Figure 4a showed the enriched pathways identified from the top 100 FastMix neutrophil-specific DEGs for Engerix-B high responders. The interferon (IFN) immune signaling pathways were substantially

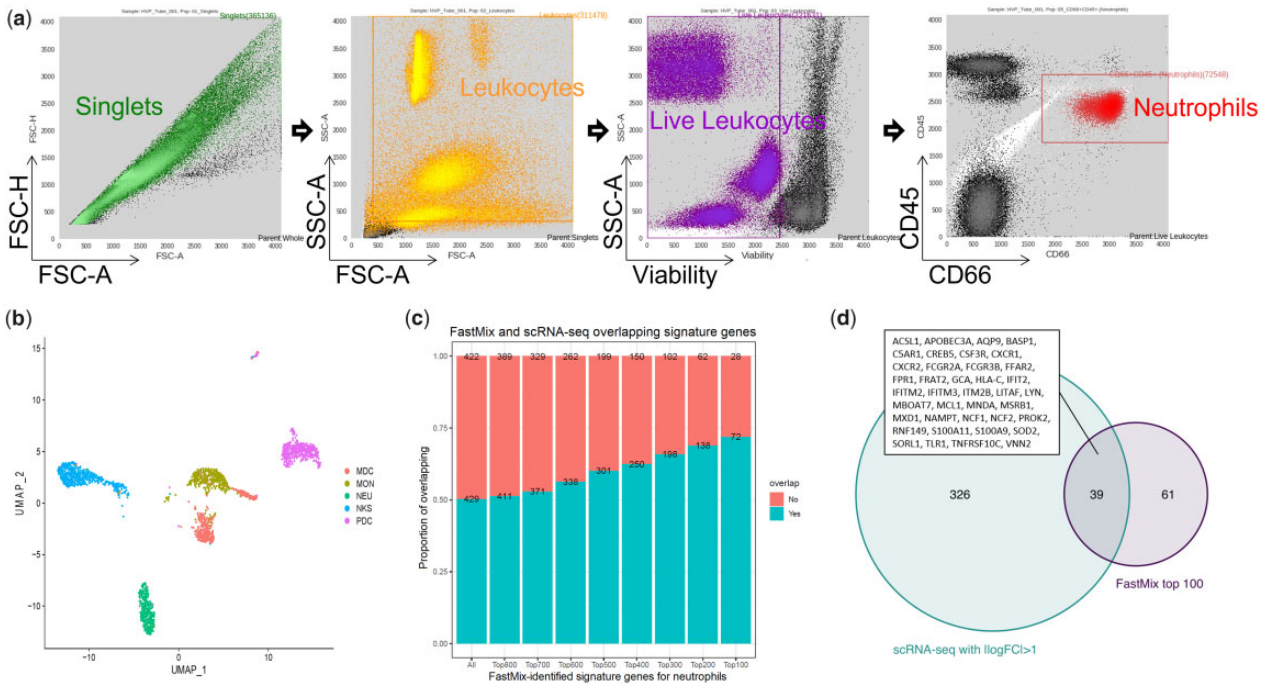

Fig. 2. FastMix and scRNA-seq results for HVP01 study. (**a**) DAFi gating strategy to identify singlets, leukocytes, live leukocytes, CD66- CD45+ population (parent: live leukocytes) and CD66+ CD45- population (parent: live leukocytes). (**b**) UMAP visualization of scRNA-seq cell type clusters. Cells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

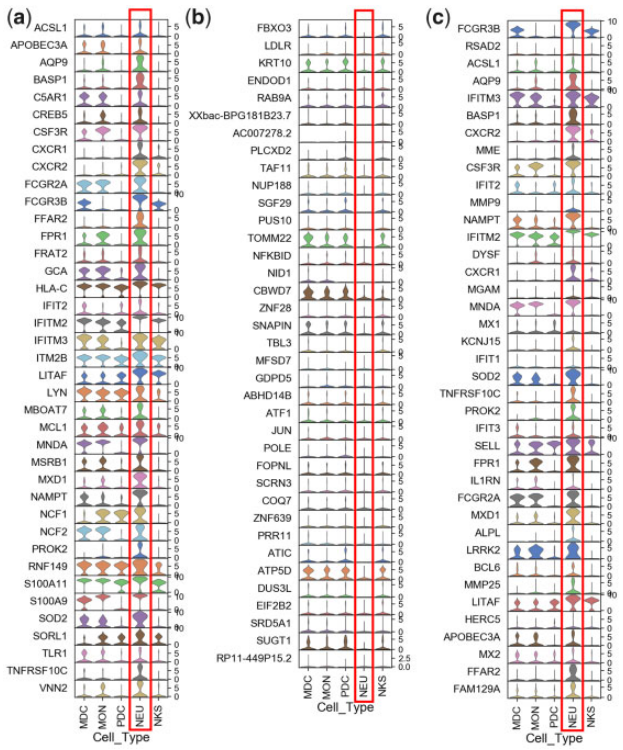

Fig. 3. Expression of neutrophil-specific signature genes in the scRNA-seq experiment. (**a**) The 39 common signature genes identified by `FastMix` and scRNA-seq analysis (same in Figure 2c). (**b**) The bottom 39 genes ranked by `FastMix`. (**c**) The top 39 genes ranked by `FastMix`

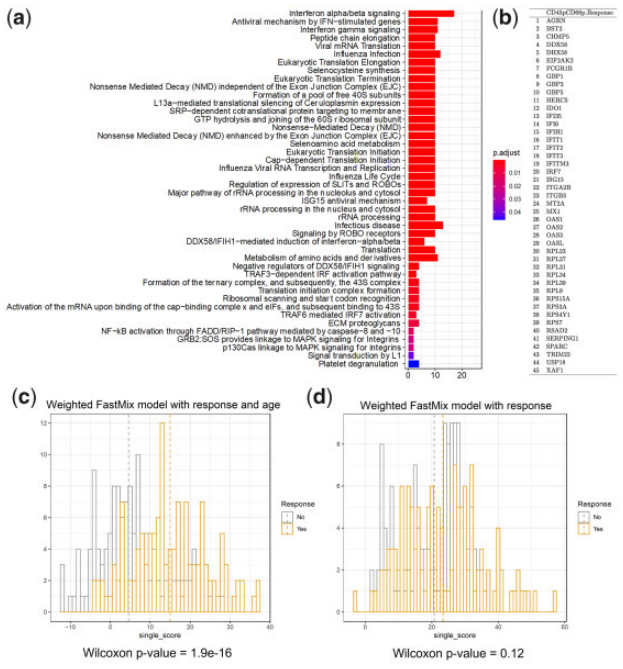

Fig. 4. Pathway enrichment analysis for HVP01 study. (**a**) Enriched pathways identified by the top 100 `FastMix` neutrophil-specific DEGs for high responders. (**b**) Unique genes from the CD45pCD66p.Response (i.e. neutrophil and high response) interaction DEG list that are identified in the enriched pathways in (**a**). (**c**) Scores of weighted FastMix when both age and response are included as covariates, with significant *P*-value indicating age is a factor in vaccine response. (**d**) Scores of weighted FastMix without age. The *P*-value is insignificant

presented (the top 4 pathways). In contrast, using the top 100 csSAM genes identified 0, 1 (neutrophil degranulation) and 5 enriched pathways for the neutrophils, non-neutrophils and rest population, respectively, which seem problematic (Supplementary Table S16). Further, *BST2* (Tetherin/CD317) is found in the FastMix top 100 gene list. Tetherin is a key host cell defense molecule in response to stimuli from IFN pathway (Blasius *et al.*, 2006; Sarojini *et al.*, 2011). Traditional understanding of *BST2* expression is with mature B cells and plasmacytoid dendritic cells while it has cell type-dependent variation (Miyagi *et al.*, 2009). Our analysis showed that *BST2* was also expressed in neutrophils, whose increased expression level is correlated with the high anti-HB levels after Dose 3 of Engerix B (estimated linear coefficient = 1.016).

3.2.3 Inferring cell type-specific temporal pattern from longitudinal data

We have also evaluated how FastMix can be applied to identify signature expression patterns of immune cell-specific genes over time points before and after vaccination. We downloaded SDY180 from ImmPort [56] (www.immport.org/shared/home), which represents typical systems immunology approaches for studying Influenza (102 samples from 12 subjects across at least 8 time points) and Pneumococcal (Pneumovax23; 100 samples from 12 subjects over at least 8 time points, Supplementary Table S17) vaccines (Obermoser *et al.*, 2013). We chose to focus on lymphocytes and neutrophils (Fig. 5a and Supplementary Fig. S2) based on previous findings (Tang *et al.*, 2019). In addition to the 8 time points in the design of SDY180, subject age is also included as a covariate for weighted `FastMix` modeling of cell type-specific immune responses.

The temporal change of proportions of the lymphocytes, granulocytes, monocytes and rest population for the Influenza arm are shown in Figure 5b. With the flow cytometry data only, we noticed that the proportion of lymphocytes had a substantial drop on Day 1 after vaccination and was recovered by Day 3. The 'ground truth' we used to interpret the FastMix findings is the gene modules identified by the original study (Obermoser *et al.*, 2013). Using a pre-post (i.e. between day 0 and day 1) comparison, the original analysis (Obermoser *et al.*, 2013) curated an interferon module, namely M1.2, which includes genes (*CXCL10*, *IFIT1* and *LAMP3*) showing significant *global* changes in blood transcript abundance between the baseline Day 0 and Influenza Vaccine Day 1 at the bulk level (Fig. 5c and Supplementary Fig. S3) and returning to baseline after Day 3.

FastMix identified both *specific* cell populations and their signature genes that are associated with the temporal activation of these M1.2 interferon genes. Among the 24 gene in M1.2, `FastMix` identified 22 genes with significant *P*-values ($P < 0.05$) for lymphocyte-specific differential expression (Fig. 5d and Supplementary Table S18); the top 9 M1.2 genes were found in the top 1% (out of 10 732 genes) of the lymphocyte-specific DE list by FastMix. Interestingly, the majority of the M1.2 genes showed no significance for granulocytes and monocytes in FastMix analysis (Supplementary Table S18), which suggest that the activation of the interferon genes is lymphocyte-specific: the differential expression of interferon signaling genes is driven by the up-regulation of the lymphocyte-specific expression. This increased expression is not because of the abundance of lymphocytes. In fact, the proportion of lymphocytes decreased on Day 1 after vaccination (Fig. 5b, in rectangle) when the bulk expression of M1.2 genes increased on Day 1 (Fig. 5c). Further, `FastMix` produced positive estimated coefficients for lymphocytes for all M1.2 genes (Supplementary Table S19), confirming the up-regulation of the cell type-specific gene expressions. The lymphocyte-specific interferon activation was observed only in the Influenza arm, but not in the Pneumococcal arm (Supplementary Tables S18 and S20), agreeing with the existing knowledge regarding differences between virus and bacterial infections as reported in the original study (Obermoser *et al.*, 2013).

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 have

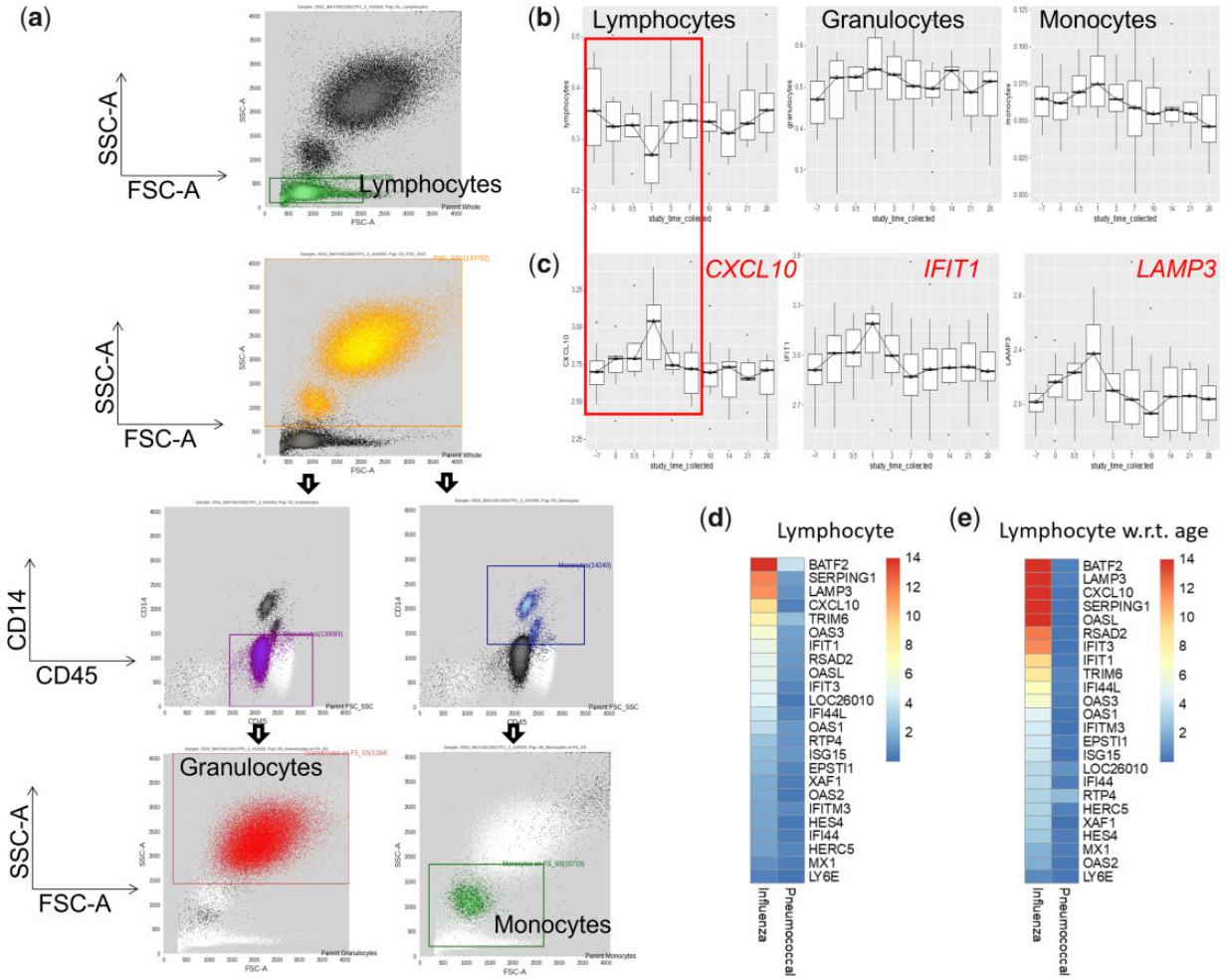

Fig. 5. `FastMix` analysis for SDY180. (**a**) DAFi gating strategy to identify lymphocytes, granulocytes and monocytes, CD45+ CD14- (parent: granulocytes and monocytes), CD45+ CD14+ (parent: granulocytes and monocytes), granulocytes (parent: CD45+ CD14-) and monocytes (parent: CD45+ CD14+). (**b**)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

significant *P*-values, Figure 5e and Supplementary Table S20), whose coefficient estimates showed negative association between the subject age and lymphocyte-specific expression (Supplementary Table S19 and Fig. S4).

For completeness of method comparison, we tried to apply csSAM (Shen-Orr *et al.*, 2010) to compare the pre- and post-vaccination groups for identifying csDEGs. Unlike FastMix, the design of csSAM does not include a way to handle the within-subject correlation across multiple time points. Therefore, no significant cell type-specific DEGs w.r.t. the pre- and post-vaccination groups were identified by csSAM at the 5% FDR level. Using the top 100 csSAM genes for each cell type, no enriched pathway was identified for any cell type.

3.2.4 Discriminant analysis with `FastMix`

Finally, we used HVP01 data to demonstrate the utility of discriminant analysis based on `FastMix`. For illustration, two versions (with and without age) of 'single_sparse_score' that predict the response to HBV vaccination were computed and plotted in Figure 4c and d. There was significant (Wilcoxon *P*-value <0.0001) difference between the responding (dark grey) and non-responding (grey) groups only when age was included in the analysis (Fig. 4c). It suggests that age is highly relevant in host immune response to the HBV vaccine.

4 Discussion

In this study, we developed an efficient and robust data integration framework called `FastMix` based on large-scale LMER models and MM. We demonstrate the utility of `FastMix` by applying it for integrating flow cytometry, bulk transcriptomics and clinical data, with both cross-sectional and longitudinal data. Classical LMER fitting algorithms, such as lme4, use iterative EM-based algorithm to fit the model. In comparison, `FastMix` fits the model with a robust non-iterative algorithm with built-in trimming and bias-correction. Using extensive simulation studies, we showed that `FastMix` produced more accurate estimates than lme4 and other competing methods, with only a fraction of computational cost (Tables 1 and 2).

Inspired by competitive tests used in gene set enrichment analyses (Gatti *et al.*, 2010; Wu and Smyth, 2012; Zhang *et al.*, 2017), we designed a quasi-*P*-value to rank and select csDEGs that have significantly larger/smaller random effects (cell-type-specific effects) than most other genes. `FastMix` can also produce discriminative scores which quantify the contributions of model variables to the

classification of samples. This is a practical feature that previous methods have not provided.

We compared the type-I error rate and statistical power of `FastMix` for csDGEA with a reference pipeline, csSAM [56], using both simulations and real data. `FastMix` achieved better statistical power with much lower type-I error than csSAM, using about 10% of csSAM's run time (see Table 3a–c). A subsequent benchmark simulation study suggests that the runtime of FastMix is linear to the dimensions of the input data (number of genes and samples). These results can be found in Supplementary Table SA, Supplementary Data S6. We also notice that our method is robust to measurement error in cell proportion (Supplementary Tables SB–SE, Supplementary Data S6).

We applied `FastMix` to analyze multi-modal data from two clinical studies (Obermoser *et al.*, 2013; Shannon *et al.*, 2020) that measured host responses to three different vaccines (influenza, pneumococcal and hepatitis B). Input data included bulk gene expressions, FCM and various clinical covariates.

Due to the lack of ground true in real data, multi-modal data integration methods were mostly evaluated by subjective interpretation of the results. We addressed this issue by using scRNA-seq data as an objective gold standard. Excitingly but not surprisingly, csDEGs selected by `FastMix` overlapped significantly with those selected by the scRNA-seq analysis (Fig. 3). Furthermore, `FastMix`-identified biomarker genes are complementary to results from the scRNA-seq analysis. For example, `FastMix` identified the neutrophil-specific genes *MMP9* and *RSAD2*/Viperin (Fig. 3c), which were not found in the scRNA-seq analysis (Fig. 3a). MMP9 is a regulatory factor in neutrophil migration (Kolaczkowska *et al.*, 2009) and Viperin is an important anti-viral protein induced in neutrophils (Hinson *et al.*, 2010) (Fig. 3c). Also, `FastMix` identified the IFIT gene family members (*IFIT1, IFIT2* and *IFIT3*) that can limit the HBV replication (Pei *et al.*, 2014). In summary, `FastMix` provides an *in silico* alternative when scRNA-seq data is unavailable or unreliable.

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 analyzing and integrating these data are suboptimal. FCM data analysis mainly relies on subjective manual gating analysis, which is difficult to be integrated with other computational modules. csDGEA relies on predefined marker genes in the transcriptomics data, without utilizing the FCM data that provide canonical phenotypic definitions of the cell types. We recently developed an automated and objective FCM data analysis pipeline, DAFi, that produces more accurate proportions of cell populations. Combining DAFi and `FastMix` (Fig. 1d) produces an end-to-end, unbiased solution for immunologists to investigate the interplay between FCM, transcriptomics data and clinical covariates.

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. This challenge can potentially be solved if there are replicates of the same measurement, which unfortunately are usually unavailable in most studies. Inference of rare cell populations can also benefit from longitudinal (and repeated) measurements when they are available, which will be investigated in our future work.

Although `FastMix` is developed based on the normality assumption, it is applicable to properly pre-processed RNA-seq data (e.g. HVP01 data used in this study) based on our experiment results. While raw RNA-seq reads are discrete, common pre-processing steps can make the RNA-seq data much more normal. Recent comparative studies (Cui *et al.*, 2021; Law *et al.*, 2014; Rapaport *et al.*, 2013; Ritchie *et al.*, 2015; Smyth *et al.*, 2005) showed that DGEA designed for continuous data can achieve comparable, sometimes even slightly better, performance than those based on discrete models. We plan to extend `FastMix` based on generalized linear mixed-effects regression, so that it can integrate high-dimensional data that cannot be approximated by normal distributions.

By design, FastMix has reduced complexity inherent to deconvolution methods such as csSAM. However, FastMix still will fail if L (total number of covariates) is greater than n (sample size). This limitation motivates us to explore ways to combine regularization techniques designed for LMER (Adjakossa and Nuel, 2017; Li *et al.*, 2018) with the moment-based estimation methods in `FastMix` in a future study.

The generic design of `FastMix` allows it can potentially be applied to address many other bioinformatics problems. For example, based on microbial community composition data and bulk metabolomics data, we may adapt `FastMix` to infer contributions of individual species to the metabolomic profiles, modulated by clinical covariates. Individual components of `FastMix`, such as the bias-correction for trimmed parameter estimation based on Mahalanobis distance; and the quasi-P-value for hypothesis tests that involves random effects, can also be useful for other applications.

Funding

This work was partially funded by the National Institute of Allergy and Infectious Diseases [NIAID grant number UH2AI132342], the National Center for Advancing Translational Sciences [NCATS grant number U01TR001801], the Human Vaccines Project [HVP01 and HVP Innovation Award], the Respiratory Pathogens Research Center [NIAID contract number HHSN272201200005C] and the University of Rochester [NCATS CTSA award number UL1TR002001] from the National Center for Advancing Translational Sciences of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: none declared.

Data availability

Both DAFi and `FastMix` are freely accessible (https://github.com/JCVenterInstitute/DAFi-gating and https://github.com/terrysun0302/FastMix) as open-source software packages. Real data used in this study are available at https://clinicaltrials.gov/ct2/show/NCT03083158 (HVP01) and http://www.immport.org (SDY180). Specifically, RNA-Seq data of HVP01 is available at NCBI's Gene Expression Omnibus (GEO) under GSE155198. Flow cytometry data of HVP01 is available at flowRepository with the ID FR-FCM-Z2R9. Raw and expression matrix for the single cell RNAseq data of HVP01 are available at dbGaP accession number phs002508.v1.p1, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002508.v1.p1. Data preprocessing details are provided in Supplementary Material.

References

Adjakossa,E. and Nuel,G. (2017) Fixed effects selection in the linear mixed-effects model using adaptive ridge procedure for L0 penalty performance. *arXiv preprint arXiv:170501308*. https://doi.org/10.48550/arXiv.1705.01308.

Aevermann,B.D. *et al.* (2021) Machine learning-based single cell and integrative analysis reveals that baseline mDC predisposition correlates with hepatitis B vaccine antibody response. *Front. Immunol.*, **12**, 690470.

Bates,D. *et al.* (2015) Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Blasius,A.L. *et al.* (2006) Bone marrow stromal cell antigen 2 is a specific marker of type I IFN-producing cells in the naive mouse, but a promiscuous cell surface antigen following IFN stimulation. *J. Immunol.*, **177**, 3260–3265.

Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **2008**, P10008.

Cao,K. *et al.* (2020) Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, **36**, i48–i56.

Cui,Z. *et al.* (2021) Super-delta2: an enhanced differential expression analysis procedure for multi-group comparisons of RNA-seq data. *Bioinformatics*, **37**, 2627–2636.
Efron,B. *et al.* (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.
Gatti,D.M. *et al.* (2010) Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics*, **11**, 574–510.
Gaujoux,R. and Seoighe,C. (2012) Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. *Infect. Genet. Evol.*, **12**, 913–921.
Hinson,E.R. *et al.* (2010) Viperin is highly induced in neutrophils and macrophages during acute and chronic lymphocytic choriomeningitis virus infection. *J. Immunol.*, **184**, 5723–5731.
HIPC-I Consortium. (2017) Multicohort analysis reveals baseline transcriptional predictors of influenza vaccination responses. *Sci. Immunol.*, **2**, eaal4656.
Hoerl,A.E. and Kennard,R.W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.
Horn,R.A. *et al.* (1994) *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, England.
Jin,S. *et al.* (2020) scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol.*, **21**, 1–19.
Khanam,A. *et al.* (2017) Blockade of neutrophil's chemokine receptors CXCR1/2 abrogate liver damage in acute-on-chronic liver failure. *Front. Immunol.*, **8**, 464.
Kolaczkowska,E. *et al.* (2009) Neutrophil elastase activity compensates for a genetic lack of matrix metalloproteinase-9 (MMP-9) in leukocyte infiltration in a model of experimental peritonitis. *J. Leukocyte Biol.*, **85**, 374–381.
Lähdesmäki,H. *et al.* (2005) In silico microdissection of microarray data from heterogeneous cell populations. *BMC Bioinformatics*, **6**, 54.
Law,C.W. *et al.* (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.
Le,P.-H. *et al.* (2017) Clinical predictors for neutrophil-to-lymphocyte ratio changes in patients with chronic hepatitis B receiving peginterferon treatment. *In Vivo*, **31**, 723–729.
Lee,A.J. *et al.* (2018) 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*, **93**, 597–610.
Li,Y. *et al.* (2018) Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. *Stat. Interface*, **11**, 721–737.
Li,Y. *et al.* (2021) Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. *Brief. Bioinf.*, **22**, bbab024.
Liu,Y. *et al.* (2017) Super-delta: a new differential gene expression analysis procedure with robust data normalization. *BMC Bioinformatics*, **18**, 582.
Maldonado,Y.M. (2009) Mixed models, posterior means and penalized least-squares. *Lect. Notes Monograph Ser.*, **57**, 216–236.
Maronna,R.A. and Yohai,V.J. (1995) The behavior of the Stahel–Donoho robust multivariate estimator. *J. Am. Stat. Assoc.*, **90**, 330–341.
Maronna,R.A. and Zamar,R.H. (2002) Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, **44**, 307–317.
McCall,M.N. *et al.* (2021) A systems genomics approach uncovers molecular associates of RSV severity. *PLoS Comput. Biol.*, **17**, e1009617.
Miyagi,E. *et al.* (2009) Vpu enhances HIV-1 virus release in the absence of Bst-2 cell surface down-modulation and intracellular depletion. *Proc. Natl. Acad. Sci. USA*, **106**, 2868–2873.
Mohammadi,S. *et al.* (2017) A critical survey of deconvolution methods for separating cell types in complex tissues. *Proc. IEEE*, **105**, 340–366.
Newman,A.M. *et al.* (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, **12**, 453–457.
Noecker,C. *et al.* (2016) Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *MSystems*, **1**, e00013–e00015.
Obermoser,G. *et al.* (2013) Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. *Immunity*, **38**, 831–844.
Pei,R. *et al.* (2014) Interferon-induced proteins with tetratricopeptide repeats 1 and 2 are cellular factors that limit hepatitis B virus replication. *J. Innate Immun.*, **6**, 182–191.
Peng,C. *et al.* (2020) A latent unknown clustering integrating multi-omics data (LUCID) with phenotypic traits. *Bioinformatics*, **36**, 842–850.
Picelli,S. *et al.* (2014) Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.*, **9**, 171–181.
Pinu,F.R. *et al.* (2019) Systems biology and multi-omics integration: viewpoints from the metabolomics research community. *Metabolites*, **9**, 76.
Qiao,W. *et al.* (2012) PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput. Biol.*, **8**, e1002838.
Qiu,X. *et al.* (2005) Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article34.
Qiu,X. *et al.* (2013) The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinformatics*, **14**, 124.
Qiu,X. *et al.* (2014) Evaluation of bias-variance trade-off for commonly used post-summarizing normalization procedures in large-scale gene expression studies. *PLoS One*, **9**, e99380.
Quon,G. and Morris,Q. (2009) ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics*, **25**, 2882–2889.
Quon,G. *et al.* (2013) Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med.*, **5**, 29.
Rapaport,F. *et al.* (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, R95.
Repsilber,D. *et al.* (2010) Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. *BMC Bioinformatics*, **11**, 1–15.
Ritchie,M.E. *et al.* (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
Robinson,G.K. (1991) That BLUP is a good thing: the estimation of random effects. *Stat. Sci.*, **6**, 15–32.
Rousseeuw,P.J. and Driessen,K.V. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.
Sarojini,S. *et al.* (2011) Interferon-induced tetherin restricts vesicular stomatitis virus release in neurons. *DNA Cell Biol.*, **30**, 965–974.
Shannon,C.P. *et al.* (2020) Multi-omic data integration allows baseline immune signatures to predict hepatitis B vaccine response in a small cohort. *Front. Immunol.*, **11**, 578801.
Shen-Orr,S.S. *et al.* (2010) Cell type–specific gene expression differences in complex tissues. *Nat. Methods*, **7**, 287–289.
Singh,A. *et al.* (2019) DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, **35**, 3055–3062.
Smyth,G. *et al.* (2005) Limma: linear models for microarray data. In: Gentleman,R. *et al.* (eds.) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, pp. 12837–12842.
Tang,B.M. *et al.* (2019) Neutrophils-related host factors associated with severe disease and fatality in patients with influenza infection. *Nat. Commun.*, **10**, 1–13.
Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodological)*, **58**, 267–288.
Tomic,A. *et al.* (2019) SIMON, an automated machine learning system, reveals immune signatures of influenza vaccine responses. *J. Immunol.*, **203**, 749–759.
Venet,D. *et al.* (2001) Separation of samples into their constituents using gene expression data. *Bioinformatics*, **17**, S279–S287.
Wu,D. and Smyth,G.K. (2012) Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.*, **40**, e133.
Xu,R. *et al.* (2016) Low expression of CXCR1/2 on neutrophils predicts poor survival in patients with hepatitis B virus-related acute-on-chronic liver failure. *Sci. Rep.*, **6**, 38714–38719.
Yu,G. and He,Q.-Y. (2016) ReactomePA: an R/bioconductor package for reactome pathway analysis and visualization. *Mol. BioSyst.*, **12**, 477–479.
Zhang,S. *et al.* (2020) MatchMixeR: a cross-platform normalization method for gene expression data integration. *Bioinformatics*, **36**, 2486–2491.
Zhang,Y. *et al.* (2017) FUNNEL-GSEA: FUNctioNal ELastic-net regression in time-course gene set enrichment analysis. *Bioinformatics*, **33**, 1944–1952.
Zhang,Y. *et al.* (2019a) Highly efficient hypothesis testing methods for regression-type tests with correlated observations and heterogeneous variance structure. *BMC Bioinformatics*, **20**, 185.
Zhang,Y. *et al.* (2019b) The effect of tissue composition on gene co-expression. *Brief. Bioinf, 22, 127–139.*
Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67**, 301–320.