Fall 10-1-2021

# Integrative computational methodologies on single cell datasets

Wenxuan Deng
*Yale University Graduate School of Arts and Sciences*, wenxuan.deng@yale.edu

# Abstract

Integrative Computational Methodologies on Single Cell Datasets

Wenxuan Deng

2021

High throughput single cell sequencing has seen exciting developments in recent years. With its high resolution characterization of genetics, genomics, proteomics, and epigenomics features, single cell data offer more insights on the underlying biological processes than those from bulk sequencing data. The most well developed single cell technologies are single cell RNA-seq (scRNA-seq) on transcriptomics and flow cytometry on proteomics. Many multi-omics single cell sequencing platforms have also emerged recently, such as CITE-seq, which profiles both epitope and transcriptome simultaneously. But some well known limitations of single cell data, such as batch variations, shallow sequencing depth, and sparsity also present many challenges. Many computational approaches built on machine learning and deep learning methods have been proposed to address these challenges. In this dissertation, I present three computational methods for joint analysis of single cell sequencing data either by multi-omics integration or joint analysis of multiple datasets.

In the first chapter, we focus on single cell proteomics data, specifically, the antibody profiling of CITE-seq and cytometry by time of flight (CyTOF) applied to single cells to measure surface marker abundance. Although CyTOF has high accuracy and was introduced earlier than scRNA-seq, there is a lack of computational methods on cell type classification and annotations for these data. We propose a novel automated cell type annotation tool by incorporating CITE-seq data from the same tissue, publicly available annotated scRNA-seq data, and prior knowledge of surface markers in the literature. Our new method, called automated single cell proteomics data annotation approach (ProtAnno), is based on non-negative matrix factorization. We demonstrate the annotation accuracy and robustness of ProtAnno through extensive applications, especially for

peripheral blood mononuclear cells (PBMC).

The second chapter introduces an integrative method improving bulk sequencing data decomposition into cell type proportions by harmonizing scRNA-seq data across multiple tissues or multiple studies. As a Bayesian model, our method, called tranSig, is able to construct a more reliable signature matrix for decomposition by borrowing information from other tissues and/or studies. Our method can be considered an add-on step in cell type decomposition. Our method can better derive signature gene matrix and better characterize the biological heterogeneity from bulk sequencing datasets.

Finally, in the last chapter, we propose a method to jointly analyze scRNA-seq data with summary statistics from genome wide association studies (GWAS). Our method generates a set of SNP (single nucelotide polymorphism)-level weight scores for each cell type or tissue type using scRNA-seq atlas. These scores are combined with risk allele effect sizes to decompose polygenic risk score (PRS) into cell types or tissue types. We show through enrichment analysis and phenome-wide association study (PheWAS) that the decomposed PRSs can better explain the biological mechanisms of genetic effects on complex traits mediated through transcription regulation and the differences across cell types and tissues.

# Integrative Computational Methodologies on Single Cell Datasets

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Wenxuan Deng

Dissertation Director: Hongyu Zhao

December 2021

2021

## Acknowledgments

I want to first express my sincere gratitude to my advisor, Prof. Hongyu Zhao. Over the past four years since I joined Zhao lab, Prof. Zhao always offers tremendous support. As a mentor, he guides the students to choose the research directions independently, trains us to do research critically, and teaches us how to collaborate efficiently. Whenever we met difficulties, he would always respond very quickly and be available for a discussion, although he is much busier than us most of the time indeed. His mentorship and research training enabled me to think critically and scientifically. My research is greatly influenced by his taste and biological scientific way of thinking. Moreover, Prof. Zhao is the most hard-working and humble person I have met. It is my great fortune to be his student and have him as my role model.

I would like to thank my dissertation committee member, Prof. Naftali Kaminski, for his helps through the past few years. Prof. Kaminski provided the rich data resources for me to play with after he knew I was interested in single cell methodologies. As an expert in pulmonary disease, he regularly scheduled committee meetings with me on my research progress and always proposed suggestions from a clinician perspective, which made great help. I would also like to thank my committee member, Prof. Leying Guan. Prof. Guan offered her help since we were both at Tsinghua University. She graduated one year earlier and gave me many suggestions on research and academia when I was a junior undergraduate. After she joined the department of biostatistics at Yale, she became my committee member and always gave me quick and helpful feedback on my dissertation. I'm grateful to have Prof. Kaminski and Prof. Guan as my committee members. I also want to thank Prof. Shuangge Steven Ma, Prof. Forrest Crawford, Prof. Joshua Warren and Prof. Heping Zhang from the Department of Biostatistics, for being great course instructors during my first two years as a graduate student. I want to thank Prof. Maria Ciarleglio for being my academic advisor before I joined Zhao lab. She is the first faculty I met at Yale and helped me go over the courses and academic plan in the first year. I must thank Melanie Elliot, who is the best administrative director one can ever imagine.

I would like to extend my thanks to other great scholars I have been working with. I would like to thank Dr. Avraham (Rami) Unterman, Dr. Maor Sauler, Taylor Adams, and Prof. Xiting Yan from Kaminski lab. I worked with Rami and Taylor when I first got exposed to single cell data. They were experienced in transcriptomics data and gave me hands-on help from data quality control to downstream data analysis. Maor is a pulmonary doctor in ICU and had a busy schedule, especially during the pandemic. But he still managed to take time to discuss with us with his expert knowledge in COPD. I am very thankful to Prof. Yan for giving me many insightful advices when we had a bottleneck in research. Moreover, I have gained considerable help from Prof. Tomokazu Sumida. We had a pleasant collaboration on covid-19 research project and he kindly shared their data permission to me for methodology development, which is finally a significant part of one of my projects.

The past five years have been very joyful. It is also my privilege to work closely with my labmates in Zhao lab. I want first to thank Biqing Zhu, Bolun Li, and Yixuan Ye for my dissertation. Biqing helped done the laborious work on simulations and real applications when I was developing ProtAnno. She always gave me timely feedback and quick results. On tranSig method, I worked closely with Bolun. He is a talented and reliable collaborator, and it was always efficient to discuss with him. Yixuan helped me formulate the PRS project, and I benefited a lot when working with her. Besides, I am thankful to Wei Liu and Ming Chen, who collaborated with me through the past two years. It is an excellent opportunity to work with them and explore the possibility of integrating GWAS and single cell analysis with them together. I also want to thank Daiwei Tang, Jiawei Wang, Jerome Yu, Yiliang Zhang, Chi Zhang, Dr. Wei Jiang, Dr. Ying Zhu, Dr. Dingjue Ji, Dr. Xiaochen Wang, and Dr. Xinyue Li, who often had valuable discussions with me on research through the past few years. I want to thank Dr. Ning Sun, who is just like our parents in lab. She always gave us suggestions on research and provided gentle care during the pandemic when we were far from our family.

Next, I would also like to thank all my friends I have met at Yale. They are the best presents in

New Haven through these few years. Firstly, I want to thank Ming Chen, Wei Liu, and Yixuan Ye again for being my best friends in New Haven in the lab and daily life. We have unlimited topics to talk about and laugh about. It is hard to feel lonely when having them. I would also like to thank Shuwen Deng and Chang Liu, another two best friends in Yale. They are always the first ones to help me whenever I meet difficulties. I always feel comfortable and relieved to be with them when I am stressed.

Finally but most importantly, I want to thank my parents and my fiance for their unconditional support. My parents gave me sufficient freedom to make my own decisions even they were extremely anxious about me when I first left my home country and lived independently overseas. I clearly knew how hard they were at that time. But they never stopped believing in me and giving me courage when I doubted myself. And I will never survive my graduate study without Zhiwei's support. His companionship and patience encouraged me to keep going when I hesitated. We have lived in two different countries for more than five years, but the distance is never a problem. I own them more than I can say.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# ProtAnno, an Automated Cell Type Annotation Tool for Single Cell Proteomics Data that integrates information from Multiple Reference Sources

## Abstract

Compared with sequencing-based global genomic profiling, cytometry labels targeted surface markers on millions of cells in parallel either by conjugated rare earth metal particles or Unique Molecular Identifier (UMI) barcodes. Correct annotation of these cells to specific cell types is a key step in the analysis of these data. However, there is no computational tool that automatically annotates single cell proteomics data for cell type inference. In this chapter, we propose an automated single cell **prot**eomics data **anno**tation approach called **ProtAnno** to facilitate cell type assignments without laborious manual gating. ProtAnno is designed to incorporate information from annotated single cell RNA-seq (scRNA-seq), CITE-seq, and prior data knowledge (which can be imprecise) on biomarkers for different cell types. We have performed extensive simulations to demonstrate the accuracy and robustness of ProtAnno. For several single cell proteomics datasets that have been manually labeled, ProtAnno was able to correctly label most single cells. In summary, ProtAnno offers an accurate and robust tool to automate cell type annotations for large single cell proteomics datasets, and the analysis of such annotated cell types can offer valuable biological insights.

*Keywords*: CyTOF; scRNA-seq; CITE-seq; Gating; Data Integration.

## 1.1 Introduction

Recent years have seen the developments of many single cell platforms [1] that have enabled researchers to collect high throughput -omics profiles at the individual cell level, including genomics [2][3], transcriptomics [4], proteomics [5], and [6]. These data can reveal biological heterogeneity across different biological conditions. They offer a direct approach to studying cell type compositions and functional cell states. The most well-developed platforms are scRNA-seq for transcriptomics and flow cytometry (CyTOF) [7][8][9] for proteomics. Notably, the rise of scRNA-seq has generated rich data and resources on single cell transcriptomics [10][11].

In addition to collecting single cell data for one specific -oimcs data type, it is possible to collect multi-omics data simultaneously at the single cell level, e.g., CITE-seq [12] that measures mRNA and antibody counts simultaneously by UMI barcode. These data can characterize the cellular relationship between transcript and cell surface marker abundance. Methods have been developed to bridge these two types of omics data. For example, cTP-net [13] is a transfer learning approach under a deep learning framework to predict surface protein levels from scRNA-seq data. The generation of diverse types of single cell data poses many computational challenges due to their high dimensionality and large sample sizes. In this chapter, we focus on the annotation of single cell proteomics data to their corresponding cell types, which is a critical step in single cell analysis.

A major advantage of single cell proteomics data compared with scRNA-seq data is their high sensitivity and specificity. Based on surface marker expression patterns, manual gating can be used to identify different cell populations by the expression distributions. However, cell type labeling by manual gating on single cell data is labor intensive and subjective as it highly depends on the expert who annotates the cells. Although some methods can analyze these data based on state-of-the-art machine learning methods [14][15][16] for clustering cells, most of these methods are unsupervised and unable to identify cell populations automatically.

For scRNA-seq data, a number of methods have been developed to assign cells to different cell types based on expression profiles. For example, SingleR [17] trains on an extensive collection of large annotated reference transcriptomics single cell data. Preliminary labeling by SingleR can vastly accelerate cell type inference. However, there is no similar automated tool for single cell proteomics data annotation. A recently developed tool named CellGrid [18] applied the same idea to CyTOF data but needed a large number of labeled proteomics datasets as input. However, there is a lack of labeled single cell proteomics data although well-annotated scRNA-seq data are more broadly available. For the CITE-seq data, it is still necessary to annotate the transcriptomics and proteomics data separately.

To overcome the lack of labeled single cell proteomics data, we introduce an automated single cell **prot**eomics data **anno**tation approach called ProtAnno, based on non-negative matrix factorization (NMF) to incorporate data from different reference sources. The only essential input of ProtAnno is some prior knowledge on cell type-specific biomarkers. To further improve annotation accuracy, ProtAnno can take advantage of publicly available CITE-seq data and annotated scRNA-seq data. This enables ProtAnno to perform cell type annotation with no prior characterization between cell types and surface proteins by leveraging these external references.

We have evaluated the performance of ProtAnno through simulations under different settings. The results showed the robustness of ProtAnno to biological variability, technical noise, cell type number, and incomplete and inaccurate expert knowledge. We then applied ProtAnno to three real datasets: peripheral blood mononuclear cell (PBMC) paired stimulated B cell receptor CyTOF data, PBMC CITE-seq data from healthy subjects, and longitudinal whole blood covid-19 CyTOF data grouped by patients' disease severity. In the analyses of these real data, ProtAnno provided fast and accurate labeling as demonstrated through comparisons with manual annotations or downstream biological investigations. In summary, ProtAnno is a computationally efficient and statistically robust approach for automated cell type annotation when only limited expert knowledge is available for single cell proteomics data.

## 1.2 Methods

### 1.2.1 Automated Single cell Proteomics Data Annotation Model

The workflow is sketched in Fig 1.1. ProtAnno deconvolutes the proteomic expression profile $X \in R^{D\times C}$ into the product of the cell type-specific signature matrix, $W \in R^{D\times K}$, and cell type assignment matrix, $H \in R^{K\times C}$, i.e. $X =$ *WH*. In the model, we have $D$ surface markers for $C$ cells in the proteomics data, e.g., the cytometry data and antibody profile in CITE-seq. We denote $K$ as the number of cell types. The columns of $W$ are matched with the known cell types in the same order. Due to the non-negative requirement on the estimations of $W$ and $H$, ProtAnno implemented NMF for solving $X = WH$.

In ProtAnno, we integrate information from both prior knowledge encoded in a matrix $A_0$ and relationship between protein markers and RNA-seq data encoded in a matrix $A$. In our model, we denote $A \in R^{D\times G}$ as the protein-RNA association matrix inferred from a CITE-seq data by elastic net [39], where $G$ is the number of genes considered. A desirable CITE-seq data should include a large number of measured antibodies. ProtAnno implements two internal dictionary-like CITE-seq datasets from [19] and [20]. Both datasets have more than $180$ antibody tags. The transcriptome signature matrix $S \in R^{G\times K}$ is generated by well-annotated scRNA-seq data. The $K$ columns of $S$ should be matched with $W$. We recommend using the denoised scRNA-seq data to impute drop-out events for better annotation. Specifically, we deploy SAVERx [21] for imputation due to its superior performance.

The expert knowledge matrix $A_0 \in R^{D \times K}$ is designed based on known cell type surface markers. $A_0$ is a discrete matrix containing three possible values, $+1$, $-1$, and $0$. If biomarker $i$ should have high expression level in cell type $j$, we set $A_0^{ij} = 1$; if the biomarker is not expressed in this cell type, $A_0^{ij} = -1$; and if there is no constraint on the biomarker and cell type, then we set $A_0^{ij} = 0$.

To estimate the signature matrix $W$ in the order of given cell type list, ProtAnno adds constraint on $W$ with respect to the above two proteomic signature matrices, $AS$ and $A_0$.

With all the notations introduced above, the ProtAnno model is formulated as follows:

$$\begin{aligned} \min_{W \geq 0, H \geq 0} L(W, H) = \frac{1}{2} \parallel X - WH \parallel_F^2 - \lambda_1 tr(W^T AS) - \lambda_2 tr(W^T A_0) \\ + \frac{\mu}{2} \parallel W \parallel_F^2 + \frac{\eta}{2} \parallel 1_k^T H - 1_N^T \parallel_2^2 \end{aligned} \tag{1.1}$$

Note that ProtAnno adds regularization penalty on $W$ and $H$ to improve performance. We use $1_K$ and $1_N$ to denote the $K$-dimensional and $N$-dimensional column vectors with all-ones. The fourth term in ProtAnno is to control the scale of $W$, and the last term is to force the column sum of $H$ to be 1.

We optimize this overall objective function based on the multiplicative update algorithm [22][23] to guarantee non-negativity. The algorithm requires the specifications of the penalty parameters, $\lambda_1, \lambda_2, \mu$, and $\eta$. In each iteration, ProtAnno updates $W$ by rows and $H$ by columns. The details of

ProtAnno are provided in Algorithm 1.

```
Algorithm 1: ProtAnno Optimization Algorithm
  Result: Optimal non-negative $W$ and $H$
  Model input: Normalized expression profile matrix $X$, discrete prior knowledge matrix
    $A_0$, public labeled single cell expression profile, and penalty parameters $\lambda_1, \lambda_2, \mu, \eta$;
  Step 1: set $t = 0$ and generate intial non-negative $W$ and $H$ with all ones;
  Step 2: Update $W$ by rows and $H$ by column:
    $$\begin{aligned} w_{ij}^{t+1} &= w_{ij}^{t} \frac{[xH^T]_j^+ + \lambda_1[(AS)_i]_j^+ + \lambda_2[(A_0)_i]_j^+}{[w_i(HH^T + \mu I)]_j + [xH^T]_j^- + \lambda_1[(AS)_i]_j^- + \lambda_2[(A_0)_i]_j^-} \\ h_j^{t+1} &= h_j^{t} \frac{[W^T x]_j^+ + \eta 1_k}{[(W^T W + \eta 1_k 1_k^T)h^t]_j + [W^T x]_j^-} \end{aligned} \qquad (1.2)$$
  Step 3: Update $t = t + 1$ until the convergence or reaching the given iteration times.
```

It can be shown that the algorithm converges in theory with the details of the theorems and converging rates provided in the supplementary materials B.2.

### 1.2.2 Choices of Penalty Parameters

Since the penalty parameter values are critical to ProtAnno, we developed the following algorithm to tune their values iteratively.

To ensure that ProtAnno can group expression profiles into reasonable clusters, we use the default unsupervised Louvain algorithm for preliminary clustering. We then set the initial parameter$\eta$ by the KKT condition; and then search the initial $\lambda_1$, $\lambda_2$, and $\mu$ values by choosing from among 0.1, 1, 10, and 100. To find the initial $\lambda_1$ and $\lambda_2$ values, we maximize the ARI between the Louvain clusterings and ProtAnno results. We consider a novel index to select the initial value for $\mu$:

$$D(\mu) := X_{mean} - W_{mean} \tag{1.3}$$

The above function considers the difference between the mean values of $X$ and $W$. The main purpose of $D(\mu)$ is to ensure that $W$ is on approximately the same scale as $X$. Otherwise, the deconvolution would be unstable since we penalize the column summation of H to make the cell type proportion estimations meaningful.

To set more precise penalty terms, ProtAnno uses binary search to determine the final optimal output as detailed in Algorithm 2, and allows the specified search depth depending on the running time. If the final ARI is lower than $0.75$, the algorithm will restart the binary search at higher resolution.

```
Algorithm 2: ProtAnno Parameter Automated Tuning Algorithm
  Result: Optimal penalty parameters: $\lambda_1, \lambda_2, \mu, \eta$
  Model input: Normalized expression profile matrix $X$, discrete prior knowledge matrix
    $A_0$, public labeled single cell expression profile;
  Step 1: Estimate Louvain cluster;
  Step 2: Initialize $W$ and $H$ by an arbitrary optimization, and initialize $\eta$ by KKT
    conditions:
        $\eta :=\| (W^T WH - W^T X)/(1_K 1_K^T H - 1_K 1_N^T) \|_{\text{median}}$    (1.4)
  Step 3:Initialized $\lambda_1$ and $\lambda_2$ by choosing from $(0.1, 1, 10, 100)$ and minimizing Adjusted
    Rank Index (ARI) with Louvain clustering;
  Step 4:Initialized $\mu$ by estimated signature matrix W reliability by a developed metric:
        $D(\mu) := X_{\text{mean}} - W_{\text{mean}}$    (1.5)
  Step 5: Select the best $\lambda_1$, $\lambda_2$ and $\mu$ by binary search;
  Step 6 (Optional):If the final ARI with Louvain cluster is still lower than 0.75, we re-do
    the binary search on all the penalty parameters.
```

## 1.3 Results

### 1.3.1 Simulation Setup

In the following, we describe how we simulated the protein expression profiles for each single cell. We first generated the entries in the expert matrix $A_0$ from a multinomial distribution with three categories, $+1$, $-1$, and $0$, corresponding to biomarker knowledge for a cell type, i.e., high expression, no expression, and no information. The matrix will be built with the angles between vectors as large as possible, so that the simulated single cell expression profile is a nonnegative linear combination of the proteomics signature matrix [22]. To achieve it, we generate $100$ matrices randomly and minmax the inner products between column vectors to get the optimal one. This design captures the nature of signature genes. However, the prior knowledge may not be in line with true relationships between markers and cell types due to technical and biological variations. We, therefore, regenerated an intermediate discrete matrix $\tilde{A}_0$ based on $A_0$ containing three possible discrete values, 2 (high expression), $1$ (low expression), and $0$ (no expression), by random walk and the protocols observed in real data (details in Methods). We then derived the signature matrix $W$ based on $\tilde{A}_0$ from two truncated normal distributions with high and low mean values, respectively, and varying relationships between variances and means that reflect the signal noise ratio for a biomarker. The difference of the means of the two truncated normal distributions together with their variances dictate the informativeness of a specific biomarker for cell type specification. The proteomics expression profile $X$ was sampled from the truncated normal distributions based on the signature matrix that defines the mean and associated variance for each biomarker in each cell type. Finally, to simulate the predicted proteomics signature matrix from transcriptomics data, which is the product of the protein-RNA association matrix $A$, and the transcriptomics signature matrix $S$, $[AS]$, we sampled the element $[AS]_{ij}$ in the matrix with mean $W_{ij}$, the expression level of signature gene $i$ in cell type $j$, and variance $W_{ij}/(2 * corr)$. A smaller $corr$ results in a weaker correlation between the two proteomics signature matrices. All the simulation details can be found in Materials and Methods.

### 1.3.2 Model performances and comparisons

We considered six simulation scenarios to cover a wide range of biological and technical variations in real data, with data quality varying from high (scenario $1$) to low (scenario $6$). We tuned the parameters of the normal distributions to decrease the cell type distinction and enlarge the expression profile divergence. To simulate the situations when $AS$ or $A_0$ is too noisy to have a good prediction power, we also added more uncertainty to $AS$ and $A_0$ that led to a lower correlation with real signature matrix $W$. The overall noise increased from scenario $1$ to $6$. In scenario $5$, we increased the noisy level of expert matrix $A_0$ so that it is more aberrant compared with the single cell proteomics data expression pattern. In addition, the correlation of $W$ and $AS$ was set at only $0.2$. These large noises from references would reduce the annotation accuracy. Furthermore, we increased expression profile variations by lowering the mean-variance ratio in scenario $6$, making the annotation more challenging. All the parameters settings and details are listed in Table A.1. In the columns, big_w_mean and big_tau_w are the mean and standard error of the truncated normal distribution for high expression biomarker; small_w_mean and small_tau_w are the mean and standard error of the truncated normal distribution for low expression biomarker; p.0 is the probability of random walk from 0 in $A_0$ to 2 in $A_1$; q.0 is the probability of random walk from 0 in $A_0$ to 1 in $A_1$; p.neg1 is the probability of random walk from -1 in $A_0$ to 2 in $A_1$; q.neg1 is the probability of random walk from -1 in $A_0$ to 2 in $A_1$; mean_var_ratio is signal noise ratio of a celltype-specific biomarker expression; corr is correlation of $W$ and $AS$

We compared the performance of six models: the full ProtAnno model, the ProtAnno model without transcriptomics information $AS$, the ProtAnno model without expert knowledge $A_0$, the unsupervised ProtAnno model, the unsupervised Louvain clustering, and cell type assignment by non-negative least squares (NNLS) when the protein signature matrix $W$ is known. The last one is the best result an NMF model can achieve. We consider five metrics for annotation accuracy:

Adjusted Rand Index (ARI), annotation accuracy assigned by the largest value for every single cell, Normalized Mutual Information (NMI), cosine similarity between the estimated and real assignment, and Average Silhouette Width.

The simulation results in Fig 1.2A show that the full ProtAnno model achieved the best annotation accuracy, especially compared with the NNLS model. Specifically, from scenarios $1$ to $4$, ProtAnno was able to achieve the same performance an NNLS model could have when the true signature matrix is known. These results suggest the importance of having precise information for $A_0$ and $AS$. Besides the full ProtAnno model, the partial ProtAnno model with the expert matrix had the second-best performance, just slightly worse than the full model. The partial ProtAnno model only having the transcriptomics reference was not competitive with the first two models. But it was still much better than the unsupervised clustering, even in terms of ARI and NMI. In general, both full and partial ProtAnno models could obtain accurate cell type assignments even in the presence of considerable noise.

We investigated the numerical convergence of ProtAnno in our simulation studies. Fig 1.2B shows an example of ProtAnno cell type assignment in a UMAP plot with $29$ cell types, where the accuracy was $98.8\%$. The high agreement of clustering between the real labels (left) and the ProtAnno annotation (right) demonstrates the power of ProtAnno.

### 1.3.3 Algorithm convergence and robustness

ProtAnno generally converges after $100$ iterations (Supp A.1A). To filter the high confidence annotation, the subsetting step will keep the cell whose estimated $H$ is greater than $0.5$ after column normalization. The subsetting step was able to slightly enhance the annotation accuracy (Supp A.1B).

ProtAnno's robustness was investigated with respect to three parameters: mean-variance ratio (Fig 1.3A), correlation between $W$ and $AS$ (Fig 1.3A, supp A.1C), and cell type number $K$ (Fig

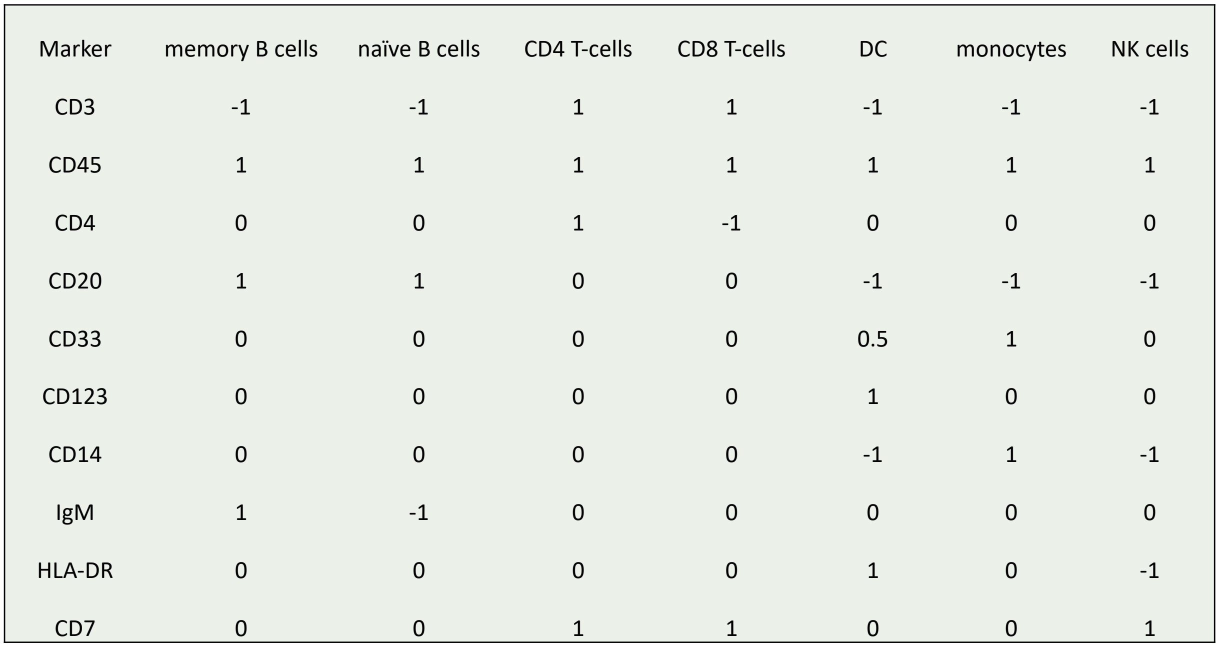
1.3C). Since a protein with a higher mean expression level tends to have a higher variance, we used mean-variance ratio to characterize the signal noise ratio of a protein. As expected, the prediction accuracy dropped with a reduced mean-variance ratio (Fig 1.3A). Overall, the full ProtAnno model was very robust even when the $AS$ prediction power was weak (supp A.1C), implying the critical role of $A_0$. We also explored how the transcriptomics information $AS$ could help the annotation when $A_0$ is not available (supp A.1C). The accuracy rate was almost linearly associated with the prediction correlation. Therefore, scRNA-seq data and dictionary CITE-seq data with high quality are essential for the good performance of the partial ProtAnno model when no reliable expert knowledge is available. As for the impact of the number of cell types, ProtAnno was robust to the large cell type number as long as the cell counts of each cell population were more than 100 in our simulation (Fig 1.2B, Fig 1.3C).

The good performance of ProtAnno depends on the appropriate choices of the penalty parameters. Simulation results suggest that our proposed parameter tuning algorithm was able to find good penalty values (Fig 1.3D, Supp A.1D) as shown in the relationship between algorithm performance and tuning parameter values in these figures in most cases.

Finally, we can rewrite the ProtAnno objective function in an equivalent form as (see supplementary materials B.3):

$$\begin{aligned}\min_{W\geq 0, H\geq 0}\frac{1}{2}\parallel X-WH\parallel_F^2+\frac{\mu}{4}\parallel W-\frac{2\lambda_1}{\mu}(AS)\parallel_F^2\\+\frac{\mu}{4}\parallel W-\frac{2\lambda_2}{\mu}A_0\parallel_F^2+\frac{\eta}{2}\parallel 1_k^T H-1_N^T\parallel_F^2.\end{aligned} \tag{1.6}$$

The above objective function form illustrates the imposed relationships between $W$, $AS$, and $A_0$. The second penalty terms will force the ratio of elements in matrices $W$ and $AS$ to be approximately equal to $\frac{2\lambda_1}{\mu}$. Specifically, the ratio of $W_{ij}/\left[AS\right]_{ij}$ should be approximately $\frac{2\lambda_1}{\mu}$ for any entry with row index $i$ and column index $j$. Similarly, the ratio of $W$ and $A_0$ should be approximately $\frac{2\lambda_2}{\mu}$. Therefore, we empirically examined whether the ratios of matched elements in



| Marker | memory B cells | naïve B cells | CD4 T-cells | CD8 T-cells | DC | monocytes | NK cells |
|---|---|---|---|---|---|---|---|
| CD3 | -1 | -1 | 1 | 1 | -1 | -1 | -1 |
| CD45 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| CD4 | 0 | 0 | 1 | -1 | 0 | 0 | 0 |
| CD20 | 1 | 1 | 0 | 0 | -1 | -1 | -1 |
| CD33 | 0 | 0 | 0 | 0 | 0.5 | 1 | 0 |
| CD123 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CD14 | 0 | 0 | 0 | 0 | -1 | 1 | -1 |
| IgM | 1 | -1 | 0 | 0 | 0 | 0 | 0 |
| HLA-DR | 0 | 0 | 0 | 0 | 1 | 0 | -1 |
| CD7 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

Table 1.1: $A_0$ for BCR application.

matrices $W$ and $AS$ (or $A_0$) approached the theoretical ones (Supp A.2A, A.2B). The figures show that the empirical distribution of these ratios compared with $\frac{2\lambda_1}{\mu}$ and $\frac{2\lambda_2}{\mu}$, suggesting that estimated $W$ was close to $\frac{2\lambda_1}{\mu}(AS)$ and $\frac{2\lambda_2}{\mu}A_0$ at the same time.

### 1.3.4 Application to paired stimulated B cell receptor/Fc receptor cross-linker (BCR) cytometry data

We first evaluated ProtAnno's performance on a benchmarking BCR CyTOF dataset with eight paired samples from PBMC in [24]. Compared with the reference group, the B cell population was stimulated by BCR in the treatment group. All the samples were collected from healthy individuals, and $10$ cell type markers were measured in a total of $172791$ cells. This dataset was labeled in [25] after the authors applied an unsupervised clustering method FlowSOM [15] and manually merged clusters. We evaluated the assignments of these cells to five major cell types based on the $10$ cell type markers. Other inputs, the association matrix and the transcriptomics signature matrix $S$, were generated by the CITE-seq data in [20]. To match with annotated cell types by Nowicka et al., we manually gated the five major cell types in the scRNA-seq. We also constructed an expert-guided matrix $A_0$ (Table 1.1). ProtAnno assigned a cell to the cell type with the largest value in matrix $H$.

The overall median accuracy of ProtAnno for this data set was $83\%$ (Fig 1.4A) across the samples. For most of the samples, the accuracy could be as high as $80\%$ to $85\%$ (Supp3), except an outlier, patient 8, with decreased the average. For this data set, even the partial ProtAnno with only $A_0$ or only $AS$ could achieve an accuracy of over $70\%$ overall. The subsetting step could slightly improve accuracy by keeping high-confidence cells (Supp A.3, Supp A.4). Because the BCR dataset does not have clear clustering patterns, making it challenging to annotate cells. This can be seen in the UMAP plot, where some of the cells from different cell types are located in the same region (Fig 1.4E). Similar observation was made in the benchmarking study of unsupervised clustering accuracy by [26]. At the lowest clustering resolution of 9 clusters, the false positive rate was around $20\%$, where unsupervised clustering could not achieve accurate annotation. Therefore, an overall accuracy of $80\%$ may be considered satisfactory for this dataset. We further investigated the misclassification patterns by the confusion matrix for patient 1 in the stimulated group (Fig 1.4B). It can be seen that most of the cells could be correctly assigned by ProtAnno. However, some of the natural killer cells (NK cells) were annotated as CD8 T cells. That is partly because the BCR data did not measure the biomarker CD8, which offers important information for CD8 T cell assignment. For this data set, the only biomarker that was informative to distinguish these two cell populations is CD3, which makes computational annotation difficult. The UMAP plot also shows that the annotation is highly consistent with the manually annotated cell types, although the boundary of NK cells and CD8 T cells is hard to recognize (Fig 1.4E).

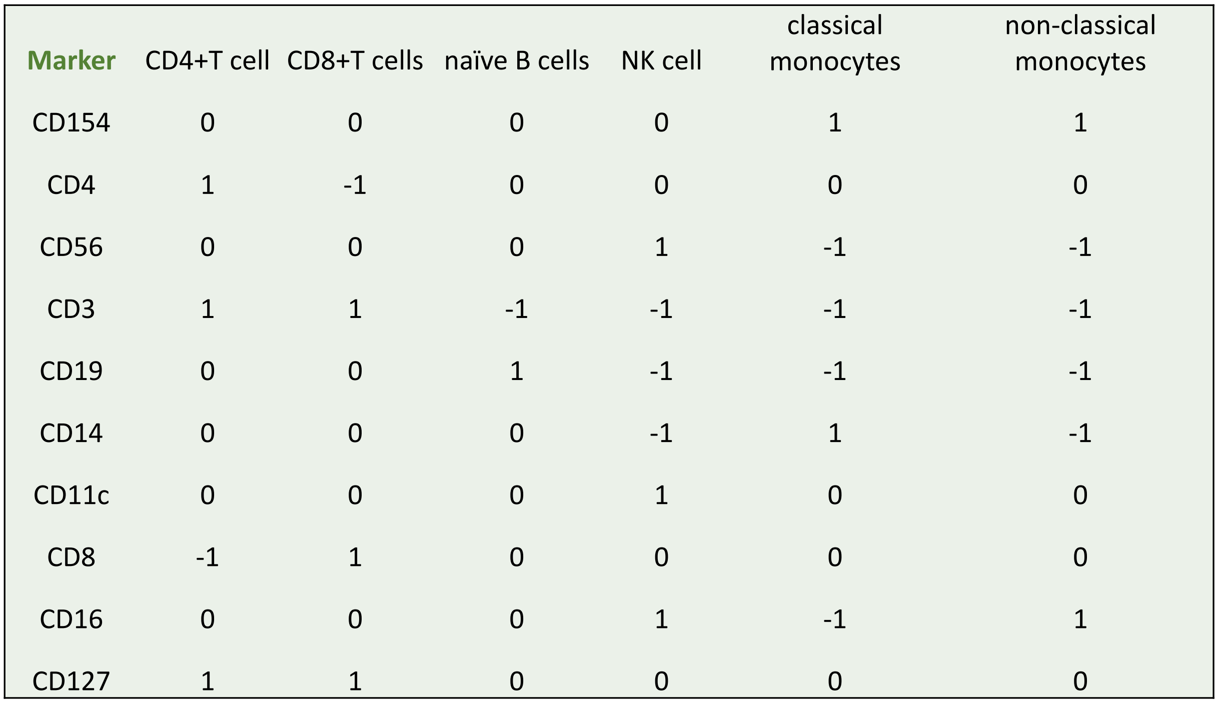
Since many genes had significant differential expression levels across groups in the BCR dataset, we studied the W matrix variations to test whether ProtAnno could accurately estimate the signature matrix. The most differential expression directions in the stimulated group by ProtAnno are consistent with prior biological knowledge (Fig 1.4C). For example, biomarker HLA-DR had a significantly elevated expression in most cell types after BCR stimulation. The inferred cell type proportions also suggest that ProtAnno can provide an accurate estimation in the comparative study. For example, the proportions of naive B cells were lower in the stimulated group (Fig 1.4D). These downstream analyses suggest accurate annotation by ProtAnno.

### 1.3.5 Application to PBMC CITE-seq data

To investigate the performance of ProtAnno on the proteomic expression profile in CITE-seq data, we analyzed a PBMC CITE-seq dataset reported in [27]. This dataset measured $10$ surface markers in $1372$ cells from a healthy subject. We used ProtAnno to annotate $1153$ cells with six cell types, because we cannot annotate the other $219$ cells to any cell type based on the available limited $10$ markers. These cells were manually labeled and the details are provided in the online methods and materials 1.5.4. We used the CITE-seq data in [19] to derive the transcriptomics signature matrix and the CITE-seq data in [20] to obtain the association matrix and construct $A_0$(Table 1.4). The automated annotation accuracy was around $99\%$ with $100\%$ accuracy for most cell types, except for non-classical monocytes. ProtAnno failed to recognize this cell type since it is a rare population whose cell type proportion is below $1\%$ (Fig 1.5A). These results suggest that ProtAnno could be applied to multiple single cell proteomics sequencing platforms, though cytometry and CITE-seq are different technologies with distinct features.

### 1.3.6 Application to longitudinal whole blood covid-19 cytometry data

Finally, we applied ProtAnno to a longitudinal whole blood covid-19 CyTOF data set reported in [28]. This dataset included $37$ adult covid-19 patients classified into ICU, non-ICU, and recovery groups. All the hospitalized (ICU and non-ICU) patients were measured at multiple time points. Each sample had $15,000 \sim 20,000$ cells. The authors used an in-house annotation method Cell-Grid [18] to train a subtyping strategy with high resolution to annotate over $50$ cell types. Since we do not have access to such a large training CyTOF dataset and comprehensive knowledge on

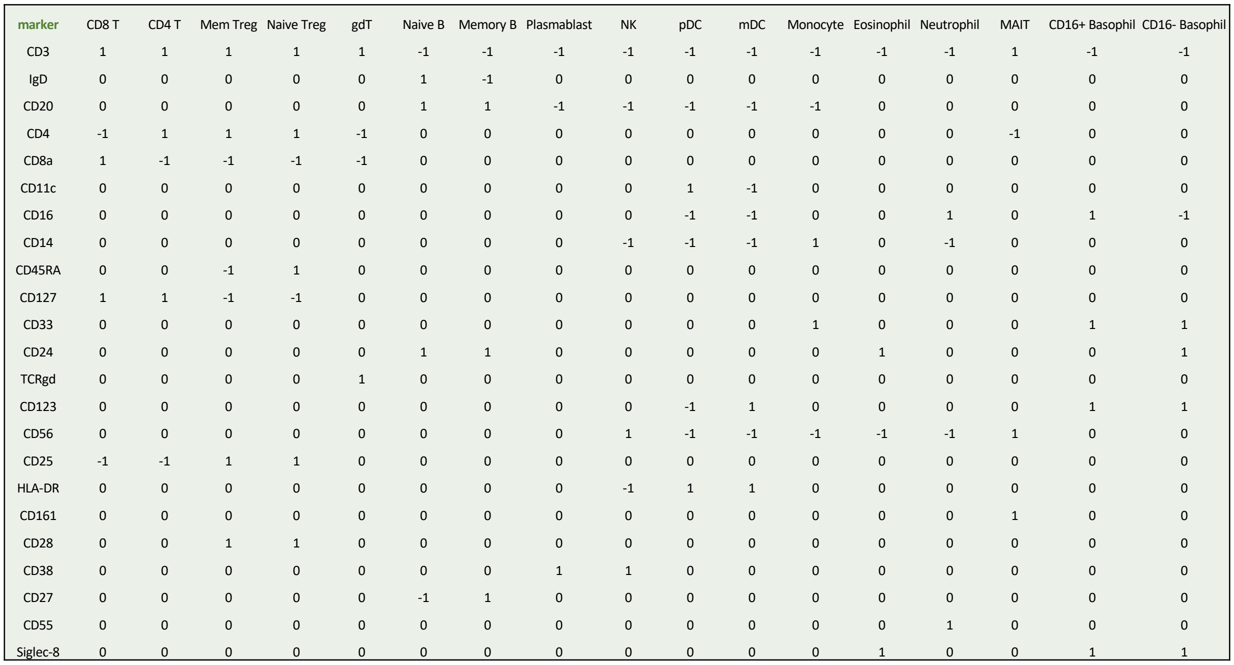


| Marker | CD4+T cell | CD8+T cells | naïve B cells | NK cell | classical monocytes | non-classical monocytes |
|---|---|---|---|---|---|---|
| CD154 | 0 | 0 | 0 | 0 | 1 | 1 |
| CD4 | 1 | -1 | 0 | 0 | 0 | 0 |
| CD56 | 0 | 0 | 0 | 1 | -1 | -1 |
| CD3 | 1 | 1 | -1 | -1 | -1 | -1 |
| CD19 | 0 | 0 | 1 | -1 | -1 | -1 |
| CD14 | 0 | 0 | 0 | -1 | 1 | -1 |
| CD11c | 0 | 0 | 0 | 1 | 0 | 0 |
| CD8 | -1 | 1 | 0 | 0 | 0 | 0 |
| CD16 | 0 | 0 | 0 | 1 | -1 | 1 |
| CD127 | 1 | 1 | 0 | 0 | 0 | 0 |

Table 1.2: $A_0$ for PBMC CITE-seq application.

biomarkers, we selected 23 surface markers to define 17 common cell types in whole blood. For this dataset, we only considered the partial ProtAnno model with $A_0$ since there is no whole blood CITE-seq dataset available (Table **??**). Since we do not have ground truth on cell type annotations, we primarily evaluated the performance of ProtAnno through downstream analysis. More specifically, we studied the cell counts and cell types of neutrophils and lymphocytes.

The neutrophiles are strongly positively correlated with patient disease severity [29]. In this dataset, we treated the patient group as the severity indicator. In the ProtAnno results, the neutrophil cell counts of recovery patients were significantly lower than those of the hospitalized patients (Fig 1.5B, 1.5C) with the Wilcoxon test p-value smaller than 0.05 for both raw output and subsetting output of ProtAnno, consistent with our expectation. It is also known that the inflammatory symptoms are dramatically elevated in severe cases [30][31][32][33]. Longitudinal analysis of this dataset showed a slight decrease of neutrophils proportions over time after admission, except with one patient, COV-34, having a slight increase at the end of the study (Fig 1.5D, 1.5E). The



| marker | CD8 T | CD4 T | Mem Treg | Naive Treg | gdT | Naive B | Memory B | Plasmablast | NK | pDC | mDC | Monocyte | Eosinophil | Neutrophil | MAIT | CD16+ Basophil | CD16- Basophil |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CD3 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 1 | -1 | -1 |
| IgD | 0 | 0 | 0 | 0 | 0 | 1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CD20 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | -1 | -1 | -1 | -1 | -1 | 0 | 0 | 0 | 0 | 0 |
| CD4 | -1 | 1 | 1 | 1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 | 0 |
| CD8a | 1 | -1 | -1 | -1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CD11c | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 |
| CD16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | 0 | 0 | 1 | 0 | 1 | -1 |
| CD14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 |
| CD45RA | 0 | 0 | -1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CD127 | 1 | 1 | -1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CD33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| CD24 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| TCRgd | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CD123 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| CD56 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | -1 | -1 | -1 | -1 | -1 | 1 | 0 | 0 |
| CD25 | -1 | -1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| HLA-DR | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| CD161 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CD28 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CD38 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CD27 | 0 | 0 | 0 | 0 | 0 | -1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CD55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Siglec-8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |

Table 1.3: $A_0$ for whole blood covid-19 cytometry application.

changes between day 1 to day 12 were significant, especially for the ICU group with a p-value of $0.029$. These results suggest that most ICU patients were recovered after treatment.

In contrast, the lymphocyte proportion is expected to decline in severe covid-19 patients. Lymphopenia has been used as a biomarker to define a patient's morbidity [34][35][36]. When we applied ProtAnno to infer CD4 T and CD8 T cells, although there was no statistically significant difference across groups, the CD4 T cell count was indeed elevated if the patients recovered (Fig 1.5F). In the longitudinal curves of cell counts, almost all the patients had increased CD4 T cell and CD8 T cell counts after treatment (Fig 1.5G). This implies that the lymphopenia was alleviated by treatment.

The analysis of this data set shows the usefulness of ProtAnno even when the data are highly noisy.

## 1.4 Discussion

In this chapter, we have developed a computationally efficient and statistically robust method, ProtAnno, for automated annotation for single cell proteomic data. Using protein expression profiles, this computationally efficient method annotates cells based on cell type information gathered through (imprecise) prior knowledge, publicly annotated single cell datasets, and CITE-seq data. ProtAnno only requires simple and accessible references, e.g. publicly annotated scRNA-seq. When only limited references are available, ProtAnno can be applied either without prior biological knowledge or public transcriptomics references. ProtAnno can also resolve heterogeneity among cell populations. For instance, ProtAnno was able to detect signature matrix variations across different sample groups in the BCR study. In real data applications, we showed that we can gain biological insights through downstream analysis, such as cell type-specific differential expression analysis and cell type proportion comparisons after annotations by ProtAnno.

Simulation results showed that ProtAnno is robust to the number of cell types and noises in cytometry and CITE-seq antibody expression profiles. However, ProtAnno may be limited to major cell types and unstable when the data are very noisy. Specifically, it is common that the single cell data does not have a clear separated clustering. For example, ProtAnno does not always have a reliable assignment for all samples (supp A.3). When a specific cell type is hard to recognize even for manual gating, ProtAnno may misclassify its entire cell subpopulation. The annotation performance also dropped substantially in simulations with large noise in X and/or more limited information from references $A_0$ and $AS$. Therefore, ProtAnno may be suitable for preliminary assignments and may require manual inspection.

Future work can incorporate hierarchical cell type structure to enhance the classification resolution and accuracy. Specifically, ProtAnno could be extended with multiple layers for classification. The cell populations with low resolution, i.e., neutrophils and lymphocytes, could be identified first, and the cell subtypes, i.e., naive CD4 T cell and memory CD8 T cell, can be classified in the next few steps. Additionally, some steps in the parameter tuning algorithm could be replaced by advanced computational clustering methods, i.e., FlowSOM [15] and SPADE [37]. This may be helpful when the boundaries of cell types are unclear. Lastly, although ProtAnno does not need training data, a more unbiased association matrix inferred from more data could be a helpful add-on item to ProtAnno.

In summary, ProtAnno can provide a robust preliminary automated cell type annotation. Its performance could be further improved through adopting more advanced clustering approaches and more reliable references.

## 1.5 Materials and Methods

### 1.5.1 Optimization Procedure

**Step 1: Update $w$ with Lagrangian Multiplier**

In this step, we update the $K$-dimensional non-negative row vector $w_i$. Thus we have the new loss function by adding the Lagrangian multiplier as equation. Here to discriminate the primary constraint optimization with the dual Lagrangian function, we use $E$ to denote the latter.

$$\min_{w_i \geq 0} L(w_i; H) = \frac{1}{2} \parallel x_i - w_i H \parallel_F^2 - \lambda_1 tr(w_i^T (AS)_i) - \lambda_2 tr(w_i^T (A_0)_i) + \frac{\mu}{2} \parallel w_i \parallel^2 . \quad (1.7)$$

$$\min_{w_i} E(w_i; H) = \frac{1}{2} \parallel x_i - w_i H \parallel_F^2 - \lambda_1 tr(w_i^T (AS)_i) - \lambda_2 tr(w_i^T (A_0)_i) + \frac{\mu}{2} \parallel w_i \parallel^2 + \Phi w_i . \quad (1.8)$$

Here $x_i$ and $w_i$ are the $i$th row of $X$ and $W$, respectively. To minimize the loss function by updating row $w_i$, we have

$$\nabla E(w_i; H) = w_i(HH^T + \mu I) - x_i H^T - \lambda_1(AS)_i - \lambda_2(A_0)_i + \Phi. \tag{1.9}$$

By using the property of $\Phi w_i = 0$, we have

$$\begin{aligned} w_{ij}\nabla E(w_i; H)_j &= w_{ij}[w_i(HH^T + \mu I) - x_i H^T - \lambda_1(AS)_i - \lambda_2(A_0)_i + \Phi \\ &= w_{ij}[w_i(HH^T + \mu I) - x_i H^T - \lambda_1(AS)_i]_j - \lambda_2(A_0)_i]_j. \end{aligned} \tag{1.10}$$

We set the derivative to be zero. To get the non-negative minimizer, we decompose the items that are not non-negative in into positive and negative parts as:

$$\begin{aligned} x_i H^T = [x_i H^T]^+ - [x_i H^T]^- \lambda_1(AS)_i = [\lambda_1(AS)_i]^+ - [\lambda_1(AS)_i]^- \lambda_2(A_0)_i \\ = [\lambda_2(A_0)_i]^+ - [\lambda_2(A_0)_i]^-. \end{aligned} \tag{1.11}$$

By simple algebra, we have the following equation

$$\begin{aligned} w_{ij}([w_i(HH^T + \mu I)] + [x_i H^T]^- + [\lambda_1(AS)_i]^- + [\lambda_2(A_0)_i]^-)_j = w_{ij}([x_i H^T]^+ + [\lambda_1(AS)_i]^+ \\ + [\lambda_2(A_0)_i]^+)_j. \end{aligned} \tag{1.12}$$

Thus, we update $w_i$ by

$$w_{ij}^{t+1} = w_{ij}^{t} \frac{[xH^T]_j^+ + \lambda_1[(AS)_i]_j^+ + \lambda_2[(A_0)_i]_j^+}{[w_i(HH^T + \mu I)]_j + [xH^T]_j^- + \lambda_1[(AS)_i]_j^- + \lambda_2[(A_0)_i]_j^- + 10^{-16}}. \tag{1.13}$$

**Step 2: Update $H$ with multiplicative update algorithm**

In this step, we optimize $h$ column-wise.

$$\min_{h_j \geq 0} L(h_j; W) = \frac{1}{2} \parallel x_j - Wh_j \parallel_F^2 + \frac{\eta}{2} \parallel 1_K^T h_j - 1_N^T \parallel_2^2, \tag{1.14}$$

where $W$ is the constant matrix and $x_j$ is the $j$-th column vector of $X$. Similarly, by adding the Lagrangian multiplier, we have the following equations:

$$\min_{h_j} E(h_j; W) = \frac{1}{2} \parallel x_j - Wh_j \parallel_F^2 + \frac{\eta}{2} \parallel 1_K^T h_j - 1_N^T \parallel_2^2 + \Phi h_j \tag{1.15}$$

and

$$\nabla E(h_j; W) = (W^T W + \eta 1_K 1_K^T) h_j - W^T x_j - \eta 1_K + \Phi h_{ij} \nabla E(h_j; W) = 0. \tag{1.16}$$

We can update by

$$h_j^{t+1} = h_j^{t} \frac{[W^T x]_j^+ + \eta 1_k}{[(W^T W + \eta 1_K 1_K^T) h^t]_j + [W^T x]_j^- + 10^{-16}}. \tag{1.17}$$

### 1.5.2 Simulation Details

The expert matrix $A_0$ was generated from a multinomial distribution containing three categories, $+1$, $-1$, and $0$. We generated $100$ matrices $A_0$ and selected the optimal one that had the smallest inner product between the vectors by minimax. Such construction can more likely enable the expression profile $X$ to be a nonnegative linear combination of basis vectors.

In most cases, the expert matrix may be biased. We therefore generated an intermediate discrete matrix $\tilde{A}_0$ with three possible discrete values, 2, 1, and $0$, to represent the expert knowledge matrix used. They were matched with three biomarker distributions within a cell population: high, low, and no expression. In practice, if the entry in $A_0$ was $1$, the corresponding gene would have high expression levels in real data from our observations. Thus, the corresponding element in $\tilde{A}_0$ kept the value, $1$. However, if the entry in $A_0$ was $-1$, it is possible to have low or medium expression level instead of no expression completely. In some rare instances, the corresponding gene might even have high expression. Therefore, the values of $-1$ would random walk to $1$ with probability $q_{-1}$ and to 2 with probability $p_{-1}$in $\tilde{A}_0$. If the entry in $A_0$ was $0$, then the distribution was uncertain. In this case, the values of $0$ would random walk to $1$ with probability $q_0$ and to $2$ with probability $p_0$ in $\tilde{A}_0$. This generated the new discrete matrix $\tilde{A}_0$ based on the above random walk protocols to distinguish the different biomarkers behaviors.

Next, the signature matrix $W$ was generated based on $\tilde{A}_0$ from two positive truncated normal distributions with different expectations. When the entry in $\tilde{A}_0$ was $2$, the corresponding signature gene average expression was sampled from a positive truncated normal distribution with a large mean; when the entry in $\tilde{A}_0$ was $1$, the corresponding signature gene average expression was sampled from a half-normal distribution with a small mean; otherwise, it was set to $0.1$.

After randomly simulating the cell type labels for each single cell in the proteomics expression profile $X$, the expression level was sampled from normal distributions with the expectation of average signature gene expression in $W$ and variance calculated from sampled mean and the specified mean-variance ratio.

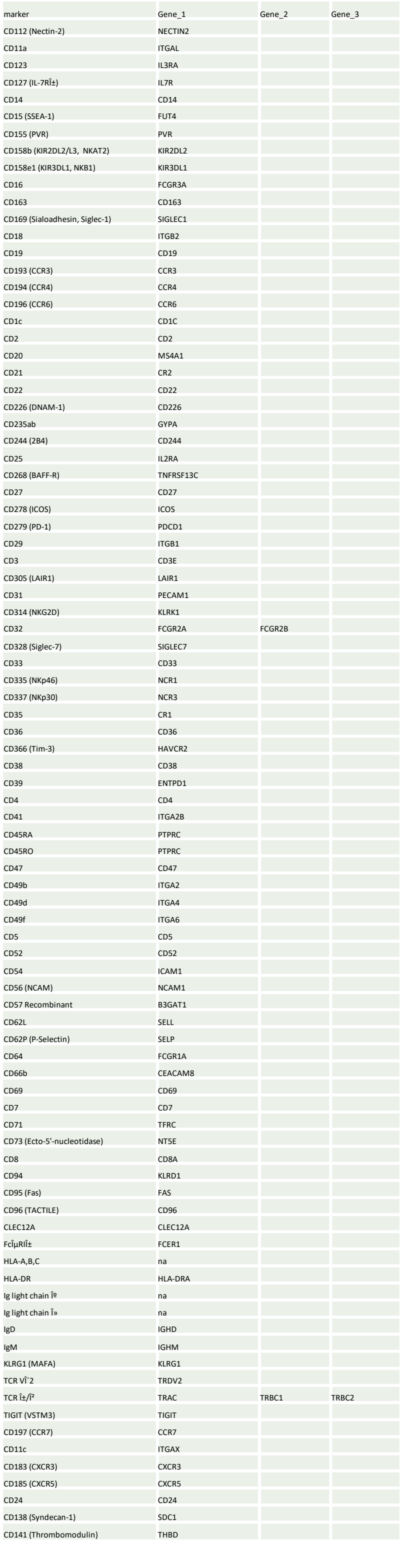
### 1.5.3 Annotation metrics

ARI was calculated using the function adjustedRandIndex from the R package mclust [38]. For NMI, the ground truth and the predicted label for each cell were converted into one-hot vectors and the function NMI in the R package aricode [39] was used to calculate the NMI for each cell. The reported result was the average across all the cells. Cosine similarity was calculated by first computing the cosine measure between the ground truth one-hot vector and the cluster assignment vector using the cosine function from the lsa package in R, and finally averaging over all the cells. ASW was calculated using the batch_sil function in the R package kBET [40].

### 1.5.4 Real Datasets and processing

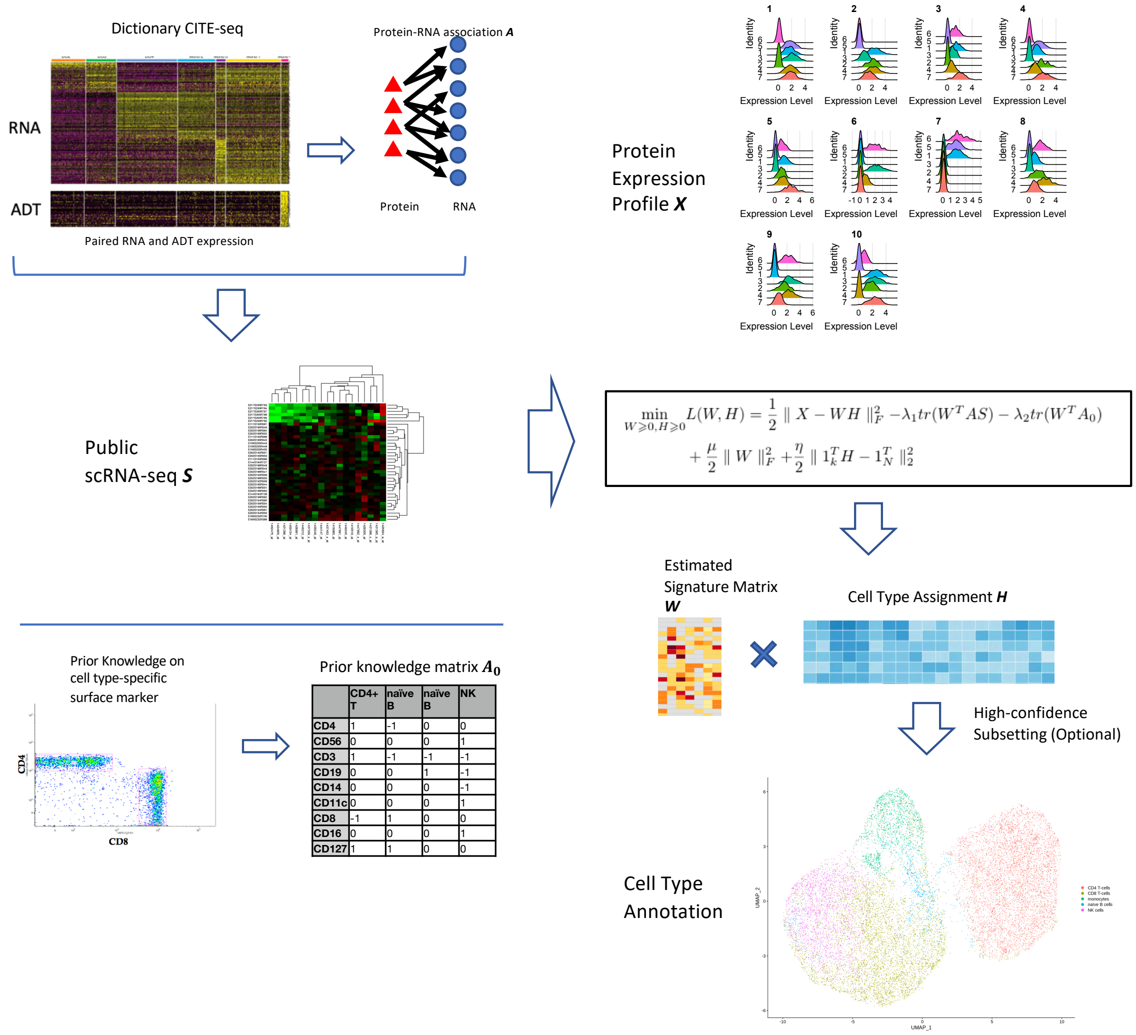
The BCR CyTOF dataset was downloaded through the HDCytoData [41] R package. The cell types were labeled by [25] and could be accessed by the HDCytoData package. The CITE-seq data in [27] were downloaded from GEO with accession ID GSE148665. The longitudinal covid-19 CyTOF data were downloaded from (`https://brodinlab.com/data-repository/`). The two dictionary CITE-seq datasets in [19] and [20] would be available after they publish the data. The CITE-seq data in [27] and [20] were manually labeled after the preliminary results by SingleR [17]. The expert-guided matrices were obtained from the biomarker panels in [19].

All the proteomics data were normalized by arcsinh function with cofactor 5. We cleaned and gated the longitudinal CyTOF data by the R package flowCore to filter the intact cells. When processing the association matrix from CITE-seq and transcriptomics signature matrix, all the scRNA-seq data are denoised by SAVERX [21]. The transcriptomics gene list for each application contains the matched protein-coding transcript genes (Table 4) and the top highly variable genes generated by the Seurat package [42].



| marker | Gene_1 | Gene_2 | Gene_3 |
|---|---|---|---|
| CD112 (Nectin-2) | NECTIN2 | | |
| CD11a | ITGAL | | |
| CD123 | IL3RA | | |
| CD127 (IL-7RÎ±) | IL7R | | |
| CD14 | CD14 | | |
| CD15 (SSEA-1) | FUT4 | | |
| CD155 (PVR) | PVR | | |
| CD158b (KIR2DL2/L3, NKAT2) | KIR2DL2 | | |
| CD158e1 (KIR3DL1, NKB1) | KIR3DL1 | | |
| CD16 | FCGR3A | | |
| CD163 | CD163 | | |
| CD169 (Sialoadhesin, Siglec-1) | SIGLEC1 | | |
| CD18 | ITGB2 | | |
| CD19 | CD19 | | |
| CD193 (CCR3) | CCR3 | | |
| CD194 (CCR4) | CCR4 | | |
| CD196 (CCR6) | CCR6 | | |
| CD1c | CD1C | | |
| CD2 | CD2 | | |
| CD20 | MS4A1 | | |
| CD21 | CR2 | | |
| CD22 | CD22 | | |
| CD226 (DNAM-1) | CD226 | | |
| CD235ab | GYPA | | |
| CD244 (2B4) | CD244 | | |
| CD25 | IL2RA | | |
| CD268 (BAFF-R) | TNFRSF13C | | |
| CD27 | CD27 | | |
| CD278 (ICOS) | ICOS | | |
| CD279 (PD-1) | PDCD1 | | |
| CD29 | ITGB1 | | |
| CD3 | CD3E | | |
| CD305 (LAIR1) | LAIR1 | | |
| CD31 | PECAM1 | | |
| CD314 (NKG2D) | KLRK1 | | |
| CD32 | FCGR2A | FCGR2B | |
| CD328 (Siglec-7) | SIGLEC7 | | |
| CD33 | CD33 | | |
| CD335 (NKp46) | NCR1 | | |
| CD337 (NKp30) | NCR3 | | |
| CD35 | CR1 | | |
| CD36 | CD36 | | |
| CD366 (Tim-3) | HAVCR2 | | |
| CD38 | CD38 | | |
| CD39 | ENTPD1 | | |
| CD4 | CD4 | | |
| CD41 | ITGA2B | | |
| CD45RA | PTPRC | | |
| CD45RO | PTPRC | | |
| CD47 | CD47 | | |
| CD49b | ITGA2 | | |
| CD49d | ITGA4 | | |
| CD49f | ITGA6 | | |
| CD5 | CD5 | | |
| CD52 | CD52 | | |
| CD54 | ICAM1 | | |
| CD56 (NCAM) | NCAM1 | | |
| CD57 Recombinant | B3GAT1 | | |
| CD62L | SELL | | |
| CD62P (P-Selectin) | SELP | | |
| CD64 | FCGR1A | | |
| CD66b | CEACAM8 | | |
| CD69 | CD69 | | |
| CD7 | CD7 | | |
| CD71 | TFRC | | |
| CD73 (Ecto-5'-nucleotidase) | NT5E | | |
| CD8 | CD8A | | |
| CD94 | KLRD1 | | |
| CD95 (Fas) | FAS | | |
| CD96 (TACTILE) | CD96 | | |
| CLEC12A | CLEC12A | | |
| FcÎµRIÎ± | FCER1 | | |
| HLA-A,B,C | na | | |
| HLA-DR | HLA-DRA | | |
| Ig light chain Îº | na | | |
| Ig light chain Î» | na | | |
| IgD | IGHD | | |
| IgM | IGHM | | |
| KLRG1 (MAFA) | KLRG1 | | |
| TCR VÎ´2 | TRDV2 | | |
| TCR Î±/Î² | TRAC | TRBC1 | TRBC2 |
| TIGIT (VSTM3) | TIGIT | | |
| CD197 (CCR7) | CCR7 | | |
| CD11c | ITGAX | | |
| CD183 (CXCR3) | CXCR3 | | |
| CD185 (CXCR5) | CXCR5 | | |
| CD24 | CD24 | | |
| CD138 (Syndecan-1) | SDC1 | | |
| CD141 (Thrombomodulin) | THBD | | |

Table 1.4: Protein-coding surface marker and transcript matching table.


Dictionary CITE-seq
RNA
ADT
Paired RNA and ADT expression
Protein-RNA association A
Protein
RNA
Protein Expression Profile X
Identity
Expression Level
Public scRNA-seq S
Prior Knowledge on cell type-specific surface marker
CD4
CD8
Estimated Signature Matrix W
Cell Type Assignment H
High-confidence Subsetting (Optional)
Cell Type Annotation
UMAP_1
UMAP_2
CD4 T cells
CD8 T cells
monocytes
naive B cells
NK cells

$$\min_{W\geqslant 0, H\geqslant 0} L(W,H) = \frac{1}{2} \parallel X - WH \parallel_F^2 - \lambda_1 tr(W^T AS) - \lambda_2 tr(W^T A_0) + \frac{\mu}{2} \parallel W \parallel_F^2 + \frac{\eta}{2} \parallel 1_k^T H - 1_N^T \parallel_2^2$$

Prior knowledge matrix $A_0$

| | CD4+ T | naïve B | naïve B | NK |
|---|---|---|---|---|
| CD4 | 1 | -1 | 0 | 0 |
| CD56 | 0 | 0 | 0 | 1 |
| CD3 | 1 | -1 | -1 | -1 |
| CD19 | 0 | 0 | 1 | -1 |
| CD14 | 0 | 0 | 0 | -1 |
| CD11c | 0 | 0 | 0 | 1 |
| CD8 | -1 | 1 | 0 | 0 |
| CD16 | 0 | 0 | 0 | 1 |
| CD127 | 1 | 1 | 0 | 0 |

Figure 1.1: Model overview. ProtAnno first generates a protein-RNA association matrix $A$ trained on CITE-seq dictionary by Elastic Net and a transcriptome signature matrix $S$ from scRNA-seq with cell type annotations. The product of $A$ and $S$ is the predicted protein level signature matrix based on mRNA counts. The other input is the discrete matrix $A_0$ to represent the general expert biomarker knowledge. The loss function is an NMF optimization program on protein expression matrix to deconvolute into estimated protein signature matrix $W$ and cell type assignment matrix $H$, with penalizations on $AS$ and $A_0$ and regularizations on $W$ scale and column sum of $H$. The matrix $H$ could be optionally trimmed to the cells which have high-confidence assignment in the presence of high noise in $X$. The final annotations of ProtAnno are based on the largest value of $H$ column-wise.

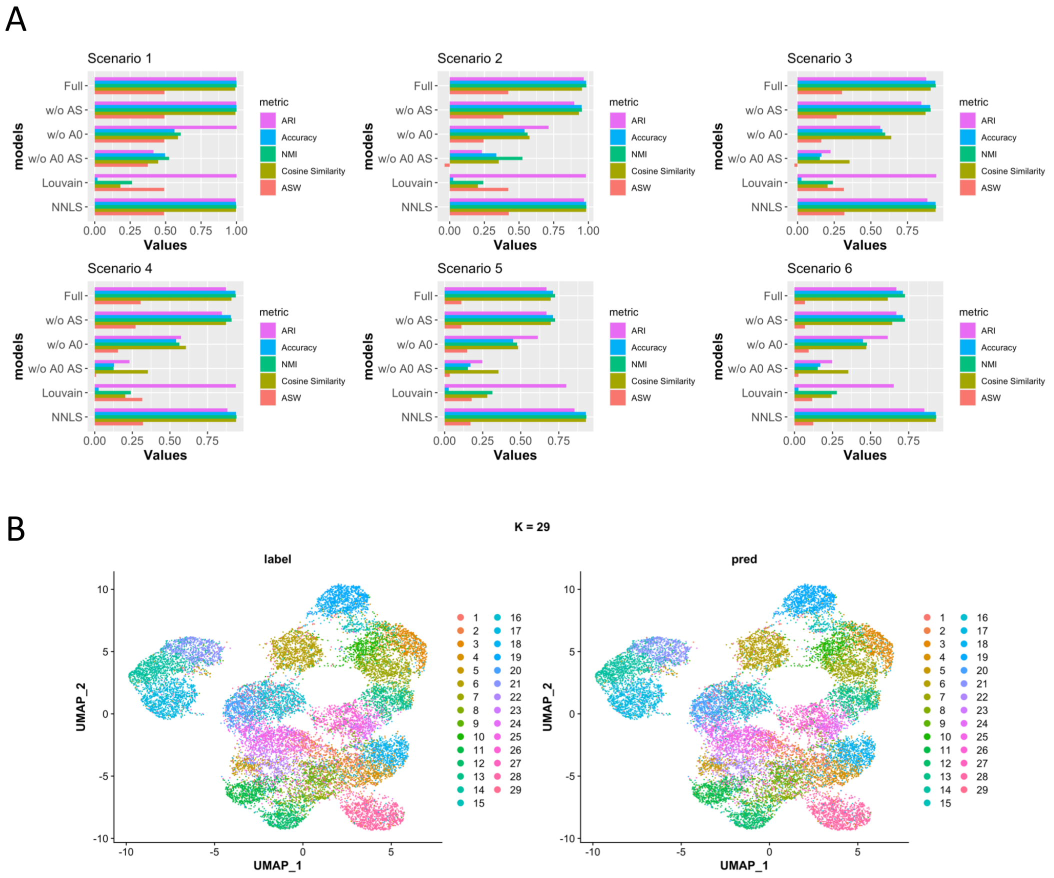


Figure 1.2: ProtAnno simulation benchmarking results. A) Comparisons of benchmarking annotation results under six simulation models. The x-axis lists the six models; the y-axis shows the annotation metric values. B) An example of ProtAnno annotation when the number of cell types $K = 29$. The left panel is the UMAP plot on true labels; the right panel is the UMAP plot of ProtAnno annotations.

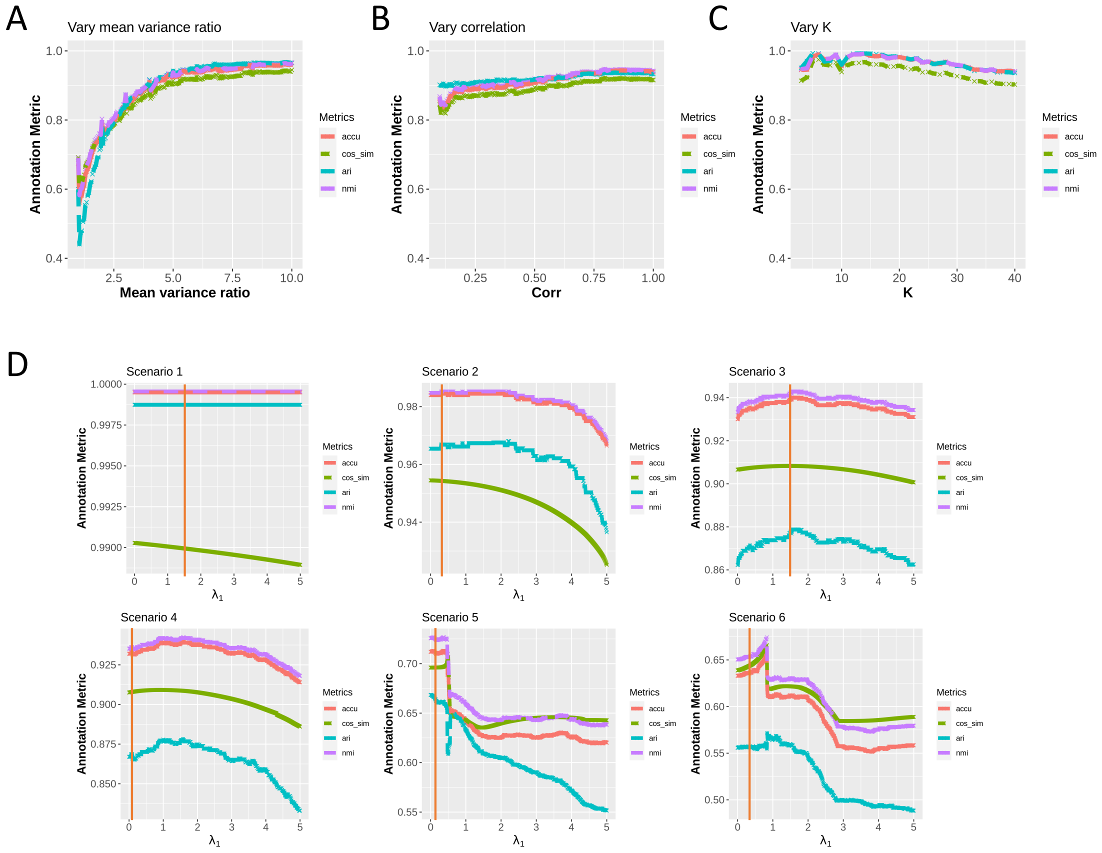


Figure 1.3: ProtAnno model robustness evaluations. A) The relationship between ProtAnno annotation accuracy and expression variation. The x-axis represents the expected ratio of mean and variance. A higher value indicates lower expression variation. The y-axis shows the values of annotation metrics. The curves represent the results from 100 simulations for each mean-variance ratio value. B) The effect of transcriptomics data on a full ProtAnno model. The x-axis represents the correlation between transcriptomics predicted signature matrix $AS$ and real protein signature matrix $W$. The y-axis represents the value of annotation metrics. C) ProtAnno robustness as a function of cell type number $K$. The x-axis represents cell type number that varies from 3 to 40. The y-axis represents the value of annotation metrics. D) Evaluation of parameter tuning algorithm. The x-axis represents the change of $\lambda_1$ from 0 to 5 under six simulation scenarios. The color represents different annotation metrics. The y-axis represents the value of annotation metrics. The red vertical line represents the optimal $\lambda_1$ by the parameter tuning algorithm for each simulation scenario.

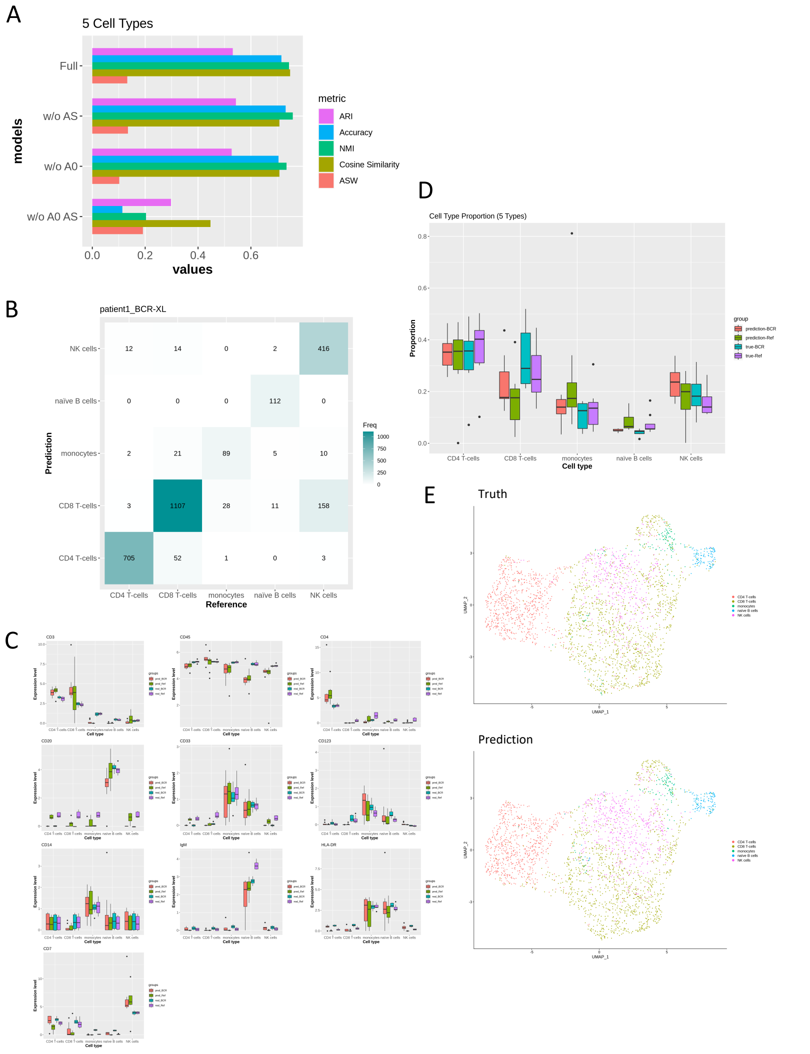


Figure 1.4: ProtAnno results on BCR cytometry data. A) Benchmarking ProtAnno on five labeled cell types. The x-axis lists the four ProtAnno models; the y-axis is the annotation metric value. B) Confusion matrix of ProtAnno annotation on BCR stimulated patient 1. The y-axis is the ProtAnno results compared with the true cell type labels as a reference in the x-axis. C) The boxplots of estimated signature expression in $W$. The x-axis represents the cell type and the y-axis is the distribution of values in the signature matrix across patients within groups. The first two columns are ProtAnno annotation results in the stimulated and unstimulated BCR groups and the last two columns are the true average signature expression distributions.D) Cell type proportions boxplots. In this figure, we compare cell type proportions estimated by ProtAnno with true proportions. The first two columns are ProtAnno annotation results in the stimulated and unstimulated BCR groups and the last two columns are the true cell type proportions in these two groups.E) The UMAP plot. The upper panel is the true annotation; the lower panel is the assignment by ProtAnno.

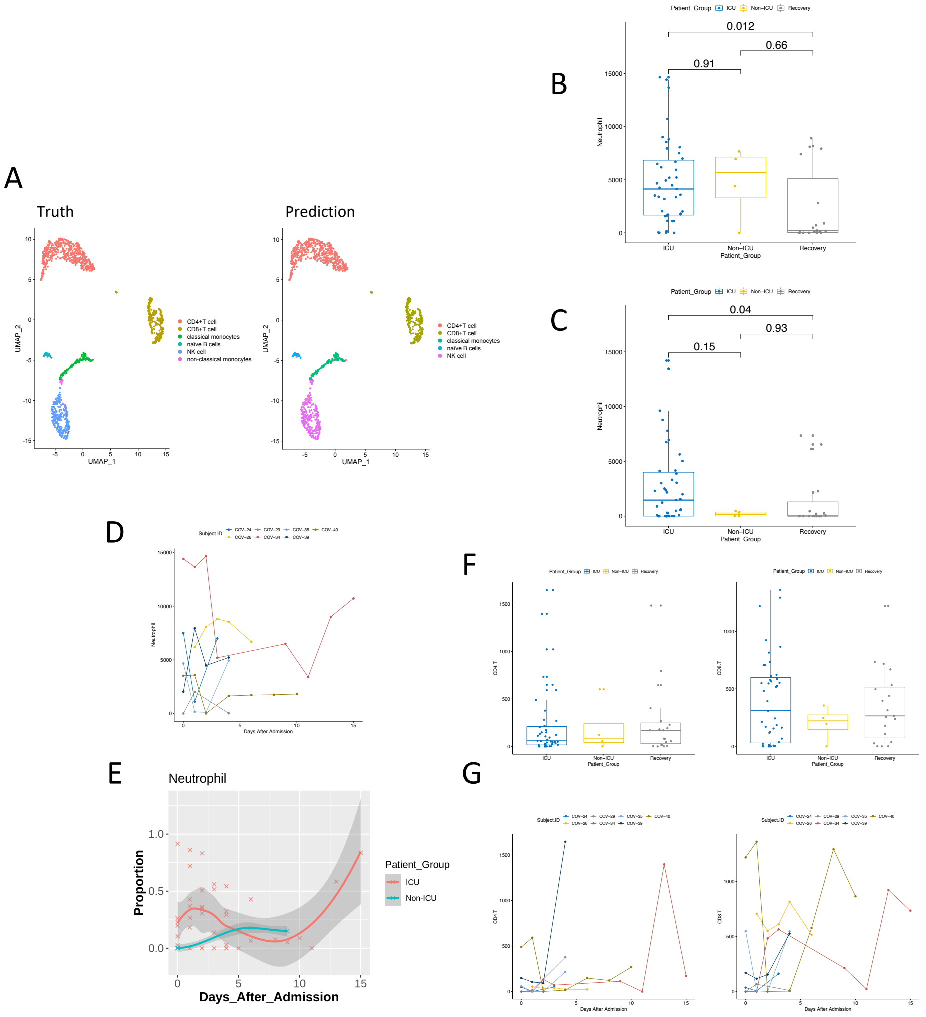


Figure 1.5: ProtAnno results on PBMC CITE-seq data and temporal whole blood covid-19 cytometry data. A) The UMAP plot of the proteomic data in the PBMC CITE-seq study in Wang et al. (2020)(X. Wang et al. 2020). The left panel is the true annotation; the right panel is the assignment by ProtAnno. The UMAP projections are colored by the cell types. B) and C) Boxplots of cell type counts of neutrophil cells from the whole blood covid-19 study in Rodriguez et al. (2020)(Rodriguez et al. 2020) by ProtAnno across different patient groups i.e., ICU, non-ICU, and recovery patients. B) is the raw output of ProtAnno. C) is the subsetted output of ProtAnno for keeping high-confidence assignments only. The paired Wilcoxon test p-values are shown on the top of the figures. D) Neutrophil cell counts estimation by ProtAnno over the days after admission. The color represents the patient who has the longitudinal cytometry data. E) Neutrophil cell type proportions estimated by ProtAnno over the days after admission. Two curves represent the ICU and non-ICU patient groups. F) Boxplots of CD4 T cell and CD8 T cell counts across groups (ICU, non-ICU, and recovery patients). G) CD4 T and CD8 T cell type proportions estimation by ProtAnno over the days after admission. The color represents the patient who has the longitudinal cytometry data.

# Chapter 2

# A Novel Bayesian Framework for Harmonizing Information across Tissues and Studies to Increase Cell Type Deconvolution Accuracy

## Abstract

Computational cell type deconvolution on mixture transcriptomics data can reveal cell type proportion heterogeneity across bulk samples. One critical factor for accurate deconvolution is the reference signature matrix for different cell types. Compared to the reference signature matrices inferred from cell lines, rapidly accumulating data from single-cell RNA-sequencing (scRNA-seq) provide another rich resource for deriving cell type signatures. However, there are challenges in scRNA-seq data because of the high biological and technical noises. In this chapter, we introduce a novel Bayesian framework, tranSig, to improve signature matrix inference from scRNA-seq by leveraging shared cell type-specific expression patterns across multiple tissues and studies. Our simulations show the robustness of tranSig with the number of signature genes and tissues specified in the model. The applications of tranSig to bulk RNA sequencing data from peripheral blood, bronchoalveolar lavage, and aorta data demonstrate its accuracy and power to characterize biological heterogeneity across groups. In summary, tranSig offers an accurate and robust approach for defining gene expression signatures of different cell types, facilitating in silico cell type deconvolutions.

*Key words*: cell type deconvolution; scRNA-seq; reference signature matrix; harmonize information

### 2.0.1 Introduction

Characterizing cell type composition in tissues can help better understand disease pathogenesis and progression. Traditional approaches for inferring cell type proportions, such as flow cytometry [43][44][45] and immunohistochemistry (IHC) [46][47], involve complicated protocols, expensive antibodies, high end platforms, and expertise. As these approaches are mainly based on antigen-antibody reaction, they are limited to the specificity of antibodies. In addition, techniques such as cytometry-based platforms may not always yield accurate cell type proportions due to different preservation rates across cell types [48]. With the rapid development of RNA sequencing (RNA-seq) technologies and accumulation of bulk RNA-seq data in recent years, there has been a surge of computational deconvolution methods [49][50][51] to factorize tissue-level RNA-seq data on their mixture gene expression profiles (MGEP) to infer cellular compositions. One critical component of these methods is the signature expression profiles (signature matrix) of known cell types largely derived from RNA-seq data of enriched or purified cell populations [52][53][54]. However, the cell types are limited to the well-defined ones, and it is hard to incorporate less-studied cell types or subtypes.

With the development of single cell RNA sequencing (scRNA-seq) technologies [52][53][54], transcriptomes can be characterized at the single cell resolution. Recent years have seen the applications of scRNA-seq to study cell types [55][56], embryonic and organic development [57][58][59], disease mechanisms, and many other fields [60][61]. There is a rapid accumulation of single cell datasets, with several cell atlas projects, such as Human Cell Atlas (HCA) [10][62] and Human Cell Landscape (HCL) [63], aiming to construct and cover the primary tissues of healthy donors. These single cell datasets and bulk RNA-seq data can facilitate cell type deconvolution through the development of a more accurate signature matrix at higher resolution. However, scRNA-seq data are more sparse and noisier than bulk RNA-seq data, and there are platform differences between scRNA-seq and bulk RNA-seq data. Furthermore, although cell type proportions can be directly obtained from scRNA-seq data, these estimates may be biased due to different capture rates across cell types.

Existing computational cell type deconvolution methods can be divided into two categories based on the references used to derive cell type signatures. The first category utilizes the signature matrix derived from bulk RNA-seq data, with CIBERSORT [50] as one representative. The second category includes methods that use scRNA-seq data to derive the signature matrix. For example, CIBERSORTx leverages single cell data and offers both S mode and B mode to address the technical differences between scRNA-seq and bulk RNA-seq data. Another popular method, MuSiC [64], aims to select the signature genes by multi-subject comparisons, which are used in the subsequent deconvolution. However, all these methods focus on using data from a specific tissue in a specific study, and are not able to capitalize on shared patterns for the same cell type across different tissues and studies. For example, immune cells show conserved cross-tissue transcriptomes [65][66] across tissues and developmental stages, and such shared patterns may provide additional information for signature matrix construction.

Here, we propose a novel Bayesian framework, called tranSig, to better infer a signature matrix from scRNA-seq data by leveraging cross-tissue information. Through benchmarking comparisons and real applications on peripheral blood, bronchoalveolar lavage, and aortic aneurysm, we show that tranSig can facilitate more accurate estimates of cell proportions and better interpretations of the pathogenesis of diseases.

## 2.1 Methods

To leverage information from multiple single cell references, we have developed a new framework, called tranSig, based on transfer learning to handle cross-platform and cross-tissue variations to derive a more accurate signature matrix for downstream cell type deconvolution (Fig 2.1).

As mentioned above, the existing methods mostly generate their signature matrices based on one scRNA-seq dataset from the target tissue. Due to the challenges of eliminating technical batch effects and discriminating biological differences, people are reluctant to integrate data from multiple tissues, or studies on the same tissue. In contrast, in tranSig, we assemble the target tissue data with various reference datasets from other tissues or studies. Specifically, we consider the Human Cell Landscape (HCL) [63] and manually cleaned the cell type annotations to cover 25 adult and five fetal tissue types as sources for both target and reference tissues. We project the reference scRNA-seq datasets on the target one by adopting the matrix factorization-based single cell batch correction method, LIGER [67][68]. Although this may alleviate some batch effects, it is still necessary to identify the remaining cross-dataset variations after batch-effect correction and remove tissue-specific signatures in the reference datasets. To accomplish this, as detailed in the Methods section, we compare every reference signature matrix with the target signature matrix to select cell type-conserved signature that may share the common distribution with the target tissue. Finally, we only keep the cell type-specific expression profiles in references that have a similar distribution with the target tissue.

Our hierarchical Bayesian model considers the batch effect-corrected single cell expression profile after batch correction, $x^t \in \mathbb{R}^{D \times M_t}$ where $M_t$ denotes the number of cells in the scRNA-seq dataset, and $t$ denotes the tissue type. When gene $d = 1, \ldots, D$ in cell type $k$ from reference tissue $t$ shares the common distribution with the target tissue and is selected in the last step,

$$x^t_{dc_k} \sim N(w^t_{dk}, \frac{1}{\tau^x_d}) \tag{2.1}$$

where $x^t_{dc_k}$is the expression level of gene $d$ in the $c_k$th cell of cell type $k$ from tissue $t$, $c_k$ indexes the individual cells of cell type $k$, explain $w^t_{dk}$ and $\tau^x_d$ is the gene-specific precision for the Gaussian distribution. We use $w^t \in \mathbb{R}^{D\times K}$, where $t = 0, 1, \ldots, T$, to denote the intermediate dataset-specific signature matrix for tissue $t$. The tranSig model assumes that the intermediate dataset-specific signatures, $w^t_{dk}$, share the same tissue-specific mixture Gaussian distribution across $t$. Mathematically, for every signature gene $d$ and cell type $k$, the corresponding value in the signature matrix follows the mixture Gaussian distribution

$$w^t_{dk} \sim N(v^0_{dk} * \gamma^0_{dk}, \frac{1}{\tau_w}) \tag{2.2}$$

where $\gamma^0_{dk}$ is a binary variable indicating whether gene $d$ is a signature gene thus has non-zero expression for cell type $k$ in the target tissue $t = 0$, and $v^0_{dk}$ is the continuous part when the gene is estimated to be a non-zero expressed signature as $\gamma^0_{dk} = 1$. And $\tau_w$ is the shared precision for all $w^t_{dk}$. Overall, our model outputs the final tranSig signature matrix as estimated $v^0 \odot \gamma^0$, which is the element-wise product of matrices $v^0$and $\gamma^0$. We implemented the State-Augmentation for Marginal Estimation (SAME) [69] to accelerate the algorithm.

To better align bulk RNA-seq data with the signature matrix inferred from single cell references, we perform batch correction similar to that of the CIBERSORTx S mode [70] and utilize the pseudo bulk mixtures derived from scRNA-seq for more accurate estimates in real applications. We generate the pseudo bulk mixtures by sampling from the single cell references, and implement the empirical Bayes (EB) batch-effect removal model, Combat [71], to adjust bulk RNA-seq mixtures to single cell dataset of the target tissue. The adjusted bulk RNA-seq data are taken as the input for downstream deconvolution along with the tranSig signature matrix.

We view the tranSig framework as an add-on step for cell type deconvolution. Therefore, it can be coupled with any existing cell type deconvolution methods, e.g. NNLS and CIBERSORTx, to estimate cell type proportions.

## 2.2 Results

### 2.2.1 Robustness evaluation through simulations

We have performed simulations to evaluate the robustness of our proposed tranSig model. Since it is challenging to simulate cross-platform or cross-tissue effects, we focused on testing the robustness assuming that both bulk RNA-seq and multi-tissue scRNA-seq batch-effects have been successfully removed.

We assumed that the true signature matrix of the target tissue $t = 0$ contains two parts: the binary matrix $\gamma^0$ indicating whether each gene is an expressed cell type-specific signature gene, and the continuous matrix $v^0$ quantifying the average expression levels of signature genes. The signatures from all the input single cell datasets, including both the target and reference datasets, share the same underlying distribution, which is the product of $v^0$ and $\gamma^0$. Specifically, we simulated $w^t_{dk}$ from a Gaussian distribution with mean $v^0_{dk} * \gamma^0_{dk}$ and let $w^t_{dk}$ be the mean of Gaussian distribution of $x^t_{dkc_k}$.

We note the possibility that an expressed signature gene may not be detected due to the low sequencing depth of scRNA-seq and other factors. The droplet-based single cell RNA-seq can theoretically detect $5000$ genes per cell as the saturated number of detected genes [72], but the number of detected genes in each cell is often lower in published studies due to the sequencing depth (median number of detected genes: 256-602 in HCL; 1973 in ATAA [73]). Therefore, in our simulations, we set the expression level for an expressed gene to be zero with a certain probability, which corresponds to the undetected rate. A higher undetected rate can result in a loss of more signature information when constructing the empirical signature matrix by averaging the single cell expression profile grouped by cell types. We evaluated the performance of the tranSig model combined with both NNLS and CIBERSORTx [50] compared with other methods, including MuSiC [64], CIBERSORTx with input empirical signature matrix (empirical + CIBERSORTx), and

NNLS from MuSiC without weights.

We investigated how the undetected rate and signature gene number can affect cell type deconvolution and the results are shown in Fig 2.2A. We can see that tranSig combined with CIBERSORTx (tranSig + CIBERSORTx) outperformed the other methods with higher correlation between the true and estimated cell type proportions. In the left panel, the performance of tranSig + CIBERSORTx was most stable with different undetected rates. When constructing a signature matrix by differential expression (DE) analysis, the number of signature genes depends on the method and threshold selected in DE analysis. We investigated how the number of signature genes could influence the cell type proportion estimation to evaluate the robustness of our method. The right panel in Fig 2.2A shows that the two tranSig methods and CIBERSORTx have good performance with a relatively small number of signature genes, e.g. 150. In contrast, MuSiC and NNLS in the MuSiC R package require multiple subjects and cross-subject variation in single cell expression profiles to have good performance.

We also show the scatter plots of the true and estimated cell type proportions to illustrate tranSig's performances across cell types (Fig 2.2B). When there were eight cell types, 500 signature genes, and an undetected rate of $25\%$, two tranSig methods with CIBERSORTx and NNLS had the best estimation, and all points centered around a single line. CIBERSORTx also had relatively good performance but was not as accurate. All the methods could not infer the proportions of rare cell types (less than $10\%$) well. The two tranSig methods tended to overestimate more prevalent cell type proportions while underestimate rare cell type proportions.

### 2.2.2 Bulk peripheral blood (PB) deconvolution to handle cross-tissue and cross-platform variations

To evaluate the performance of tranSig on real data, we first analyzed PB that is composed of easily distinguishable immune cells, including monocytes, neutrophils, T cells, B cells, and others. We applied tranSig to bulk RNA-seq data of whole blood from 12 healthy adults to estimate cell type proportions [70]. For single cell references, we took the PB as the target tissue, and bone marrow (BM), cord blood (CB), lung, kidney, and liver as the reference tissues. The union of highly expressed genes (HEGs) of each cell type in each tissue was considered as the signature gene list (details in Methods). NNLS, CIBERSORTx, and quadratic programming [74] were used for deconvolution analysis. The correlations between the estimated cell type proportions and the "ground truth" obtained from flow cytometry were calculated to assess the deconvolution performance. As shown in Fig 2.3A, tranSig had more accurate estimation than that based on the empirical signature matrix for all three deconvolution methods. Among three deconvolution methods, both NNLS and CIBERSORTx had more accurate proportion estimates for most cell types with tranSig. Overall, tranSig + CIBERSORTx was most consistent.

Within the tranSig framework, we made adjustments to both single cell references and bulk RNA-seq expression profiles and evaluated these adjustments in real applications of PB deconvolutions (Fig A.8). With LIGER implementation, the shared signature genes cross tissues were selected and improved the deconvolution results of tranSig. For the technical batch correction, the bulk mixture adjusted to the space of scRNA-seq as input of deconvolution increases the accuracy of tranSig. In addition, we compared the two types of single cell expression profiles (i.e., raw counts and transcripts per million (TPM) normalization) as the input of the tranSig model. The results suggest that using raw counts as the input may be superior to TPM normalization due to the estimation of $\tau_{x_d}$ in the tranSig model. Therefore, we used the raw counts as the input of the tranSig model in all subsequent analyses, unless stated otherwise.

To systematically benchmark the performance of different methods (Fig 2.3B), we implemented MuSiC and CIBERSORTx with the S mode and B mode. Overall, the performance of tranSig + CIBERSORTx was superior to other methods for this dataset. It is interesting to note that all three CIBERSORTx modes, including disabled batch correction mode, S mode, and B mode, accurately estimated the proportions of B cells and T cells but failed to estimate those of monocytes and neutrophils. Because of HCL datasets were generated by Microwell-seq as a UMI-based sequencing, the deconvolution results show that the S mode may substantially improve the overall estimation accuracy but perform poorly for neutrophils and monocytes (Fig A.7). Specifically, the estimated neutrophil proportions were smaller than $10\%$ when the ground truth was around $60\%$. For MuSiC, although the proportions of neutrophils, T cells, and B cells were accurately estimated, the performance of monocyte estimation was worse than either tranSig or CIBERSORTx. Taken together, the cell type deconvolution with tranSig by CIBERSORTx achieved higher accuracy as well as less variance among cell types than the other methods.

### 2.2.3 Bulk bronchoalveolar lavage (BAL) deconvolution in a Sarcoidosis cohort to identify dominant cell type

Sometimes the mixture data have only one dominant cell type. We note this case as highly unbalanced cell type proportions. It is important for a cell type deconvolution method to identify the dominant cell type and distinguish it from other cell populations in such cases. To compare the performance of different methods for this scenario, we considered a Sarcoidosis cohort [75] BAL bulk RNA-seq dataset in which the proportion of the alveolar macrophages (AM) was around $80\%$ [76].

We used adult PB in HCL as the target single cell dataset and the other five tissues (adult adipose, adult BM adult lung, CB, and fetal lung) as reference single cell datasets for joint analysis in tranSig, since all these five tissues have immune cells as their major cell populations.

The AM is critical to lung inflammation and repairment [77]. They work closely with type I and II epithelial cells, cellular and functionally different from monocyte-derived macrophages in PBMC. Therefore, we removed $48$ AM differentially expressed genes (Supplementary E.1) from the signature gene list so that only AM and macrophages commonly shared signature genes would be used for deconvolution. It can largely reduce the effects brought by the heterogeneity between AMs and macrophages in PB.

This data set has macrophages as the largest cell population and lymphocytes as the second largest. As shown in Fig 2.4, both tranSig and the NNLS model with the empirical signature matrix estimated the macrophages with proportions of around $70\%$. However, the latter failed to identify T cells and estimated the second dominant cell type as dendritic cells with the mean proportion as high as around $30\%$, which is unreasonable [75]. In comparison, tranSig successfully recognized the T cells as another major cell type in BAL.

### 2.2.4 Bulk aorta deconvolution to depict cellular pathological changes of aneurysm (AN)

We further assessed whether tranSig deconvolution can reflect the pathological changes of tissues with results shown in Fig 2.5. We applied tranSig to bulk RNA-seq of aortic or aneurysm [78] tissues from six healthy donors and six patients with ascending aortic aneurysm. Aortic aneurysm [79][80] is a permanent and localized dilation of the aorta, and is a fatal vascular disease. The pathological features [81] of aortic aneurysm were well-studied, including apoptosis of smooth muscle cells (SMCs) [82], infiltration of immune cells [83] (i.e., macrophages [84][85] and T cells [86][87], matrix metalloproteinase increase [88] and elastin degradation. First, we took the artery dataset from HCL as the target tissue and lung, BM, PB, CB, kidney and liver as the reference tissues. Deconvolution with empirical signature matrix by NNLS could infer increased macrophages and reduced SMCs and stromal cells but missed the signals of other immune cells. As for tranSig, the deconvolution results showed increased macrophages and T cells, consistent with the inflammatory infiltration, and the decreased SMCs and stromal cells, suggestive of SMC apoptosis in the pathogenesis of aortic aneurysm. Similar to the results of PB deconvolution, CIBERSORTx was more stable than NNLS. We found that CIBERSORTx failed to estimate immune cells, and there was no obvious improvement while implementing the S mode or B mode. Because of the lack of any other subject in the artery dataset from HCL, MuSiC could not be applied in this case. Then for more fair comparison, we implemented CIBERSORTx and MuSiC on another scRNA-seq expression profile of aortic tissues [73] from three healthy donors and eight patients with ascending aortic aneurysm (ATAA dataset) which had a deeper sequencing depth compared with the artery dataset from HCL. Similar to the deconvolution with the HCL artery, CIBERSORTx missed immune cells. Almost all cells in the aorta or aneurysm were estimated as SMCs by MuSiC, which cannot interpret the infiltration of inflammation in the pathogenesis of aortic aneurysm.

Additionally, we took the ATAA dataset as the target tissue and lung, BM, PB, CB, kidney, and liver from HCL datasets as the reference tissues to evaluate the performance of leveraging information across studies and platforms (Fig A.9). Consistent with the above results, deconvolution with tranSig + CIBERSORTx was more stable than NNLS. As for CIBERSORTx B mode and S mode, they failed to estimate immune cells similar to the results of HCL. However, by implementing tranSig with CIBERSORTx, the results exhibited reasonable trends of SMC and immune cells, and the substantial improvement compared with CIBERSORTx further demonstrated the benefit of tranSig across studies and platforms.

## 2.3 Discussion

In this study, we have developed tranSig, a novel Bayesian model to better infer signature matrix by transferring learning over multiple scRNA-seq datasets. In the tranSig framework, we use SAME [69] for statistical inference and estimate a more reliable signature matrix by a Gaussian mixture prior. Highly expressed genes of cell types are screened as signature genes. tranSig implements LIGER [67][68] and k-means to integrate cross-tissue and cross-study scRNA-seq datasets. Specifically, it selects target tissue-specific signature genes from multiple reference tissues. It aims to integrate informative and conserved signature genes as input of tranSig Bayesian model and removes the reference tissue-specific genes that have distinct expression distributions compared to those in the target tissue. Additionally, we adopt Combat [89][90] on bulk RNA-seq mixture and pseudo-bulk mixture derived from scRNA-seq to correct the batch effects between bulk RNA-seq and scRNA-seq. The final tranSig signature matrix and batch-effects-corrected bulk RNA-seq can be paired with NNLS or any other external cell type deconvolution tools, e.g., CIBERSORTx [70] and quadratic programming [74], as an add-on method.

To investigate the robustness of the tranSig model, we conducted a number of simulations under different conditions. Since it is challenging to simulate tissue- and platform-effects, the applications on simulations mimic the scenario after the initial two-steps batch-effect corrections of LIGER and Combat, and only examined the performances on signature matrix construction by the tranSig bayesian model. Simulations demonstrated more stability of tranSig than other methods (e.g., NNLS, CIBERSORTx, and MuSiC) for different numbers of tissues, signature genes, and non-zero expression undetected rates (Fig 2.2, Fig A.6). Notably, the robustness to undetected rates suggests that our approach can handle low data quality single cell datasets due to, for instance, low sequencing depth.

For the real application of tranSig to PB, we calculated the correlations between the estimated and "true" cell type proportions. Deconvolution with tranSig by CIBERSORTx was more accurate and stable. In addition, the effectiveness of LIGER and Combat was evaluated in this case. We deployed the tranSig pipeline on a BAL bulk data set. The tranSig could successfully identify the top two dominant cell types: alveolar macrophages and T cells. Although there was still a gap between the true and estimated cell type proportions of AMs, tranSig was the most accurate one. Unlike the first two applications, the deconvolution methods did not perform well on the aortic aneurysm data set which does not have the "true" cell type proportions estimated through sorting experiments. Therefore, we tried to validate our results indirectly by comparing the estimates between the normal aorta and aortic aneurysms. Our results showed that tranSig+CIBERSORTx could interpret the pathological changes of aneurysms, including SMC apoptosis and inflammatory infiltration.

Based on the simulation and real application results, tranSig+CIBERSORTx performed better than other methods, specifically when using scRNA-seq datasets with low sequencing depth to derive a signature matrix. In the tranSig framework, we mainly utilized the HCL datasets generated by Microwell-seq [91][92] with a low sequencing depth ($\sim 500$ detected genes in each cell) over $30$ main tissues, including rarely studied tissues. Thus, the tranSig framework takes advantage of the comprehensive tissue types and overcomes various common technical noises [93][94][95][96][97] of scRNA-seq, such as the bias caused by insufficient sequencing depth, low capture efficiency, high drop-out rate, or cell type misclassifications. Thus, the deconvolution results of tranSig are more robust than other benchmarking methods. Furthermore, we used the ATAA as the target single cell datasets and other tissues of HCL datasets as references in the real application of aortic aneurysms, which shows the robustness and effectiveness of tranSig in transferring cross-study and -platform information.

The tranSig framework requires raw counts in scRNA-seq datasets and TPM normalized data in bulk RNA-seq. Our Bayesian model allows unnormalized scRNA-seq data input because of our mixture Gaussian priors. The utilization on raw count matrix of scRNA-seq can also improve performance over TPM normalized matrix in terms of correlations between estimates and ground truth (Fig A.8).

We note that the primary goal of tranSig is to harmonize information from other tissues. We used Gaussian mixture prior distribution to model the tranSig signature matrix $\upsilon^0_{dk} * \gamma^0_{dk}$. In the tranSig model, $\upsilon^0_{dk}$ is not constrained by non-negativity, better representing the relative relationships between the signatures and signatures across cell types. As the signature matrix of the tranSig output, we calculate the average of $\upsilon^0_{dk} * \gamma^0_{dk}$ of the last iteration as the algorithm sampled multiple $\upsilon^0_{dk}$ and $\gamma^0_{dk}$ in each iteration (details can be found in online methods). Then $\gamma^0_{dk}$ is considered as the binary weight suggesting whether we take the genes as signatures, further describing the relationships between signatures and cell types. Therefore, we can consider tranSig as fine-tuning and better capturing the relationships between signatures and cell types and between signatures. It is able to perform cell type deconvolution on rarely-studied tissue when the matched single cell datasets are lacking.

We note that the appropriate tissue types should be carefully chosen while implementing the tranSig model. Tissues with similar cellular compositions may provide more information for the signature matrix. However, incorporating more tissues as input of the tranSig framework may be time-consuming. Furthermore, tranSig to cross-study datasets requires consistent cell type annotations. In the current implementation of tranSig, $\tau_v$ needs to be specified and a better choice of appropriate $v$ can further improve its performance.

In conclusion, tranSig is a novel Bayesian framework to better infer a signature matrix by leveraging cross-tissue or -study information. Deconvolution based on the signature matrix inferred by tranSig may lead to more accurate cell type proportion estimates and gain additional insights from analyzing bulk sample data. Coupled with HCL data, tranSig is applicable to the deconvolution of various tissues. In the broader scheme, our approach may be considered as transfer learning. Future directions can focus on how to better incorporate information by integrative analysis and design more plausible models to derive signature matrices.

## 2.4 Online Methods

### 2.4.1 Single Cell RNA-seq batch correction

To prepare for downstream analysis, we first obtained batch-corrected expression profiles $\widehat{E}_i$ by comparing every reference single cell dataset $i$ with the target one through LIGER [67]. We then derived the empirical signature matrix from the reference tissues expression profile $\widehat{E}_i$ by averaging the expression levels for each cell type to obtain $\widehat{W}^t$ where $t = 1, \cdots, T$, where$T$ is the number of reference datasets, and we denote the matrix from the target tissue as $\widehat{W}^0$. After comparing every reference signature matrix $\widehat{W}^{t_i}$ with the target signature matrix $\widehat{W}^0$ element-wise, we used k-means to group all the cell type-specific signature ratios $\frac{\widehat{W}^t_{ij}}{\widehat{W}^0_{ij}}$ into three groups. The groups with ratios close to $1$ or $0$ will be considered tissue-specific signatures. Therefore, in the reference dataset, the signatures whose ratio is in the middle range are assumed to share similar expression patterns with the target tissue and will be taken into the tranSig model.

### 2.4.2 tranSig Bayesian model

Given the bulk RNA-seq data of tissue $t = 0$ of $N$ subjects with $D$ signature genes, $Y^0 \in \mathbb{R}^{D\times N}$, we can do matrix factorization as the product of cell type-specific signature matrix $w^0 \in \mathbb{R}^{D\times K}$ and cell type proportion matrix $Z \in \mathbb{R}^{K\times N}$ where $K$ is the number of cell types. In general, the cell type deconvolution methods assume $Y^0 = w^0 \cdot Z$ and $w^0$ characterizes the differential signatures across cell types by averaging the expression levels of signature genes.

Additionally, $w^t_{dk}$ is the expected expression level of gene $d$ in cell type $k$ from tissue $t$, where $t = 0, 1, 2, \cdots, T$. We also add a sparsity on the signature gene matrix by introducing the Bernoulli variable $\gamma^0_{dk}$. The Gaussian mixture distribution exhibits the mixture heterogeneity of the signature matrix.

$$x^t_{dc_k} \sim N(w^t_{dk}, \frac{1}{\tau_{x_d}}) \tag{2.3}$$

$$w^t_{dk} \sim N(v^0_{dk} * \gamma^0_{dk}, \frac{1}{\tau_w}) \tag{2.4}$$

The terms in the Gaussian mixture distribution have the following prior specifications.

$$\gamma^0_{dk} \sim Ber(\pi) \tag{2.5}$$

$$v^0_{dk} \sim N(0, \frac{1}{\tau_v}) \tag{2.6}$$

The prior distributions for the other parameters follow the uninformative conjugate distributions.

The goal of tranSig Bayesian model is to estimate $v^0_{dk}$ and $\gamma^0_{dk}$ based on the single cell data from multiple tissues.

### 2.4.3 tranSig optimization algorithm

We employed SAME [69] to accelerate the algorithm and estimate the Gaussian mixture priors. Let $\theta_1 = \{w^0, w^1, \cdots, w^T\}$ and $\theta_2 = \{\frac{1}{\tau_e}, \frac{1}{\tau_x}, \frac{1}{\tau_w}, \pi, \alpha, v, \gamma\}$. We are primarily interested in the inference of the parameters $v$ and $\gamma$ in the tranSig matrix.

We set the strictly positive increasing integer sequence as $\Lambda(i) = i$. When the iteration $i = 0$, we first initiate $\theta_1^{(0)}$. As the iteration $i \geq 1$,

- For $k = 1, \cdots, \Lambda(i)$, sample $\theta_2^{(i)} \sim p(\theta_2 | y, \theta_1^{(i-1)})$
- Sample $\theta_1^{(i)} \sim q_{\Lambda(i)}(\theta_1 | y, \theta_2^{(i)}(1), \cdots, \theta_2^{(i)}(\Lambda(i)))$

Therefore, in the last iteration $M$ and $\Lambda(M) = M$, we have $v_{i,M}^0(1), \cdots, v_{i,M}^0(M)$ and $\gamma_{i,M}^0(1), \cdots, \gamma_{i,M}^0(M)$. We derive the tranSig matrix by averaging $v_{i,M}^0(i) * \gamma_{i,M}^0(i)$ across $i$. This average can enhance the confidence of tranSig estimation and each $\gamma_{i,M}^0(i)$ can be considered as the indicator whether $v_{i,M}^0(i)$ should be taken into account. By comparing with the estimation of $v_{dk}^0 * \gamma_{dk}^0$ in the last iteration, deconvolution results of averaged $v_{dk}^0 * \gamma_{dk}^0$ substantially improved and showed more consistent with the ground truth (Fig A.8).

### 2.4.4 Bulk RNA-seq batch correction

Due to the technical differences between UMI-based scRNA-seq (e.g. Microwell-seq and 10X genomics) and bulk RNA-seq, the deconvolution by the signature matrix derived from UMI-based single cell expression profiles is far from ideal. Therefore, we leverage Combat (an empirical Bayesian batch-effect remove model) and pseudo mixture constructed by single cell expression profiles to minimize the technical variation. We re-denote bulk RNA-seq $Y^{(t_0)}$ to be $Y^1$ as the first batch and denote a pseudo mixture to be $Y^2$ as the second batch. To adjust the raw bulk RNA-seq to the space of scRNA-seq by Combat, the EB model can be formulated as the following

$$Y^1_{gn} = \alpha_g + B^1_g + \delta^1_g \epsilon^1_{gn} \tag{2.7}$$

$$Y^2_{gn} = \alpha_g + B^2_g + \delta^2_g \epsilon^2_{gn} \tag{2.8}$$

where $\alpha_g$ is the overall gene expression, $B^i_g$ and $\delta^i_j$ are the batch- and gene-specific random effects. By using the parametric model of ComBat, we can get the adjusted bulk RNA-seq data as:

$$Y^{1*}_{gn} = \widehat{\alpha}_g + \widehat{B}^{2*}_g + \widehat{\sigma}_g \frac{\widehat{\delta}^{2*}_g}{\widehat{\delta}^{1*}_g}(Z^1_{gn} - \widehat{B}^{1*}_g) \tag{2.9}$$

where $Z^1_{gn}$ are the standardized data, $\widehat{\alpha}_g$ is estimated from ordinary least squares (OLS), $\widehat{B}^{i*}_g$ and $\widehat{\delta}^{2*}_g$ are estimated from EB as the first moment of their posterior distribution iteratively.

### 2.4.5 Pseudo-bulk construction by single cell expression profile

For bulk RNA-seq batch correction, we generated pseudo-bulk expression profiles based on single cell datasets. Single cell expression profiles were normalized in TPM space, and cells were sampled according to the empirical proportion of each cell type (10,000 times in total). The average of the 10,000 cells was considered as one pseudo-bulk sample (TPM space).

### 2.4.6 Simulation Setup

Although we do not assume the signature matrix continuous part $v^0$ to be non-negative, we still coerce $v^0_{dk}$ to be non-negative in simulation for convenience. We sample $v^0_{dk}$ from half-normal distribution and $\gamma^0_{dk}$ from Bernoulli distribution:

$$v^0_{dk} \sim |N(\mu_v, \frac{1}{\tau_v})| \tag{2.10}$$

$$\gamma^0_{dk} \sim Ber(\pi) \tag{2.11}$$

And we set $\tau_v = 4$, $\mu_v = 2$, and $\pi = 0.3$.

Per the model's assumptions, the intermediate dataset-specific signature gene matrix in the first layer $w^t_{dk}$ over all the tissues share the same Gaussian distribution with a mean of $v^0 \odot \gamma^0$. Therefore, we sample the elements in the matrix $w^t_{dk}$ from the following distribution:

$$w^t_{dk} \sim N(v^0_{dk} * \gamma^0_{dk}, \frac{1}{\tau_w}) * 1_{\{w^t_{dk} \geq 0.01\}} + 0.01 * 1_{\{w^t_{dk} < 0.01\}} \tag{2.12}$$

Here we set $\tau_w = 4$.

To further simulate single cell data from multiple tissues, we let $\tau^x_d$ to be sampled from a Gamma distribution. Then we have

$$\tau^x_d \sim Gamma(1, 1) \tag{2.13}$$

$$x^t_{dkc_k} \sim N(w^t_{dk}, \frac{1}{\tau^x_d})1_{x^t_{dkc_k} \geq 0} + 0.001 \times 1_{x^t_{dkc_k} < 0} \tag{2.14}$$

where $\tau^x_d \sim Gamma(1, 1)$.

Finally, we simulated cell type fraction matrix $z$ and bulk data $Y^0$. We sampled $z$ from uniform distribution to make the summation be $1$. Then we have

$$z_{kn} \sim \text{Unif}(0, 1 - \Sigma_{j<k} z_{jn}) \tag{2.15}$$

$$Y^0 = (v^0 \odot \gamma^0) * Z \tag{2.16}$$

where $v^0 \odot \gamma^0$ is the tissue-specific tranSig matrix when given a target tissue $t = 0$.

### 2.4.7 Benchmarking

We compared the results of the tranSig model with four deconvolution methods, including NNLS, quadratic programming (QP), CIBERSORTx, and MuSiC. NNLS and QP take the average expressions of signature genes in each cell type as input, and CIBERSORTx and MuSiC take single cell expression profiles as input. For CIBERSORTx, S mode and B mode were used to correct the technical variation between scRNA-seq and bulk RNA-seq. CIBERSORTx deconvolution was implemented with default parameters, and single cell references were normalized in TPM space as the CIBERSORTx input. MuSiC needs a number of subjects to select signature genes, thus "sample" in HCL metadata was used as the MuSiC parameter. However, because there is only one sample in the artery dataset of HCL, MuSiC cannot be implemented in the aneurysm application.

### 2.4.8 Data processing

We utilized the published HCL with $60$ tissue types as the single cell references to perform deconvolution on all tissues. We manually cleaned the cell types in the most common 30 tissues from HCL, including $25$ adult tissues, four fetal tissues, and CB. For instance, we annotated all macrophage subtypes as macrophages to accommodate different annotation resolutions in the $30$ tissue-specific single cell datasets. We also removed the cells that did not have a precise annotation, i.e. unknown cell clusters and distal cells in Lung. Overall, there are 96 distinct cell types across all $30$ tissues

To further validate the significance of the tranSig model for biomedical investigations, we used bulk RNA-seq of ascending aorta media from healthy donors and aneurysm tissues [73] from ascending aneurysm patients and compared the cell type proportions corresponding to aneurysm pathological changes between two groups. To evaluate the effect of the sequencing depth and single cell platform, we used the single cell dataset from three healthy donors and eight ascending thoracic aortic aneurysm (ATAA) patients by 10X genomics. The dataset includes main cell types in the aorta, i.e., endothelial cells, smooth muscle cells, fibroblasts, macrophages, T cells, B cells, NK cells, mast cells, and plasma cells.

For real applications, bulk RNA-seq of the whole blood from $12$ healthy adults from the Stanford Blood Center [70] with group truth validated by FACS and immunofluorescence were used.

The BAL bulk RNA-seq data [75] were obtained from Vukmirovic M. et al. (2021). The samples were collected from $184$ individuals in a sarcoidosis patients cohort by the Genomic Research in Alpha-1 Antitrypsin Deficiency and Sarcoidosis (GRADS) study [98]. BAL is composed of four cell populations: alveolar macrophages, eosinophils, lymphocytes, and neutrophils. The mean proportion of alveolar macrophages is $89\%$. Therefore, a reliable cell type deconvolution tool should identify the alveolar macrophages as the dominant cell type in bulk RNA-seq data.

The aneurysm bulk RNA-seq data [78] were from Chen et al. (2020). The samples included aortic tissues from six healthy donors and aneurysm tissues from six patients with ascending aortic aneurysm. For aortic tissues, each donor had two samples on the middle and distal parts. In terms of aneurysms, the neck and belly were collected for each patient.

All the bulk datasets were normalized by TPM.

### 2.4.9 Signature Gene List Construction

Based on the HCL datasets, significantly highly expressed genes of each cell type from the five tissues (artery, lung, PBMC, BM, and liver) used in the tranSig model were calculated by the “FindAllMarkers” function of Seurat R package. We set $0.25$ as the cutoff for logarithmic fold changes, and the positive fold changes are selected by the parameter “only.pos” of “FindAllMarker” function. We used the union of these DEGs as the signature gene list for the tranSig model (Supplementary E.2) which covers most of the immune cells. However, we recommend generating a customized signature gene list by finding DE genes when the immune cells are not the only major cell populations in the target tissue.

### 2.4.10 Data Availability

All the data could be downloaded from the GEO database (`http://www.ncbi.nlm.nih.gov/geo/`). The HCL can be accessed under the GEO accession number GSE134355. It could be obtained at `http://bis.zju.edu.cn/HCL/` or `https://db.cngb.org/HCL/`. The PB bulk data are available through GEO with accession number GSE127472. The GEO accession number of the BAL dataset is GSE109516. The ATAA dataset can be accessed at GSE155468 and aorta bulk data at GSE140947.

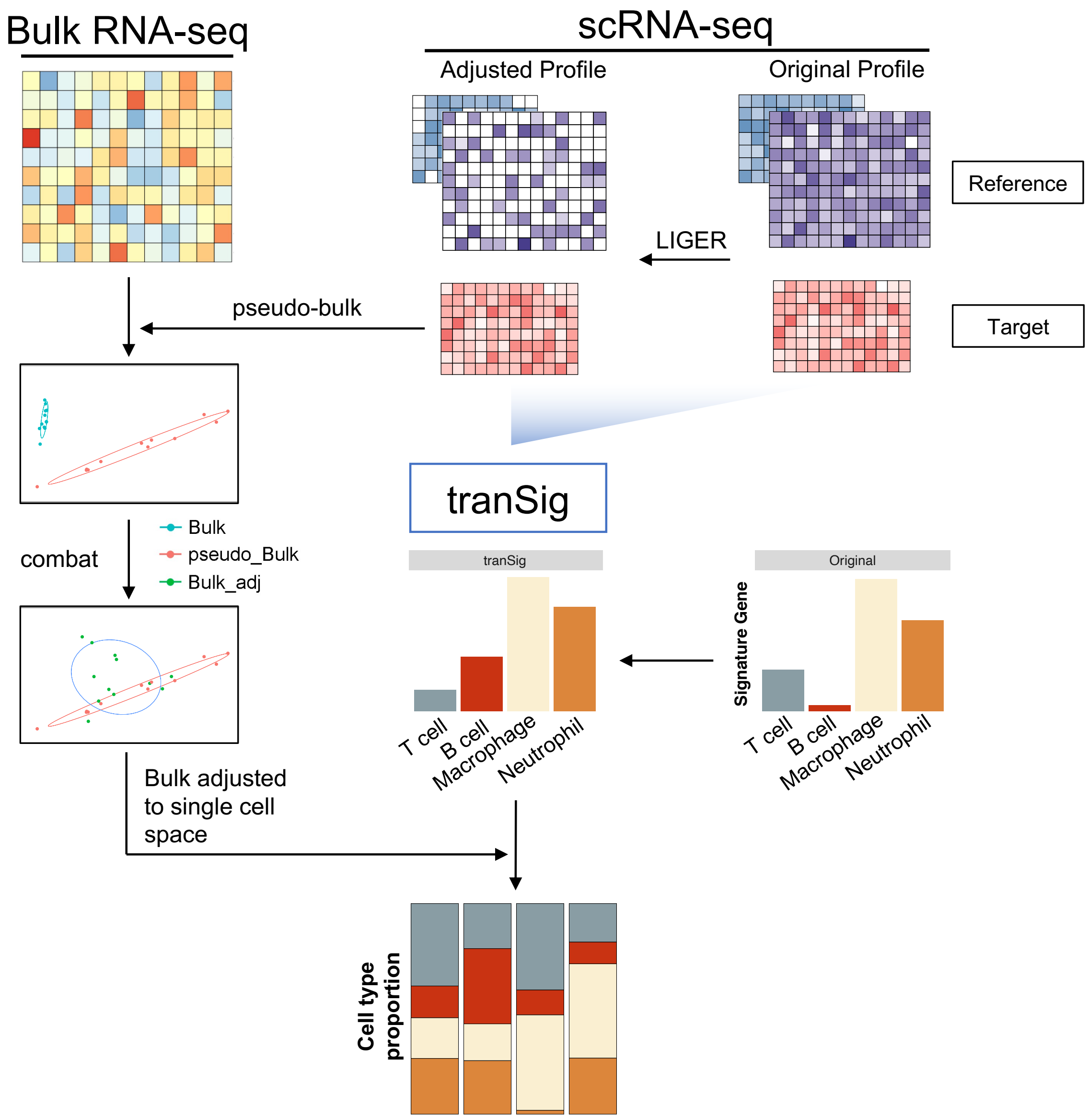


Figure 2.1: Illustration of the tranSig framework. The framework starts from selecting signatures from reference single cell datasets for downstream harmonization. Based on LIGER, we filter out the genes in reference tissues which are unlikely to share common expression distributions with the target tissue. After that, all the selected signatures and their corresponding single cell expression profiles are input into the tranSig Bayesian model to derive a more reliable signature matrix. On the other hand, we remove batch-effects between the bulk and target single cell datasets based on Combat to project the bulk data onto the space of target single cell dataset. Finally, the tranSig signature matrix and the corrected bulk RNA-seq can be coupled with other cell type deconvolution optimization tools, i.e. NNLS and CIBERSORTx.

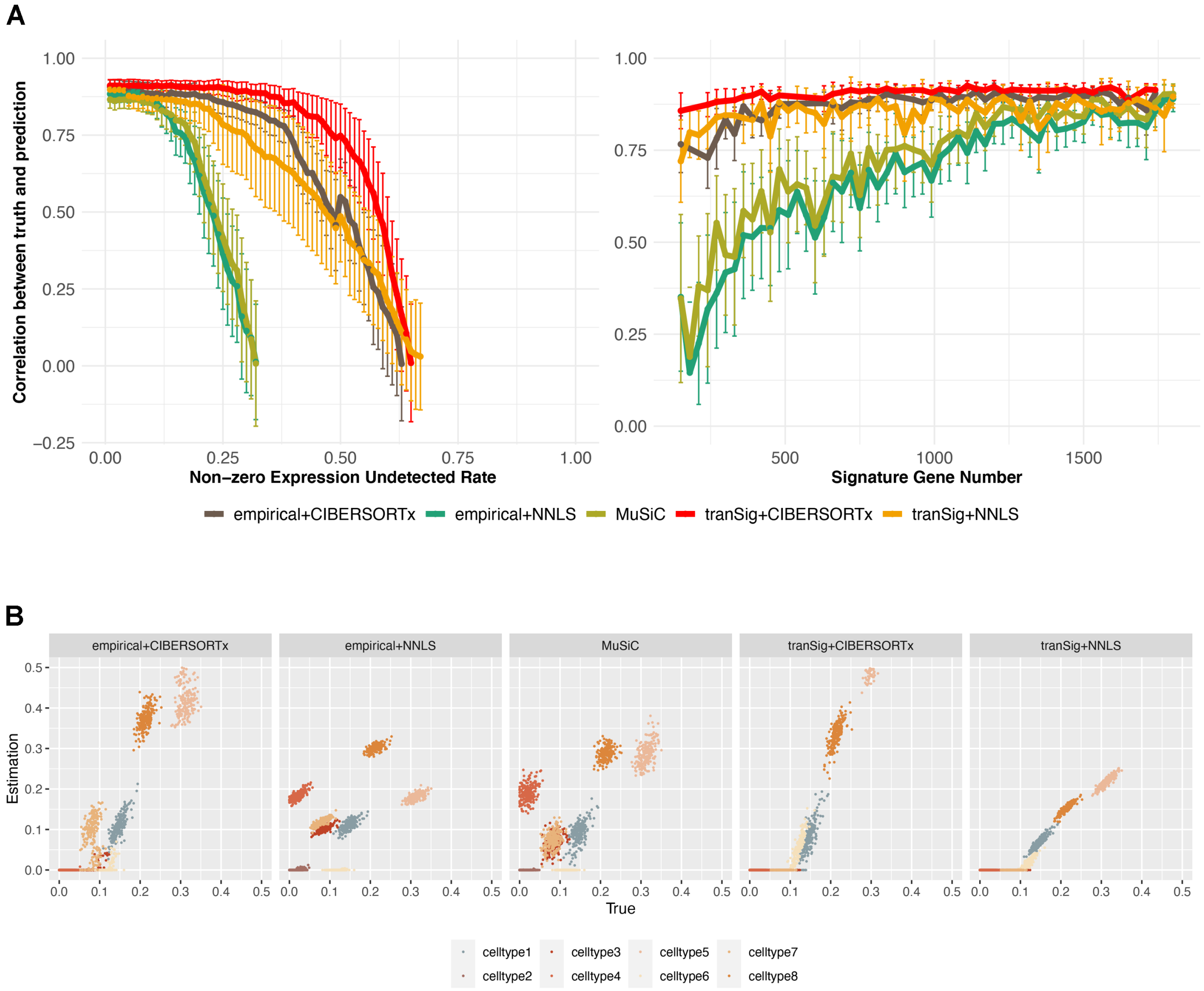


Figure 2.2: Model robustness validations on simulations. A) The curves to evaluate the robustness of tranSig signature matrix against non-zero expression undetected rate (left) and signature gene number (right). The colors code the combinations of signature matrix and deconvolution tools. The vertical lines are error bar of mean $\pm 0.5^{*}$s.d. B) The scatterplots to show the benchmarking comparisons between the true and estimated labels. The x- and y-axis are the true and estimated cell type proportions. The colors indicate the cell type labels.

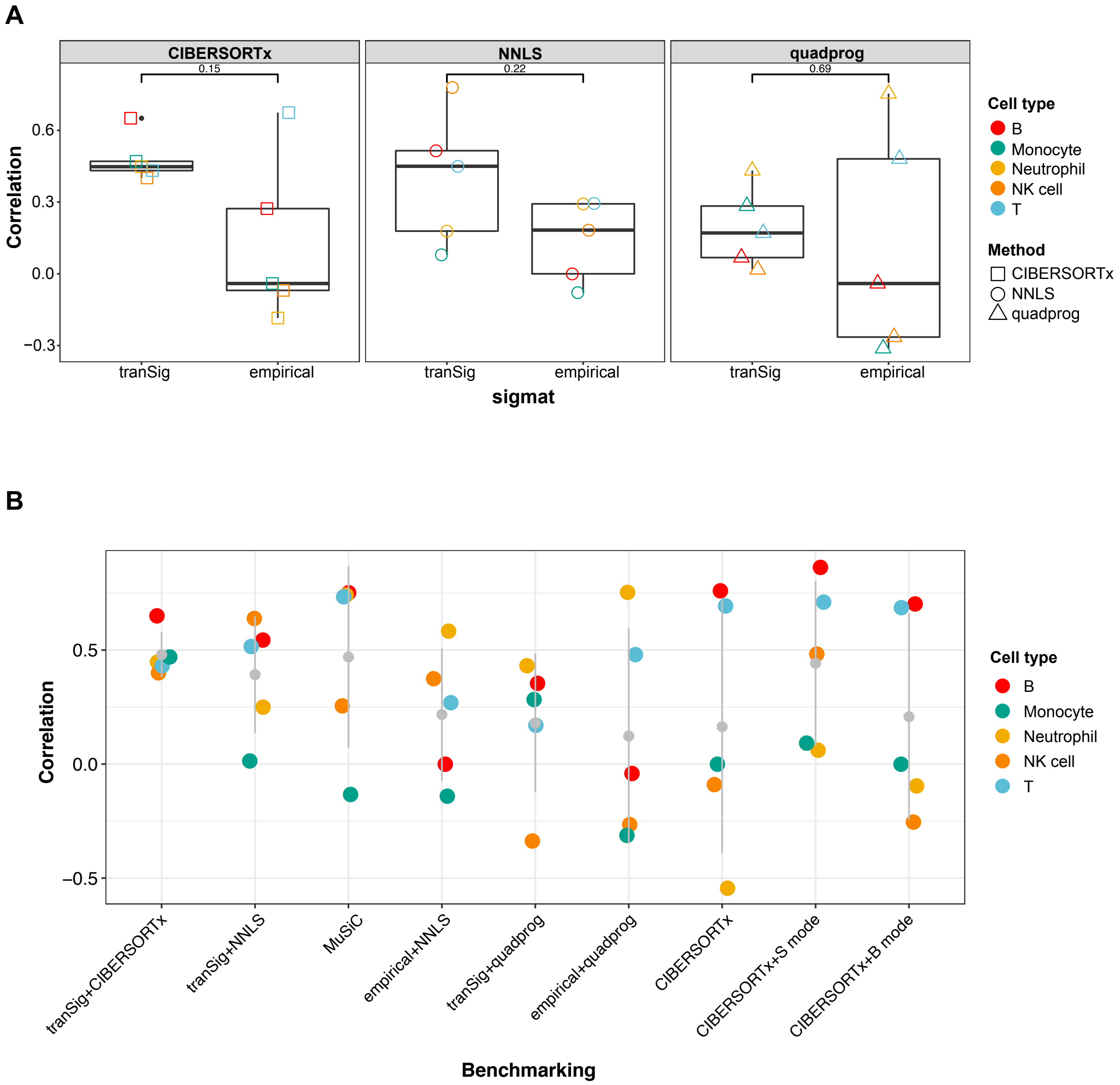


Figure 2.3: PB bulk data cell type deconvolution benchmarking analysis. A) Box plots of the correlations between estimated cell type proportions and the ground truth for Newman et al. blood samples ($n = 12$), with color coded by cell types. CIBERSORTx is denoted as square, NNLS is denoted as circles, and quadratic programming (quadprog) is denoted as triangles. Statistical significance is calculated by the Wilcoxon test. Data are presented as medians ± interquartile range. B) Benchmarking of deconvolution methods shown on jitter plots of correlations same as (A). Data are expressed as means ± s.d.

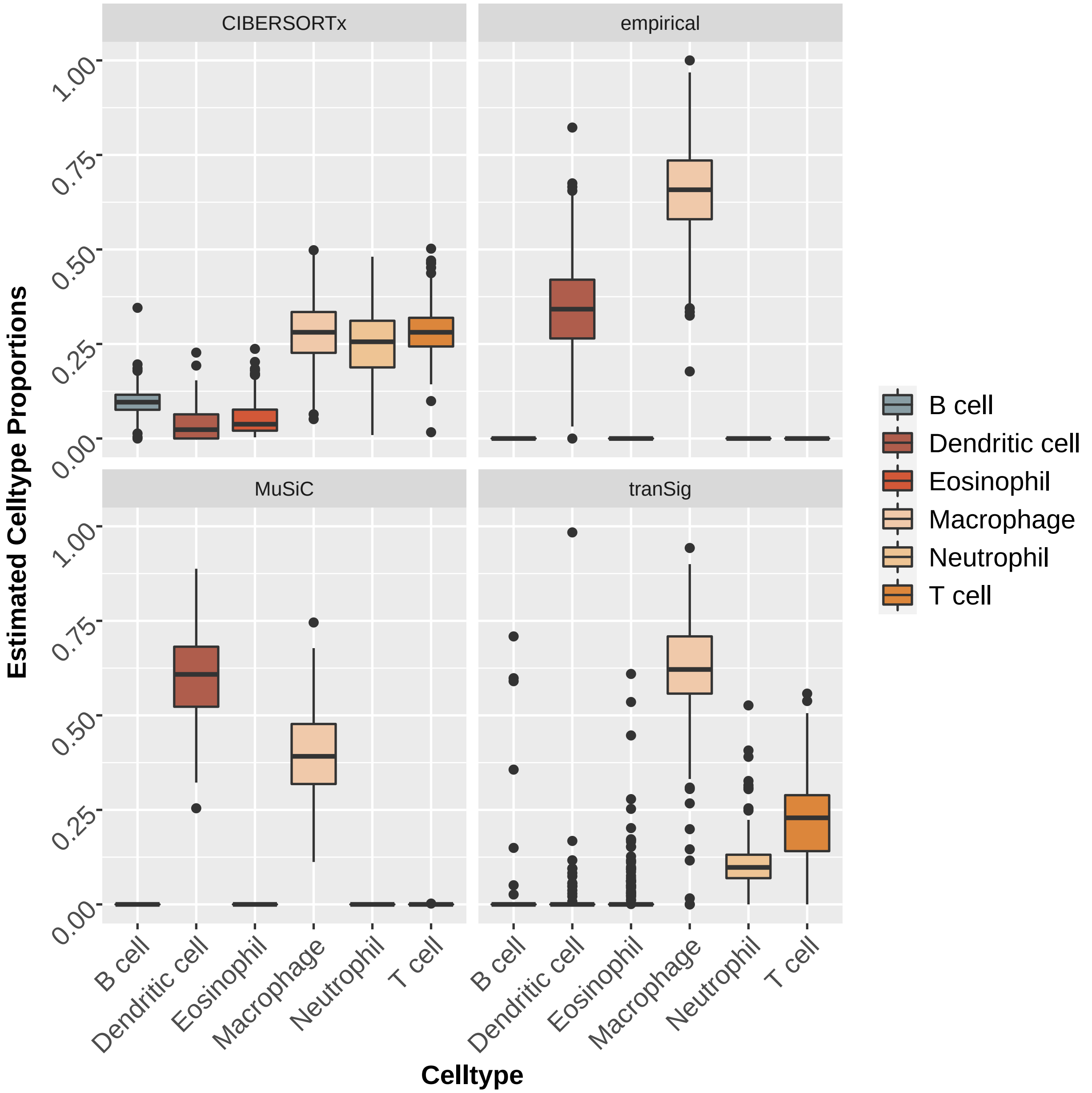


Figure 2.4: BAL bulk data cell type deconvolution and tranSig identifies macrophages as the dominant cell type. The boxplots of cell type proportions across cell types. The x-axis represents the cell types, coded by the colors. The y-axis represents the estimated cell type proportions.

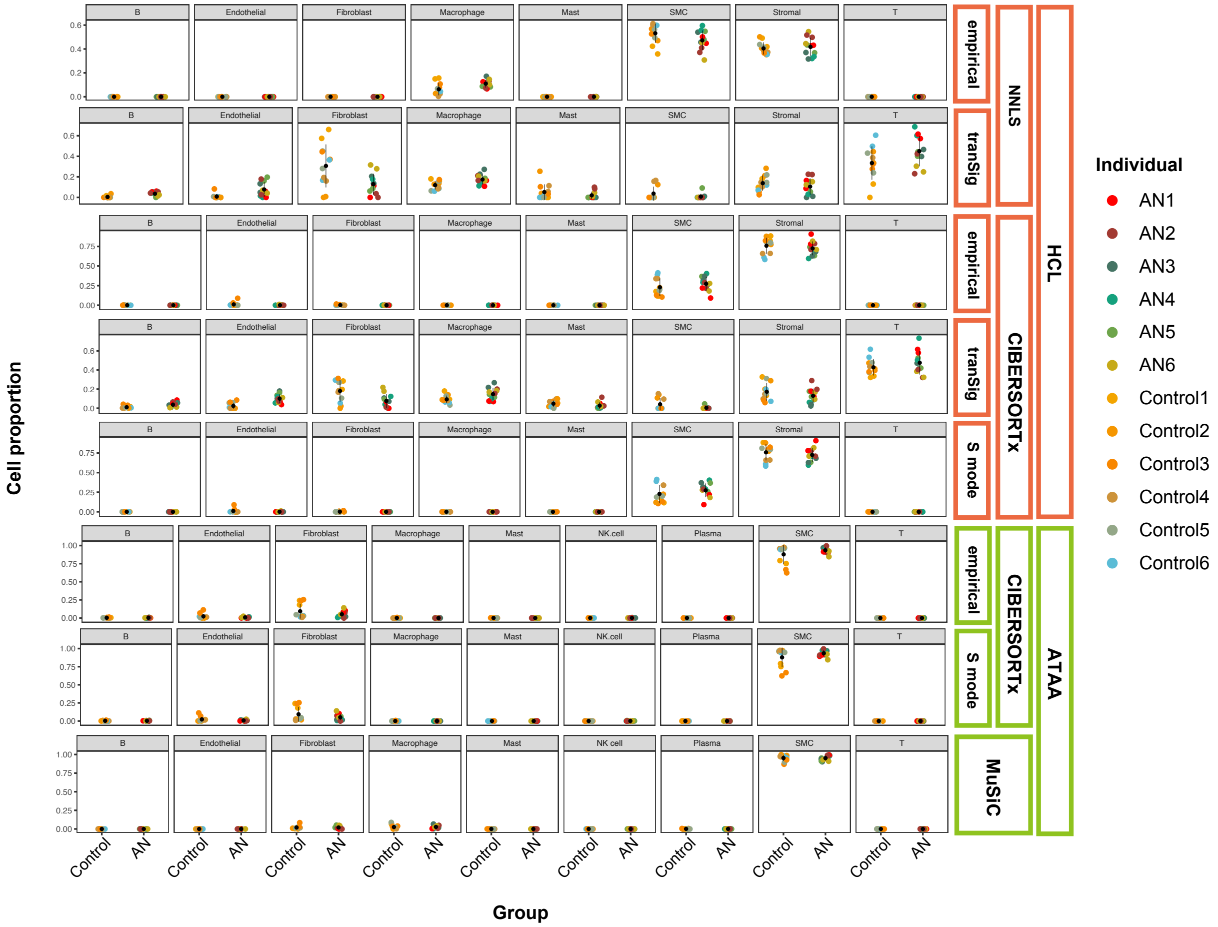


Figure 2.5: Applications on aorta to depict cellular pathological changes of aneurysm (AN). Jitter plots of estimated cell type proportions for Chen et al. subjects (aortic tissues from 6 health donors as control and aneurysm tissues from patients with ascending aortic aneurysm as AN), color coded by individuals. Color bars on the right annotate the deconvolution methods and single cell references, the artery dataset of HCL in red and the ATAA dataset in green. Data are expressed as means ±s.d.

# Chapter 3

# Polygenic Risk Scores Incorporating Single Cell RNA Sequencing Data Identified Cell and Tissue Types with Genetically Regulated Transcriptions Associated with Complex traits

## Abstract

Genome wide association studies (GWAS) have identified numerous regions in the human genome associated with thousands of complex traits. Polygenic risk scores (PRS) have been developed for many traits based on GWAS results to identify individuals with high disease risk. Most published PRS are based on statistical evidence of association derived from GWAS without leveraging the rich functional information about the markers from either experimental or computational studies. In this chapter, we describe our efforts to use information offered by the single-cell RNA-sequencing (scRNA-seq) data to construct PRS to help us identify tissues and cell types most relevant to complex traits, phenotypes that are genetically correlated through specific tissues and cell types, as well as subtypes among the patients. Our basic approach is to decompose the overall PRS based on cell type and tissue annotations on SNPs using differential expression analysis results from scRNA-seq data. The decomposed PRSs allow us to identify the cell types and tissues with strong associations to complex traits through expression regulations. Our method successfully identified T cells in the brain as relevant to Alzheimer's disease (AD), and also implicated cell types and tissues important for lipids, coronary artery disease (CAD), and artery tissue.

Keywords: scRNA-seq; Decomposed PRS; Genetics; Data Integration.

## 3.1 Introduction

Genome wide association studies (GWAS) have been very successful to identify associations between single nucleotide polymorphisms (SNPs) and complex traits[99, 100, 101]. Based on GWAS results, polygenic risk scores aggregate the risk allele effect sizes across the whole genome to predict disease risk and has the potential to improve precision medicine. Many statistical and machine learning methods have been developed, most using only summary statistics due to their easy accessibility and the retaining of much information for risk prediction. One commonly used tool is P+T (pruning+thresholding)[102]. It clumps SNPs to remove the effects of linkage disequilibrium (LD) in the selections of markers for PRS. Another method, PRS-CS, adopts a Bayesian model to accommodate complex genetic architecture and LD [103]. Incorporating various types of annotations can further improve PRS performance, with AnnoPred as a representative approach that integrates a diversity of functional and tissue-specific annotations[104].

As an individual-level proxy to genetically predicted phenotypes, PRS has many downstream applications. One potential use is patient subtyping[105, 106, 107]. However, the overall PRS cannot address disease subtypes as there is only one score for each individual, and we need to decompose the overall PRS into different components that may correspond to different subtypes. Some methods have been proposed to derive the components from singular value decomposition

(SVD) of the PRS across a large number of traits pleiotropic relationships or using pathway information [108, 109]. Based on the partitioned PRS scores, individuals with high PRSs can be partitioned into different groups based on the decomposed PRSs.

The rapid development of scRNA-seq technologies also offers another rich data source to study associations between genes and phenotypes. Many computational methods have been proposed for classification, differential expression (DE) analysis, trajectory inference, and gene regulation network based on scRNA-seq data [110, 111, 112, 113]. Efforts have been made to integrate scRNA-seq data with GWAS data to better understand genetic signals at the cell type and tissue level. For example, a lipids study conducted DE analysis on scRNA-seq and linkage disequilibrium score regression (LDSC) to associate cell types with traits[114]. Cell-type Wide Association Study (cWAS) is a method to learn how genetically-regulated cell type proportions can impact disease risk through integrated analysis of bulk RNA-seq data, scRNA-seq data, and GWAS data [115].

In this chapter, we attempt to explain PRS based on scRNA-seq data with the potential for patient subtyping. We propose a novel data integrative analysis framework to derive the cell type- or tissue-specific decomposed PRSs. The decomposed PRSs can inform us how cell types and tissues contribute to PRSs through gene expressions. In our method, we first perform DE analysis on scRNA-seq data and translate the summary statistics of DE analysis to gene-level weight scores. After mapping SNPs to genes, we derive the SNP-level weight scores through a Bayesian statistical model. Then, the decomposed PRSs are calculated based on the SNP-level weight scores and risk allele effect sizes estimated from GWAS. We have applied our framework to multiple GWAS datasets and the scRNA-seq data from the human cell landscape [63]. We were able to identify relevant tissues and cell types for lipids, heart disease, and other traits. The decomposed PRSs will be applied for patient subtyping in the future.

## 3.2 Methods

### 3.2.1 Overview

We start by describing the assumptions of cell type-specific and tissue-specific decomposed PRS, coupled with annotations based on DE analysis for scRNA-seq data. In general, the transcriptome-derived overall PRS, which is only based on the SNPs located within gene regions, can be decomposed into cell types-specific or tissue-specific PRSs by their DE gene sets. The method is based on the hypothesis that this transcriptome-derived overall PRS is predictable of complex traits since gene expression is downstream of genetic risk variants and may partly explain the phenotype heterogeneity in the population. Several studies have shown that a significant component of genetic effects on phenotypes is mediated through expression and identified the relevant tissues and cell types to the complex traits[116, 117, 118, 119, 120, 121, 122]. The genes differentially expressed across tissues and cell types are informative about their relevance for diseases. In our framework, we perform DE analysis on scRNA-seq data across cell types or tissues, which here we briefly denoted as groups for simplicity. We consider both the SNPs which are located within the gene region and SNPs in a neighborhood of 500k base pairs (bp) around the gene but chose to only present the former one. This is because we found that considering SNPs within the 500 kbp region resulted in the large overlap across groups (cell types or tissues), and led to similar and uninformative values of decomposed PRSs. The workflow of our framework is presented in Fig 3.1.

We first conducted DE analysis by marginally regressing gene expressions on the group indicator at the single cell level. The summary statistics from the linear regression DE analysis includes cell type- or tissue-specific effect sizes $\hat{\beta}_{ij}$ and the corresponding standard errors $se(\hat{\beta}_{ij})$ for gene $i$ and group $j$. To minimize the batch effects and other unwanted variations across tissues, we use the single cell data from the Human Cell Landscape (HCL)[63], which is an atlas for single cell comparative analysis.

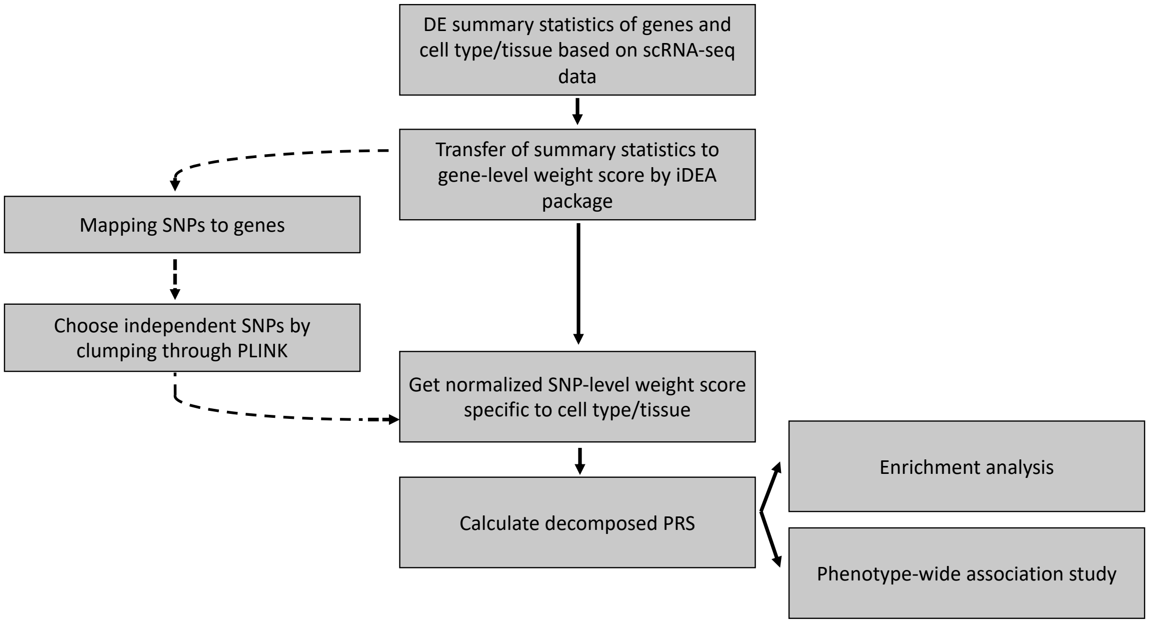

DE summary statistics of genes and cell type/tissue based on scRNA-seq data
Transfer of summary statistics to gene-level weight score by iDEA package
Mapping SNPs to genes
Choose independent SNPs by clumping through PLINK
Get normalized SNP-level weight score specific to cell type/tissue
Calculate decomposed PRS
Enrichment analysis
Phenotype-wide association study

Figure 3.1: Overview of our method. We first get the DE genes from the scRNA-seq data and use the iDEA package to get the gene-level weight scores. To translate gene-level to SNP-level weight scores, we map SNPs to genes by chromosomal locations, select independent SNPs and normalize the scores across cell types/tissues. The final decomposed PRSs are then used for downstream analysis, including enrichment analysis and PheWAS.

These DE summary statistics were then transferred to the gene-level weight score, $\pi_{ij} \in [0, 1]$ to quantify the specificity of gene $i$ to group $j$. We assumed that the observed effect size $\hat{\beta}_{ij}$ follows a Gaussian distribution with the true effect size $\beta_{ij}$ as the mean. To evaluate the DE specificity of gene $i$ to cell type $j$ or tissue $j$, we further implemented R package iDEA, which is a Bayesian framework to use gene set information to improve DE analysis results [123]. More specifically, iDEA assumes the true effect size following a mixture distributions as follows

$$\beta_{ij} \sim \Pr(\gamma_{ij} = 1)\mathrm{N}(0, se(\hat{\beta}_{ij})\dot{\sigma}_{\beta}^2) + \Pr(\gamma_{ij} = 0)\sigma_0 \tag{3.1}$$

We defined the gene-level weight score as $\pi_{ij} = \Pr(\gamma_{ij} = 1)$ as the prior probability of gene $i$ to be a DE gene for cell type $j$ or tissue $j$. The iDEA package further boosts DE analysis by bundling the gene set as the following:

$$\begin{aligned} \text{logit}(\Pr(\gamma_{ij} = 1)) &= \text{logit}(\pi_{ij}) \\ &= \log \frac{\pi_{ij}}{1 - \pi_{ij}} \\ &= \tau_0 + a_{ij}\tau_1 \end{aligned} \tag{3.2}$$

When $a_{ij} = 1$, gene $i$ belongs to the DE gene set of group $j$. Any scRNA-seq DE methods can be applied to get the DE gene sets, including t-test, MAST and DESeq2[124, 125].

The posterior inclusion probability (PIP) on gene-level weight score $\pi_{ij}$ was then mapped to the SNP-level weight score via genomic location. We then selected independent SNPs by doing clumping with PLINK version 1.90[126].

For SNP k, we assume that a total of $j_K$ genes overlap with this SNP and these genes are denoted as $i_1, i_2, \ldots, i_K$ with corresponding lengths $t_{i_1}, t_{i_2}, \ldots, t_{i_K}$, respectively. We define the normalized SNP-level weight score of the $k$th SNP for group $j$ as

$$b_k^j = \frac{\frac{\pi_{i_1 j}}{t_{i_1}} + \frac{\pi_{i_2 j}}{t_{2_1}} + \cdots + \frac{\pi_{i_k j}}{t_{i_k}}}{\Sigma_p \frac{\pi_{i_1 p}}{t_{i_1}} + \frac{\pi_{i_2 p}}{t_{2_1}} + \cdots + \frac{\pi_{i_k p}}{t_{i_k}}} \tag{3.3}$$

where P is summed over all the groups. Therefore, the sum of the normalized SNP-level weight scores, $b_k^j \in [0, 1]$, across all groups equals one. The weight scores , $b_k$, then were used to get the decomposed PRS for group $j$. More specifically, for cell type $j$ or tissue $j$, the corresponding $\text{PRS}^j = \Sigma_k x_k \beta_k b_k^j$, where $x_k$ is the $k$th SNP genotype and $\beta_k$ is the risk allele effect size for SNP $k$ estimated from GWAS. We adopted the P+T method for PRS calculation [103] by clumping SNPs whose p-values were below 0.05 and $R^2$ was greater than 0.8.

We studied the performance of our cell type-specific and tissue-specific decomposed PRSs by enrichment analysis and phenome-wide association study (PheWAS). We considered eight traits: coronary artery disease (CAD), atrial fibrillation (AF), type 2 diabetes (T2D), breast cancer (BC), HDL-cholesterol (HDL), LDL-cholesterol (LDL), total cholesterol (TC), and triglyceride (TG). We also conducted an additional cell type analysis for AD, which is more distinguishable at the cell type level.

### 3.2.2 Enrichment Analysis

To evaluate the performance of the new framework and identify the gene expression-mediated cell types and tissues, we performed enrichment analysis on nine common traits: AF, T2D, BC, HDL, LDL, TC, TG, and AD.

We first divided the SNPs by setting a score cutoff for group-specific annotations. Since there is no gold standard for the scoring setup and the enrichment analysis results may not be robust to the cutoffs, we varied the cutoff values from $0.1, 0.2$, to $.9$. The enrichment was calculated through the ratio of group-specific heritability over the overall heritability obtained by the stratified LD score regression [127]. More specifically, the enrichment ratio is defined as

$$\text{Enrichment} = \frac{\text{total heritability based on the group-specific SNPs}}{\text{overall annotated SNP coverage}} \tag{3.4}$$

For convenience, we plot heatmaps across cutoff scores and groups, visualize the negative log p-values, and highlight the significant p-values (p-val $< 0.05$) by asterisks. Cell types and tissues having consistent significant p-values across different cutoff scores are considered the relevant groups to the complex traits.

### 3.2.3 PheWAS

It is well known that many phenotypes have shared biological pathways. We clustered the decomposed PRSs based on their correlations with the phenotypes by PheWAS across a multitude of traits and diseases. Here, we considered 55 traits and manually clustered them into 14 categories, including basic, cardiovascular diseases (CVD), family history, immune system, lifestyle, lipids, liver, metabolism, obesity, protein, respiratory, sex hormone, T2D, and urinary system. The PheWAS can reveal the similarity among the cell types or tissues due to their similar underlying biological mechanisms. While comparing with the total PRS (the sum of the decomposed PRSs), we can also easily identify the dominant group which is the driving risk factor for the traits. The other advantage is that PheWAS can select a more predictable PRS as a better proxy to a trait, if there is one, instead of the regular total PRS. It is possible that the aggregation of the decomposed predictors sometimes is not as competitive as the decomposed ones, which are highly relevant to the traits.

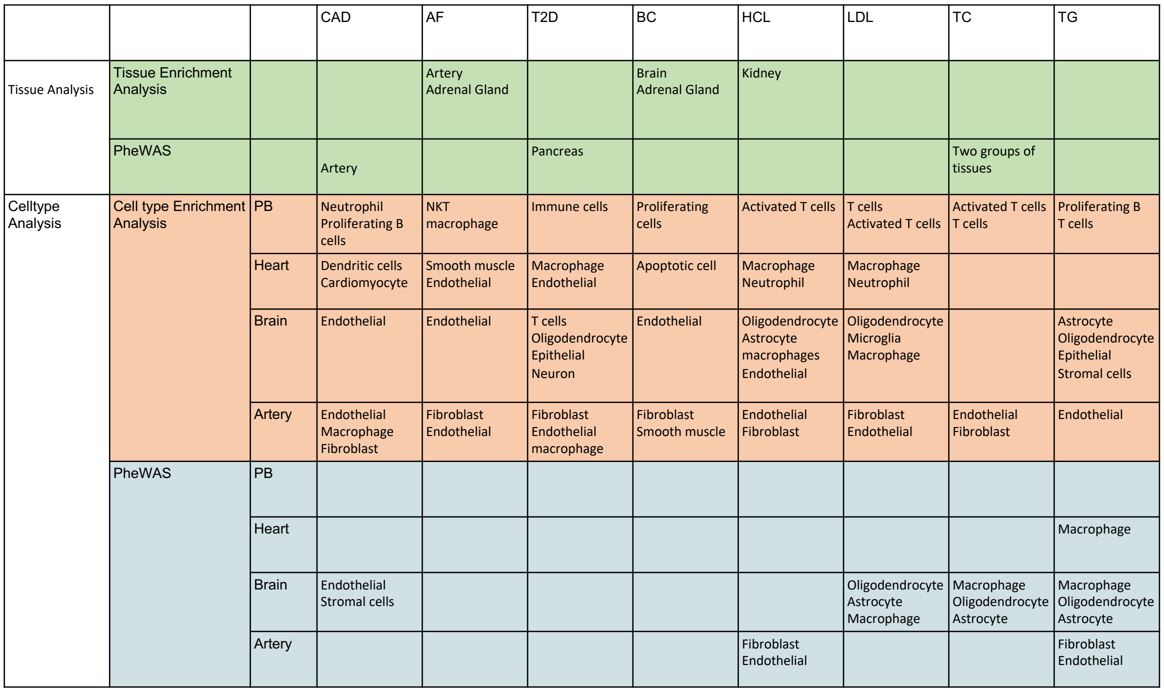
## 3.3 Results

Our proposed framework was applied for both cell type- and tissue-level analysis.

For tissue decomposed PRS analysis, we selected 26 tissues in total from HCL to cover the major parts of the human body and 13 of these tissues as a subset that is considered to have weak correlations with the phenotypes. Specifically, we focus on the results of 13 tissues since the normalized SNP-level weight scores are more reliable with these 13 tissues less correlated therefore better separations of their contributions. We conducted both tissue enrichment analysis and PheWAS on traits to evaluate our method as well as for novel biological findings.

In addition, we also performed enrichment analysis and PheWAS at the cell type-level. Specifically, we investigated the cell type decomposed PRSs from four common tissues that are critical for many complex traits, including peripheral blood (PB), heart, brain, and artery.

All the significant findings are summarized in Table 3.1. Eight complex traits are shown in columns and the rows are grouped by tissue enrichment analysis, tissue PheWAS, cell type enrichment analysis, and cell type PheWAS. The green color highlights the two tissue analyses, while the



| | | | CAD | AF | T2D | BC | HCL | LDL | TC | TG |
|---|---|---|---|---|---|---|---|---|---|---|
| Tissue Analysis | Tissue Enrichment Analysis | | | Artery<br>Adrenal Gland | | Brain<br>Adrenal Gland | Kidney | | | |
| | PheWAS | | Artery | | Pancreas | | | | Two groups of tissues | |
| Celltype Analysis | Cell type Enrichment Analysis | PB | Neutrophil<br>Proliferating B cells | NKT<br>macrophage | Immune cells | Proliferating cells | Activated T cells | T cells<br>Activated T cells | Activated T cells<br>T cells | Proliferating B<br>T cells |
| | | Heart | Dendritic cells<br>Cardiomyocyte | Smooth muscle<br>Endothelial | Macrophage<br>Endothelial | Apoptotic cell | Macrophage<br>Neutrophil | Macrophage<br>Neutrophil | | |
| | | Brain | Endothelial | Endothelial | T cells<br>Oligodendrocyte<br>Epithelial<br>Neuron | Endothelial | Oligodendrocyte<br>Astrocyte<br>macrophages<br>Endothelial | Oligodendrocyte<br>Microglia<br>Macrophage | | Astrocyte<br>Oligodendrocyte<br>Epithelial<br>Stromal cells |
| | | Artery | Endothelial<br>Macrophage<br>Fibroblast | Fibroblast<br>Endothelial | Fibroblast<br>Endothelial<br>macrophage | Fibroblast<br>Smooth muscle | Endothelial<br>Fibroblast | Fibroblast<br>Endothelial | Endothelial<br>Fibroblast | Endothelial |
| | PheWAS | PB | | | | | | | | |
| | | Heart | | | | | | | | Macrophage |
| | | Brain | Endothelial<br>Stromal cells | | | | | Oligodendrocyte<br>Astrocyte<br>Macrophage | Macrophage<br>Oligodendrocyte<br>Astrocyte | Macrophage<br>Oligodendrocyte<br>Astrocyte |
| | | Artery | | | | | Fibroblast<br>Endothelial | | | Fibroblast<br>Endothelial |

Table 3.1: Results Summary.

orange and blue colors highlight the cell type enrichment analysis and PheWAS in four common human tissues. The cell types or tissues are listed in the cells if they were found to be significant.

### 3.3.1 Enrichment analysis results

The cell type enrichment analysis in the brain identified that the T cell-specific SNPs are enriched in AD based on the GWAS results from the International Genomics of Alzheimer's Project (IGAP)[128]. The negative log p-values across different cutoff scores are shown in the left panel of Fig 3.2. It has been demonstrated in many studies that T cell plays an essential role in AD etiology [129]. The AD patients have a stronger inflammatory response to antigens than healthy individuals [130]. Notably, T cells are significantly more active and infiltrate in the brain in AD patients due to the damage[131]. Besides T cells, the other immune cells, such as B cells and neutrophils, also present enrichment at several cutoff scores, which is also consistent with previous findings on the immune cells in AD patients[132, 133, 134].

The tissue enrichment analysis also illustrates that our framework can discover tissues highly associated with the traits. In the middle panel of Fig 3.2, there is evidence of enrichment of kidneys for HDL. Kidney regulates the amount and distribution of the HDL particles by HDL metabolism, specifically, apolipoproteins and enzymes[135]. Previous studies have found that the reduction of HDL cholesterol concentration is highly related to Chronic kidney disease (CKD)[136]. Although we did not find the enrichment of heart to AF heritability and have a few NAs due to limited number of SNPs and computational singularity when we increased the cutoff, we did observe significant and consistent p-values in the artery and adrenal gland. The most common cause of AF is the damage of heart and artery. Increasing evidence shows the prevalence of AF in CAD and the high correlation between them[137, 138, 139, 140]. On the other hand, many groups have studied the association between AF and endocrine disorder[141, 142], where adrenal disorder has been found to increase the AF risk[143, 144].

In summary, the significant findings suggest that the proposed framework can discover the relevant cell types or tissues.

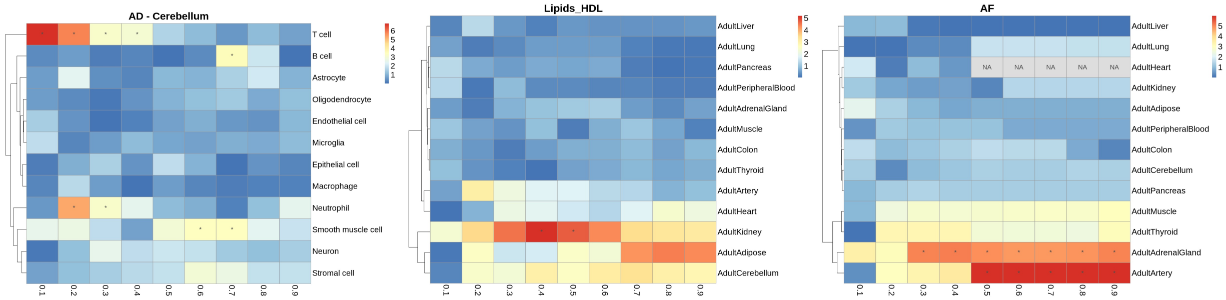


Figure 3.2: Enrichment analysis results. Results for three traits, AD, HDL, and AF, are shown in this figure. The rows are cell types/tissues and the columns are SNP-level weight score cutoff on selecting group-specific SNP set. The logrithm of negative p-values are encoded by colors. The asteroid symbols are pinned on the cell when p-val$< 0.05$.

### 3.3.2 Cell type important for different traits

One benefit of group-specific decomposed PRS is the identification of the groups (cell types or tissues) which make significant contributions to the overall PRS through gene expression.

#### Lipids

We conducted cell type enrichment analyses for the four lipids traits in four tissues, including PB, heart, brain, and artery. The four lipid traits exhibit very similar cell type patterns in all tissues we studied. Fig 3.3 shows three of them, and we will integrate the results of an artery in the following section.

The lipids traits have significant and consistent enrichment in several T cells shown in Fig 3.3A. HDL, LDL, and TC have coherent, strong cell-type-specific heritability in T cells and activated T cells. T lymphocytes can control the lipids metabolism in blood in several studies[145, 146, 147]. It is essential to learn the underlying mechanism of how T lymphocytes can manage the lipids. For instance, fatty acid (FA), a component in the biological metabolism process and the precursors of lipids complex, plays a role in the differentiation and mortality of T cells by deriving energy and store TG[145]. The proliferation of T cells can also activate lipids synthesis as well as the FA catabolic pathways[146]. Although TG does not present the significant enrichment in T cell as the other three lipids traits, we observed that the p-values of enrichment analysis in another immune cell type, proliferating B cells, are significant (p-val$< 0.05$) robust to every cutoff score screened. It has been discussed that the Fc receptor (FcR) participates in the removal of the lipid indirectly and is strongly related to B cell proliferation[148].

We also observed the high enrichment in macrophages from heart tissue to lipids traits, HDL and LDL, as shown in Fig 3.3B. The p-values are consistently significant (p-val$< 0.05$) for HDL and the strongest enrichment for LDL, although not significant. Thus, we considered the macrophages from heart disease to be the most genetically relevant cell type in heart to lipids traits through gene expression. Extensive studies are investigating the association of macrophages and lipids, especially in heart disease patients[149, 150, 151]. For instance, the LDL in macrophages can initiate atherosclerotic lesions, a heart disease[150]. And the increased exposure of macrophages to HDL can enhance the risk of CAD[151].

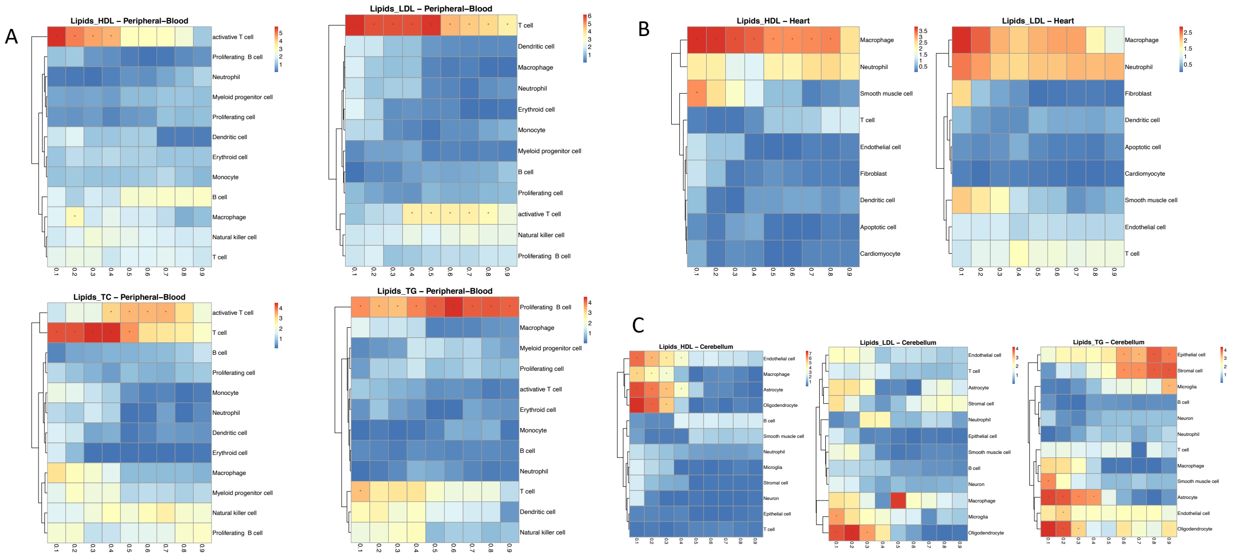


Figure 3.3: Lipids in enrichment analysis. Cell type enrichment analysis of lipids traits in PB (A), heart (B) and cerebellum of brain(C).

Several cell types in the cerebellum of the brain are also found to be enriched in the lipids traits from the results shown in Fig 3.3C. We identified oligodendrocytes, macrophages, astrocytes, and endothelials for HDL; oligodendrocytes, macroglias, and macrophages for LDL; and oligodendrocytes, astrocytes, and endothelials for TG. Their common shared cell types, oligodendrocytes, macrophages, and astrocytes, are also present the strongest correlation with lipids traits in PheWAS as shown in Fig 3.4. Especially for oligodendrocytes and astrocytes, there are numerous reports on how they require lipids in the brain [152, 153]. The primary function of oligodendrocytes in the synthesis of myelin, a lipids-riched membrane structured layer, is to enwrap and insulate nervous system axons[154, 155, 156].

Astrocytes, the other one of the most abundant cell types in brain, have a substantial contribution to myelination incorporated with lipids[157, 158]. On the other hand, we also found the importance of macrophages to lipids traits, and its decomposed PRS is highly correlated with the ones of oligodendrocytes and astrocytes in the brain. Besides the strong association between the macrophages and lipids metabolism[159], we also hypothesize that the macrophages which are not brain-resident may participate in the oligodendrocytes and astrocytes myelination at the transcriptomics level.

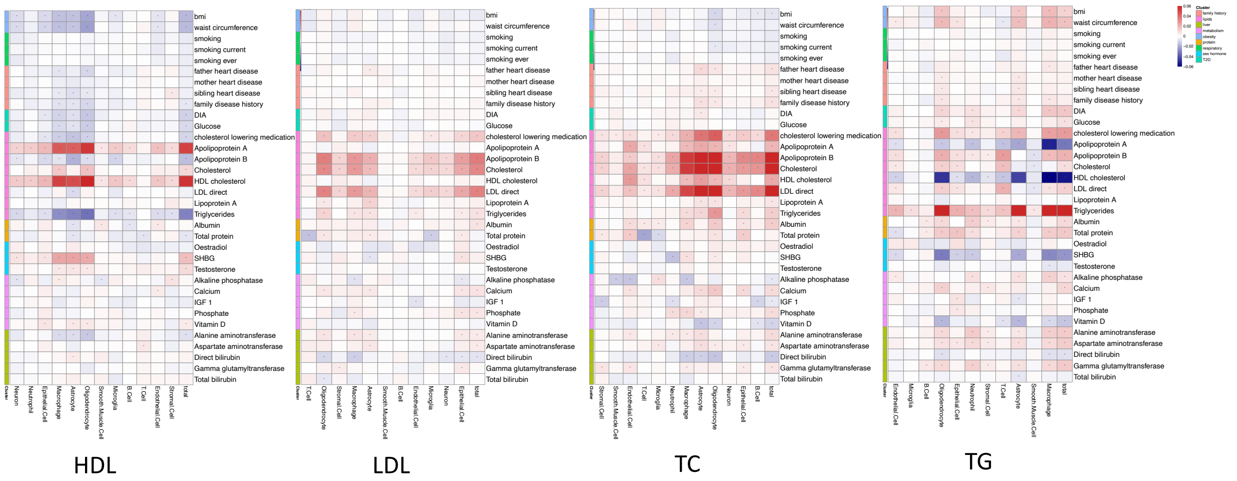


Figure 3.4: PheWAS of lipids traits in PB. The rows are manually grouped traits by their pathways and biological properties. The columns are dendrogram clustered cell types. The colors represented the correlations between the phenotypes and decomposed PRS with asteroid symbols when the Bonferroni corrected p-value$< 0.05$.

### Coronary artery disease

CAD is the leading cause of death[160, 161] and has an enormous health impact worldwide. Therefore, it is critical to learn about its possible risk factors. Specifically, one can easily ask a question about how the gene expression reflects the genetic risk factor. Therefore, in this session, we studied cell type enrichments in CAD by our framework and found signals in heart and brain tissues, two organs with high associations with CAD trait.

In Fig 3.5A, the cardiomyocytes and dendritic cells have consistent significant enrichment to CAD. Cardiomyocytes are the major cell population in the heart by mass. The growth of cardiomyocytes can result in hypertrophic cardiomyopathy (HCM), a heart disease that has high association with CAD[162, 163]. HCM is also proved to be a significant risk factor of human morbidity and mortality[164, 165, 166]. Previous reports, including longitudinal studies, have shown that up to 20% HCM patients also suffered from CAD[163, 167, 168, 169]. On the other hand, there are also some translational applications for CAD targeting cardiomyocyte in development, since the pharmacological and genetic control of the death of cardiomyocytes can effectively reduce the infarction risk to enhance the cardiac function[170]. The other signature cell type is dendritic cells (DC), which are a key cell population in the immune system with different functions. Especially DCs are associated with many cardiovascular diseases ranging from hypertension to heart failure after heart transplantation. DCs are antigen-presenting interplaying with CAD occurrence. In general, DCs can activate and regulate the immune system by all types of effect T cells, which is a typical pathological pattern of CAD[171]. Therefore, our framework indicates that genetics can increase the CAD risk by gene expression in cardiomyocytes and dendritic cells.

Furthermore, our framework discovered that endothelial cells are enriched in the cerebellum of the brain to CAD, as shown in Fig 3.5B. The endothelial dysfunction can change the white matter surrounding the blood vessels in the brain and lead to cerebral small vessel disease (SVD), a disease that has numerous shared risk factors with CAD, such as aging, hypertension, smoking, and diabetes[172]. In addition, downstream PheWAS analysis found strong associations between decomposed CAD PRS of endothelial cells with alkaline phosphatase (ALP), a homodimeric protein enzyme. A number of studies tried to understand the tissue-nonspecific alkaline phosphatase (TNAP) and found its high expression in brain microvessels, including brain microvascular endothelial cells (BMECs). The TNAP mouse model demonstrates the critical role of TNAP in aging, a well-known risk factor of CAD[173]. Integrating the results from both enrichment analysis and PheWAS in the brain, we hypothesize that the dysfunctional endothelial cells can indirectly influence the CAD risk by changing the blood vessels in the brain.

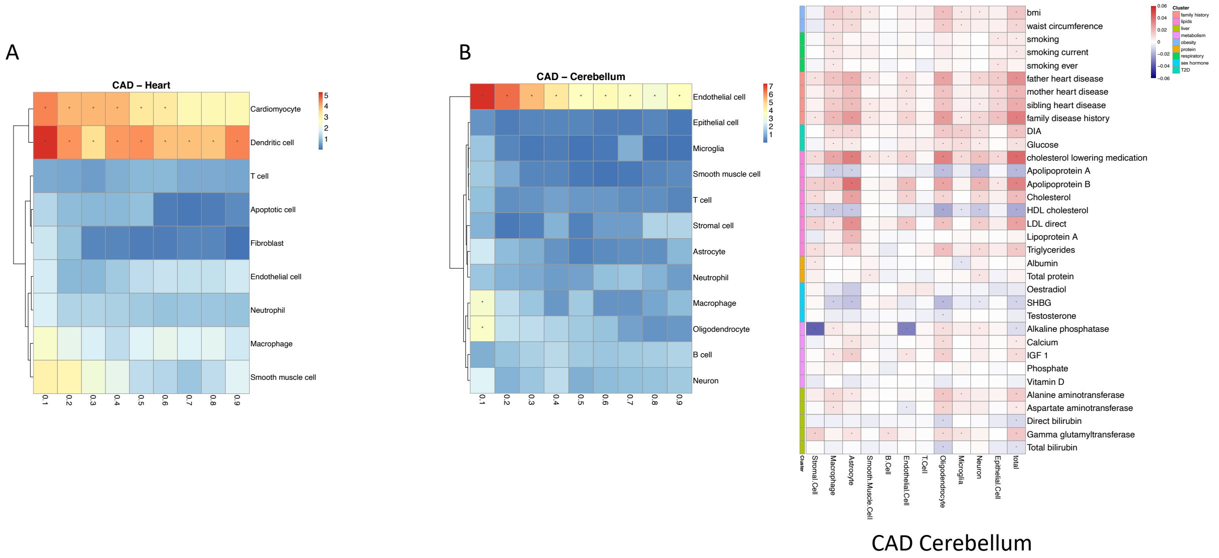


Figure 3.5: Analysis on CAD. (A) Cell type enrichment analysis in Heart. (B) Cell type enrichment analysis and PheWAS in cerebellum of brain.

### Artery

As artery is a tissue that plays a role in many common diseases, we also applied our framework on artery tissue to investigate the cell type-specific decomposed PRS on all eight traits, including CAD, AF, BC, T2D, and four lipids traits.

The cell type enrichment analysis results shown in Fig 3.6 and PheWAS results shown in Fig 3.7 suggest three major cell types in artery: fibroblasts, endothelial cells, and macrophages. Fibroblasts have enrichment in the artery to BC; both fibroblasts and endothelial cells have enrichment in AF and LDL; and for the other five traits, CAD, T2D, HDL, TC, and TG, all three mentioned cell types have significant enrichment, and the results are robust to the cutoff scores. In PheWAS analysis of TC and TG, these three cell types also have the most similar strong associations with metabolism traits compared with the overall PRS, indicating the dominance of them in the artery to CAD from the genetics perspective.

Vascular fibroblasts, located at the outer layer of the artery, participate in many biological

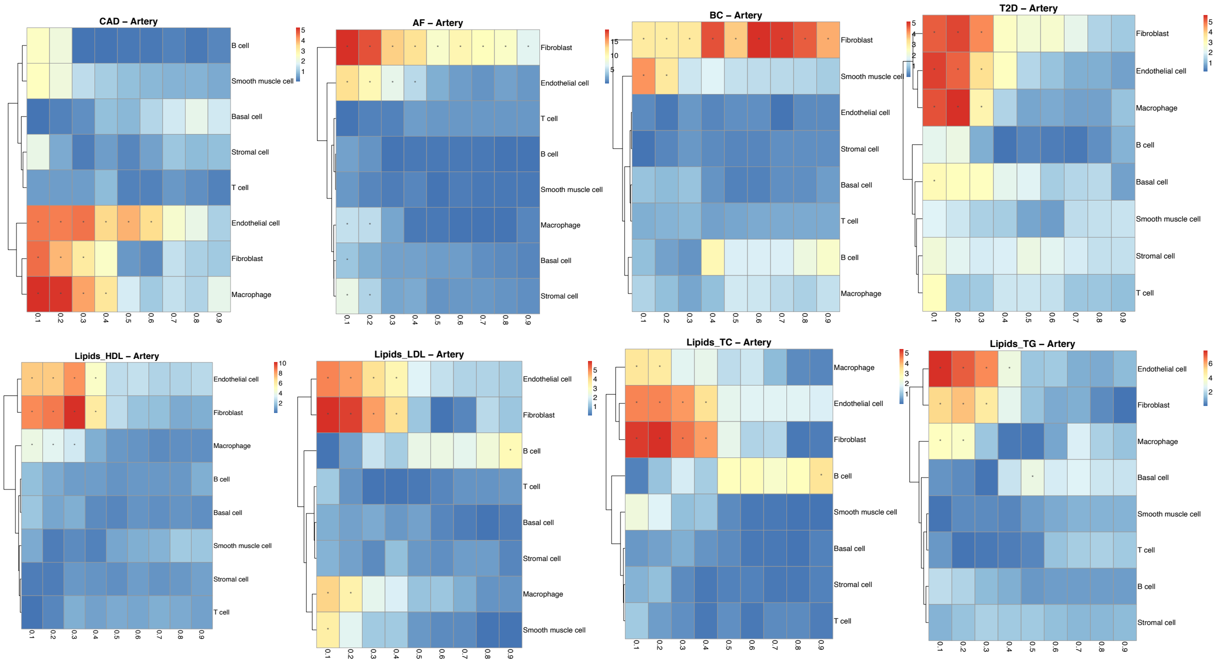


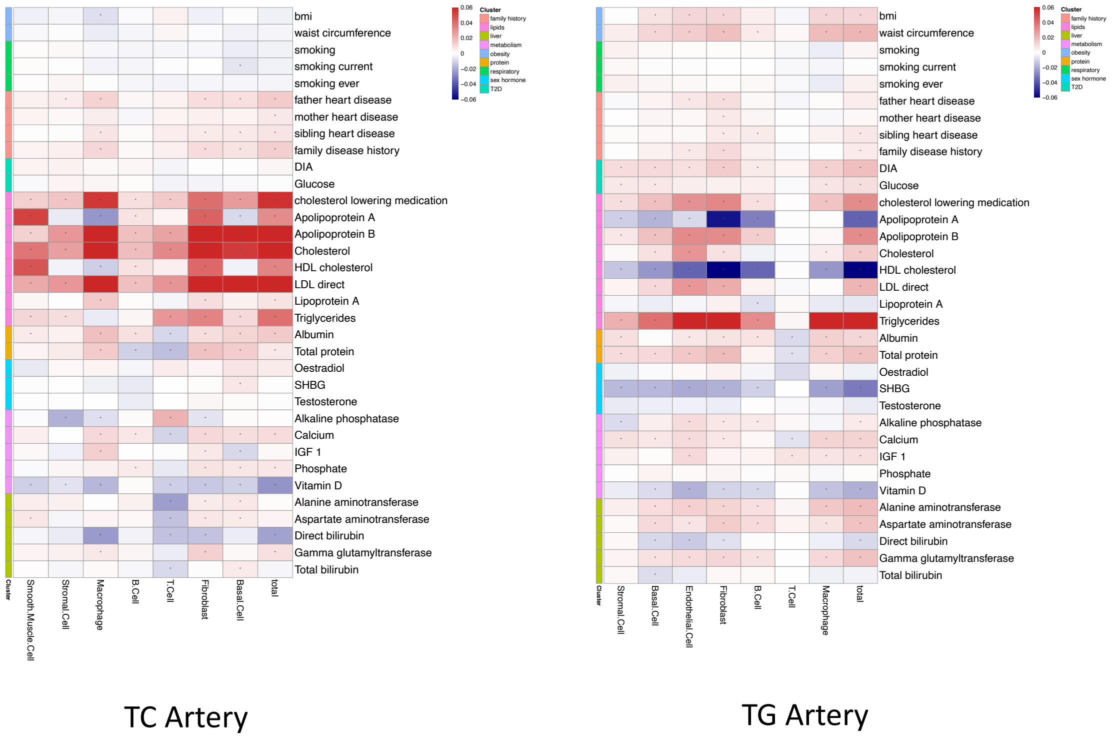
Figure 3.6: Artery in Enrichment Analysis. Cell type enrichment anlayis in artery for all eight common traits figured out the shared significant cell types, fibroblasts, endothelial cells and macrophages.

processes. A rat aorta model has shown that fibroblasts are involved in the vascular remodeling, the process of repairmen, and heart diseases, such as hypertension[174]. They also make a significant contribution to atherosclerosis, a chronic arterial disease[175]. Fibroblasts are involved in the inflammatory response and are sensitive to the stimulation of angiotensin II (ANG II), which can increase blood pressure and is a critical factor in hypertension.

The arterial innermost layer is composed of endothelial cells. And endothelial cells generate numerous cell-surface proteins in artery[176]. Thus, it is not surprising that the gene expression of endothelial cells takes a dominant intermediary role in the genetic association with multiple common traits. Endothelial cells participate in many pathological processes, including atherosclerosis and cancer[177]. Previous research also demonstrated that endothelial cells and lipids have bidirectional interactions. The endothelium is involved in lipids metabolism and, in turn, regulated and processed by the lipids/lipoprotein on the endothelial cell surface[178, 179].

Macrophages, which play a central role of many pathological processes in the artery, are also found to be a significant cell population in our downstream analysis to most traits investigated, especially for the cardiovascular disease such as CAD and atherosclerosis. Macrophages will participate in the biological processes, for example, inflammatory response and plasticity promotion[180]. Additionally, macrophages are actively engaged in the development and modification of LDL[181]. Other research also figured that macrophages are a critical player in lipid-rich carotid artery plaque[182].

In summary, if we decompose the overall PRS by the cell types in the artery, we find out these three cell types mentioned above are the significant contributors to the genetic risk factors in terms of their gene expression.



TC Artery

TG Artery

Figure 3.7: Artery in PheWAS. Lipids traits PheWAS in artery discovered the coherent cell types from enrichment analysis.

### 3.3.3 Construct more predictable decomposed PRS

The other important motivation of our framework is to construct more accurate PRS for prediction. A better decomposed PRS should have stronger correlations with traits in PheWAS compared with the overall PRS or other decomposed PRS. In Fig 3.8, we show that four novel PRSs more predictable to traits by decomposing either by tissues or cell types.

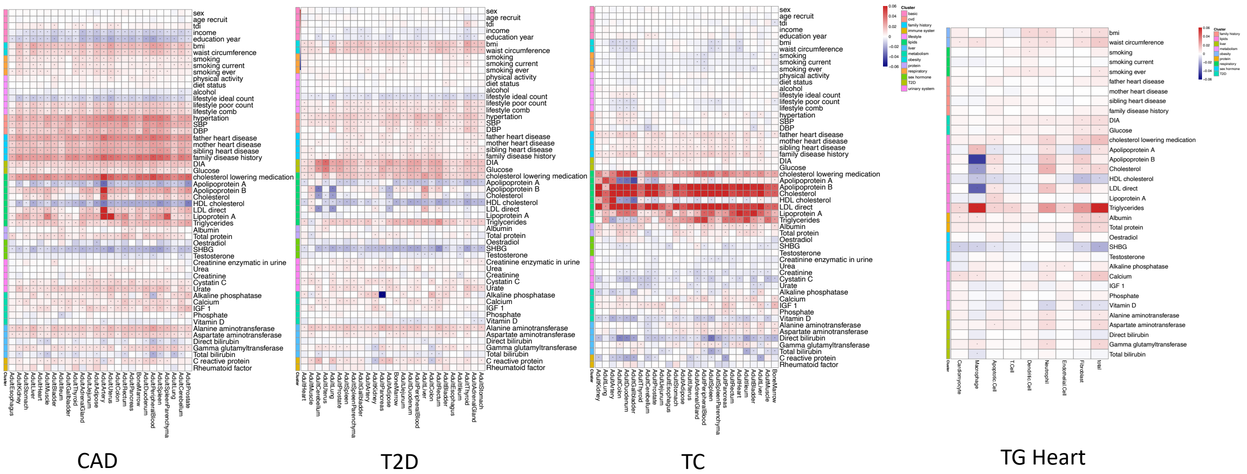


Figure 3.8: Predicable decomposed PRSs. The first three panels show tissue enrichment analysis for CAD, T2D, and TC, indicating different patterns of decomposed PRS and identifying the more predicable PRSs compared with other tissues. The last panel is the heart cell type enrichment analysis for TG and recognizes macrophage decomposed PRS having better prediction powers than the overall PRS.

We applied the framework to 26 tissues that are representative of the human body. The decomposed CAD PRS of the artery can better predict the lipids traits compared with other tissues. This decomposed PRS have the most significant positive correlations with lipids traits, including cholesterol-lowering medication, apolipoprotein B, cholesterol, LDL, and lipoprotein A. The high values of these metabolism traits can predict many diseases such as atherosclerotic cardiovascular disease (CVD). The high decomposed CAD PRS in the artery has a concordance with these traits, indicating that the high risk of CAD is more likely to be accompanied by high values of unhealthy cholesterol. On the other hand, it has the strongest negative correlations with HDL and apolipoprotein A (ApoA), which is consistent with the fact that HDL has an inverse association with CAD risk[183].

A similar analysis in cell type PheWAS of T2D discovered the strong negative correlation between decomposed T2D PRS in the pancreas and ALP trait. It is well known that intestinal alkaline phosphatase (IAP) plays an essential anti-inflammatory role in metabolism, and its absence may lead to hyperglycemia. A study on the mouse also showed that IAP in pancreatic $\beta$-cells could be a possible therapeutic target for T2D[184].

We now discuss the decomposed PRS result for TC. Among all 26 tissues, our framework discovered two subgroups based on their correlations with metabolic traits and correlation significance compared with the rest of the tissues. The first subgroup includes colon, duodenum, and gallbladder. Their decomposed TC PRS have positive correlations with all unhealthy lipids traits, i.e., TC and LDL, and have opposite correlations with HDL and ApoA. The higher prediction ability of the decomposed TC PRS in intestinal tissues is reasonable since there are multiple essential pathways involved in intestinal tissues, such as lipids absorption and regulation[185]. On the contrary, we observed that the other subgroup, including artery, lung, and kidney, have strong positive correlations with HDL and ApoA, implying that high HDL and ApoA concentrations may damage tissue functions. HDL and its major component ApoA are usually considered to remove cholesterol and carry it back to the liver. A cohort study has found a genetic variant within the gene *SCARB1* associated with both high risk of CAD and elevated HDL[186]. This study revealed that raising HDL in the artery cannot protect against heart disease. Therefore, the higher decomposed TC PRS is positively associated with HDL and ApoA traits. Similar conclusions were also found in lung, that elevated HDL will decrease lung function. A previous cohort study identified HDL accumulation associated with the decrease of forced vital capacity volume (FVC) and forced expiratory volume in the 1 s (FEV1)[187]. Although no literature can explain how decomposed TC PRS in the kidney is positively associated with HDL and ApoA, there is research studying renal handling of HDL and how it is involved in kidney damage[135, 136].

Finally, our novel framework suggested that the decomposed TG PRS in heart macrophages have better predictions based on its stronger associations with metabolism traits shown in the last panel of Fig 3.8. Compared with the overall TG PRS shown as the last column of the heatmap, it has a positive correlation with TG and significant negative correlations with most lipids traits. Therefore, heart macrophages make the majority of the contribution to the overall PRS prediction power. But the higher PRS derived from heart macrophage may point to a lower level of other lipids traits.

## 3.4 Discussion

This chapter introduces a novel statistical framework integrating scRNA-seq and GWAS data to gain better biological insights for common diseases through constructing tissue and cell type specific polygenic risk scores. Our framework aims to understand how traits are associated with SNPs through gene expression regulation and can facilitate patient subtyping (though not covered in this chapter). It starts from curating cell type-specific and tissue-specific DE genes from scRNA-seq data. Through the iDEA package, we obtain gene-level weight scores quantifying the likelihood of genes to be DE genes. After mapping the SNPs to genes, we derive normalized SNP-level weight scores that sum up to 1 across cell types/tissues. The final decomposed PRS are calculated based on the SNP-level weight scores and GWAS summary statistics. The framework was applied to several common traits either by cell type decomposition or by tissue decomposition.

We conducted two types of downstream analysis: enrichment analysis and PheWAS. The enrichment analysis utilizes the group-specific SNP lists derived from the framework to get group-specific heritability compared with the overall one. We considered different cutoff scores of SNP-level weight scores ranging from 0.1 to 0.9 to select the SNPs. The groups with consistent significant signals are considered to be strongly associated with target traits. PheWAS enables us to figure out the dominant groups that contribute to overall PRS and the groups whose decomposed PRSs are better proxies for traits prediction. All the results are summarized in Table 3.1.

We first investigated the reliability of the framework by cell type enrichment analysis in the brain for AD, and tissue enrichment analysis for HDL and AF. Our framework successfully identified high heritability enrichment in brain T cells for AD, kidney for HDL, artery, and adrenal gland for AF. Next, the framework was applied to identify the dominant cell types making significant contributions to traits. The four lipids traits share similar dominant cell types based on the results of enrichment analysis and PheWAS. Specifically, multiple types of T cells are enriched in PB; macrophages and neutrophils are enriched in the heart; oligodendrocytes, astrocytes, and macrophages are significantly enriched and dominant in the brain. From the results of CAD, we also identified cardiomyocytes and dendritic cells in the heart and endothelial cells in the brain. Another interesting finding is the leading cell types of the artery in genetics risk mediated by gene expression: fibroblasts, endothelial cells, and macrophages. Both enrichment analysis and PheWAS have coherent significant results for these three cell types across all eight common traits we studied. Finally, we listed the decomposed PRS, which are more predictable than the rest of the groups or the overall PRS for traits CAD, T2D, TC and TG.

Unlike the classical PRSs, our framework only uses the SNPs selected, annotated, and grouped by gene transcript. Thus, the decomposed PRS can only explain the differences of PRSs through gene expression. A probable and critical downstream usage of the decomposed PRS is patient subtyping. The decomposed PRS may better explain how the cell type- or tissue-specific gene expression is associated with the disease. The framework can also identify therapeutic targets for different patient subtypes based on differential decomposed PRSs. Another evaluation is necessary to see whether our framework can outperform other methods, specifically, LDSC and decomposed PRSs when pathways annotate the SNPs. In terms of cell type decomposed PRSs, we will add the analyses on kidney and liver tissues for clinical use, as they play critical roles in many complex diseases, such as T2D.

We note that our framework can be improved in a number of ways. For example, we can optimize the model parameters by deploying different methods to get DE gene summary statistics and the group-specific DE gene set. For example, it is feasible to replace marginal linear regression with joint linear regression. Another possible improvement is to add neighboring genes when mapping SNPs to genes. In this chapter, we tried the neighborhood of 500kb for genes, but the group-specific SNP lists overlapped with each other across tissues or cell types, leading to the unspecificity to groups and the similar decomposed PRSs which are not distinguishable in PheWAS. However, a smaller neighborhood can be considered to optimize the results rather than only using the original gene sequence. Besides, the framework only utilizes P+T method for PRS calculation. A more state-of-the-art risk prediction model, e.g. PRS-CS may better handle SNP effect sizes from the whole genome instead of clumping SNPs by P+T. Furthermore, we can consider other SNP-level weight score $b_k^j$.

In general, the new framework can bring a new perspective on joint analysis on scRNA-seq and GWAS. Despite the challenges from the complex biological system, our framework has been able to suggest how transcriptome can serve as an interpretable intermediate feature between the genotypes and phenotypes. There are many possibilities for the integrative analysis of multi-omics data, not only genetics and transcriptomics but also epigenetic, proteomics, and others. They can better explain how genetic factor impact disease risk and the identifications of disease subtypes.

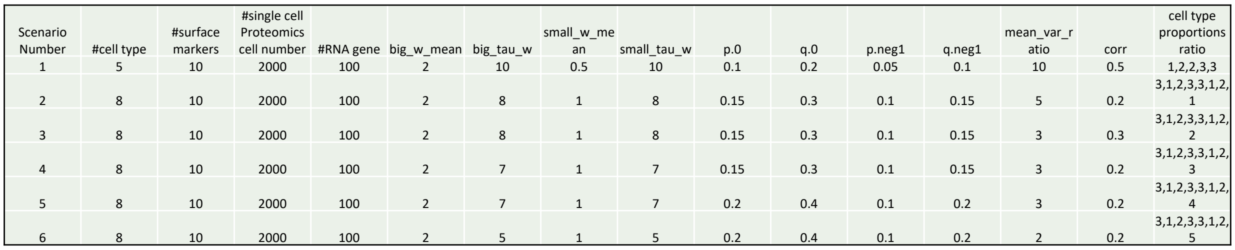
## 3.5 Online Methods and Materials

### 3.5.1 Data Availability

The HCL can be accessed under the GEO accession number GSE134355 at `http://bis.zju.edu.cn/HCL/` or `https://db.cngb.org/HCL/`. AD GWAS data was obtained from IGAP. BC GWAS data were accessed from Michailidou, Kyriaki, et al.(2017)[188]. The other GWAS data were from UK Biobank[189].

# Appendix A

# Supplementary Figures and Tables



| Scenario Number | #cell type | #surface markers | #single cell Proteomics cell number | #RNA gene | big_w_mean | big_tau_w | small_w_mean | small_tau_w | p.0 | q.0 | p.neg1 | q.neg1 | mean_var_ratio | corr | cell type proportions ratio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 10 | 2000 | 100 | 2 | 10 | 0.5 | 10 | 0.1 | 0.2 | 0.05 | 0.1 | 10 | 0.5 | 1,2,2,3,3 |
| 2 | 8 | 10 | 2000 | 100 | 2 | 8 | 1 | 8 | 0.15 | 0.3 | 0.1 | 0.15 | 5 | 0.2 | 3,1,2,3,3,1,2,1 |
| 3 | 8 | 10 | 2000 | 100 | 2 | 8 | 1 | 8 | 0.15 | 0.3 | 0.1 | 0.15 | 3 | 0.3 | 3,1,2,3,3,1,2,2 |
| 4 | 8 | 10 | 2000 | 100 | 2 | 7 | 1 | 7 | 0.15 | 0.3 | 0.1 | 0.15 | 3 | 0.2 | 3,1,2,3,3,1,2,3 |
| 5 | 8 | 10 | 2000 | 100 | 2 | 7 | 1 | 7 | 0.2 | 0.4 | 0.1 | 0.2 | 3 | 0.2 | 3,1,2,3,3,1,2,4 |
| 6 | 8 | 10 | 2000 | 100 | 2 | 5 | 1 | 5 | 0.2 | 0.4 | 0.1 | 0.2 | 2 | 0.2 | 3,1,2,3,3,1,2,5 |

Table A.1: Simulation parameters setting.

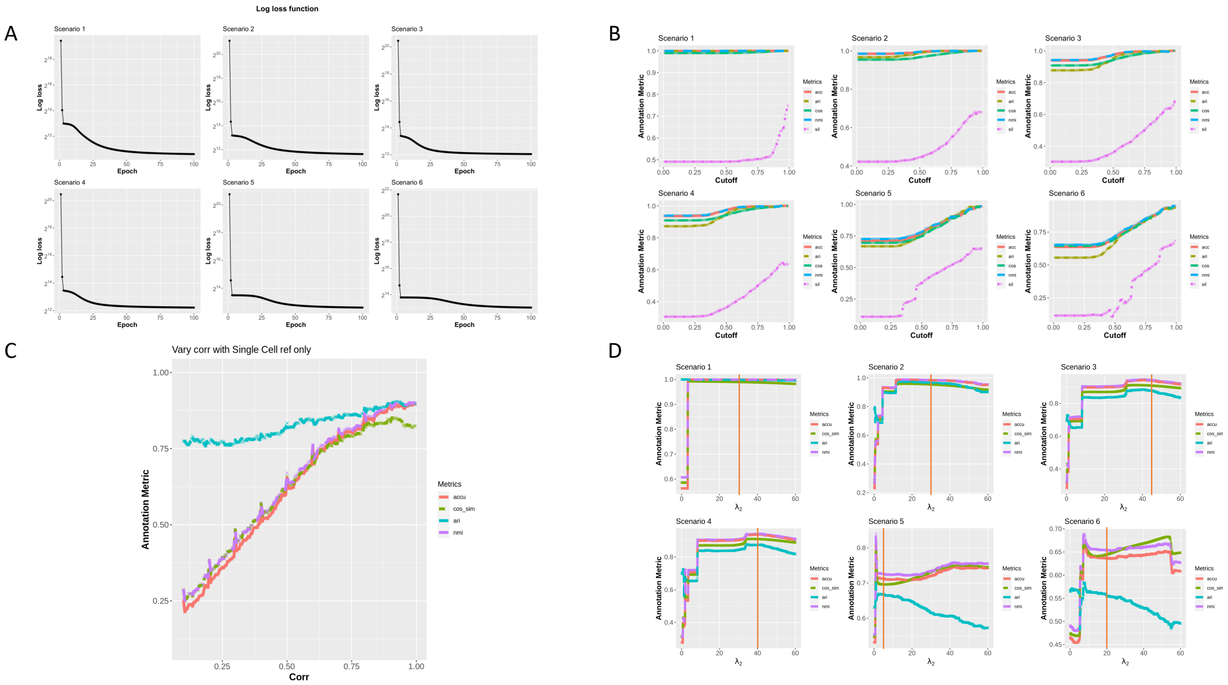


Figure A.1: ProtAnno optimization properties. A) ProtAnno convergence for six simulation scenarios. The x-axis represents iteration number. The y-axis represents log loss function values. B) The subsetting cutoff parameter. The x-axis represents the cutoff value. The y-axis represents annotation metrics value. C) The impact of transcriptomics data on annotation in the ProtAnno model with only transcriptomics penalization, $AS$. The x-axis represents the correlation between transcriptomics predicted signature matrix $AS$ and true protein signature matrix $W$. The y-axis represents the annotation metrics value. D) The evaluation of parameter tuning algorithm. The x-axis corresponds to $\lambda_2$ from 0 to 60 under six simulation scenarios. The color represents different annotation metrics. The y-axis represents the annotation metrics value. The red vertical line represents the optimal $\lambda_2$ by the parameter tuning algorithm for each simulation scenario.

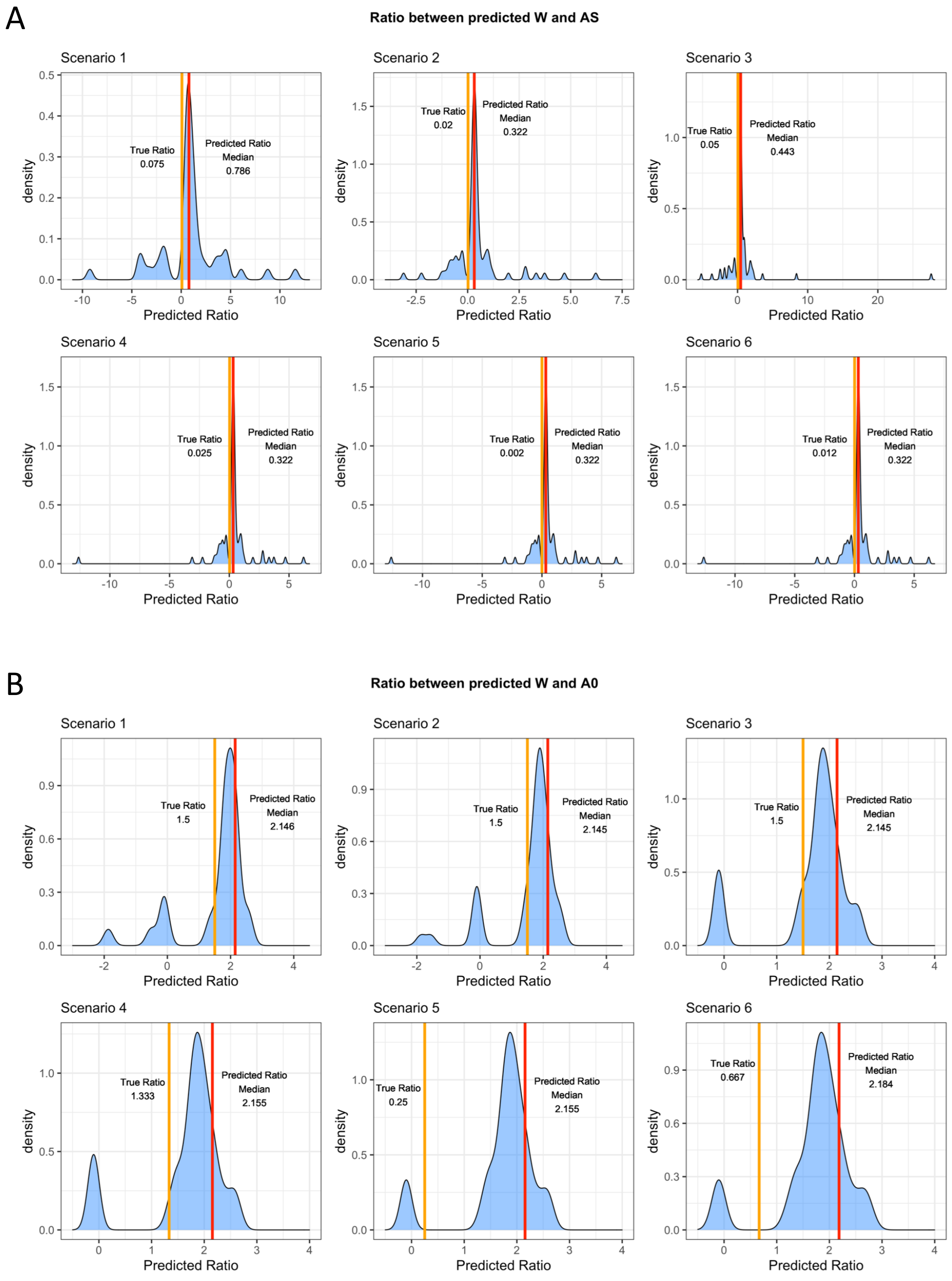


Figure A.2: Penalization powers in optimization. A) The distribution of ratios between estimated $W$ and the input, $AS$. The red vertical line is the median of the distribution. The orange vertical line is the true ratio in the rewritten optimization algorithm. B) The distribution of ratios between estimated $W$ and the input, $A_0$. The red vertical line is the median of the distribution. The orange vertical line is the true ratio in the rewritten optimization algorithm.

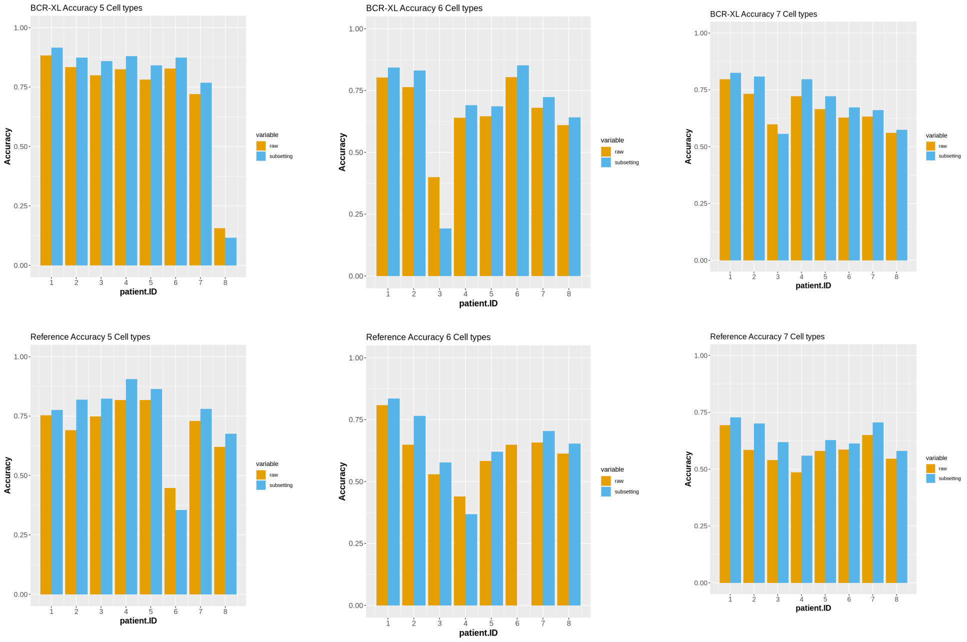


Figure A.3: The prediction power on BCR data. The annotation accuracy barplots across samples from stimulated (first row) and unstimulated (second row) groups. The columns are the cell type numbers used (5 cell types in the first column, 6 cell types in the second column, and 7 cell types in the third column). The color represents the raw output or subsetted output from ProtAnno.

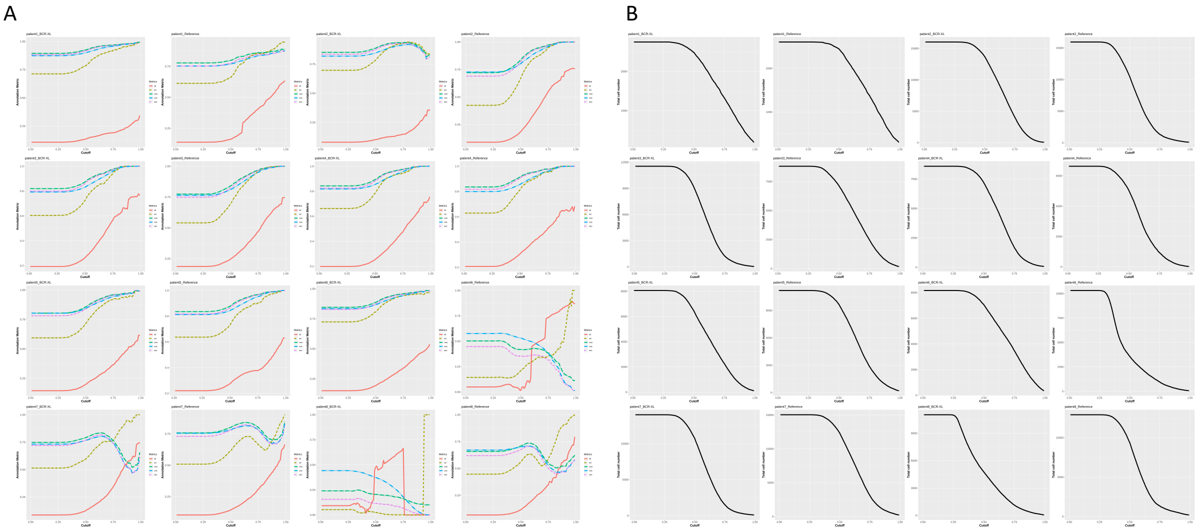


Figure A.4: Annotation accuracy improvement by subsetting. A) The annotation metrics with different subsetting cutoffs. The x-axis represents the cutoff value. The y-axis represents the annotation metric value. B) The numbers of kept cells after subsetting filter. The x-axis represents the cutoff value. The y-axis represents the remaining cell counts.

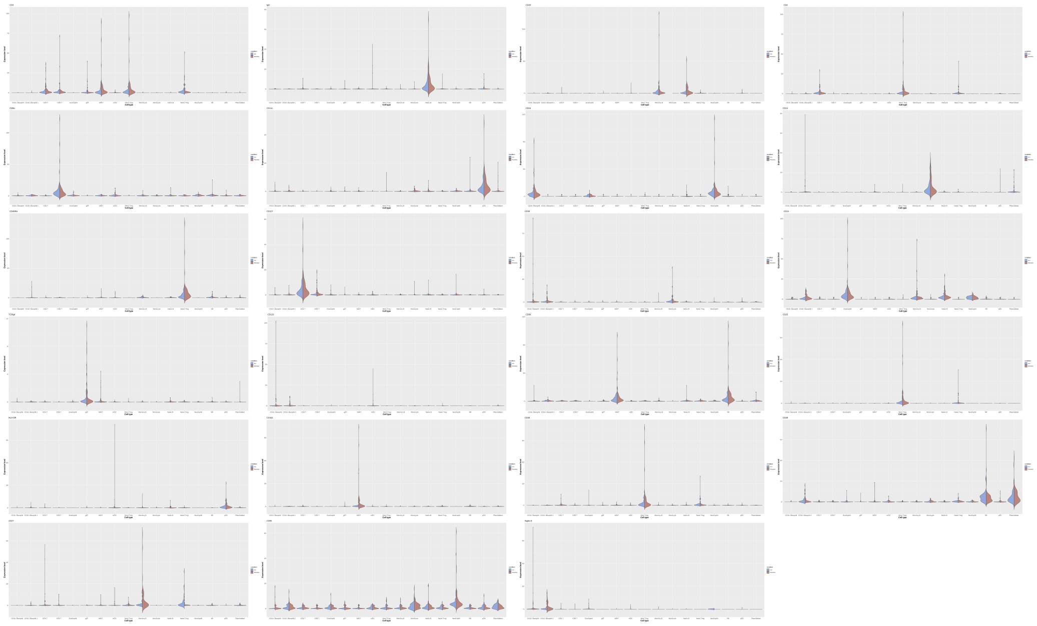


Figure A.5: The violin plots of estimated signature expression in $W$ in a longitudinal covid-19 study by ProtAnno. The x-axis represents the cell type and the y-axis is the distribution of values in the signature matrix across patients within groups. The colors represent patient groups (ICU and recovery).

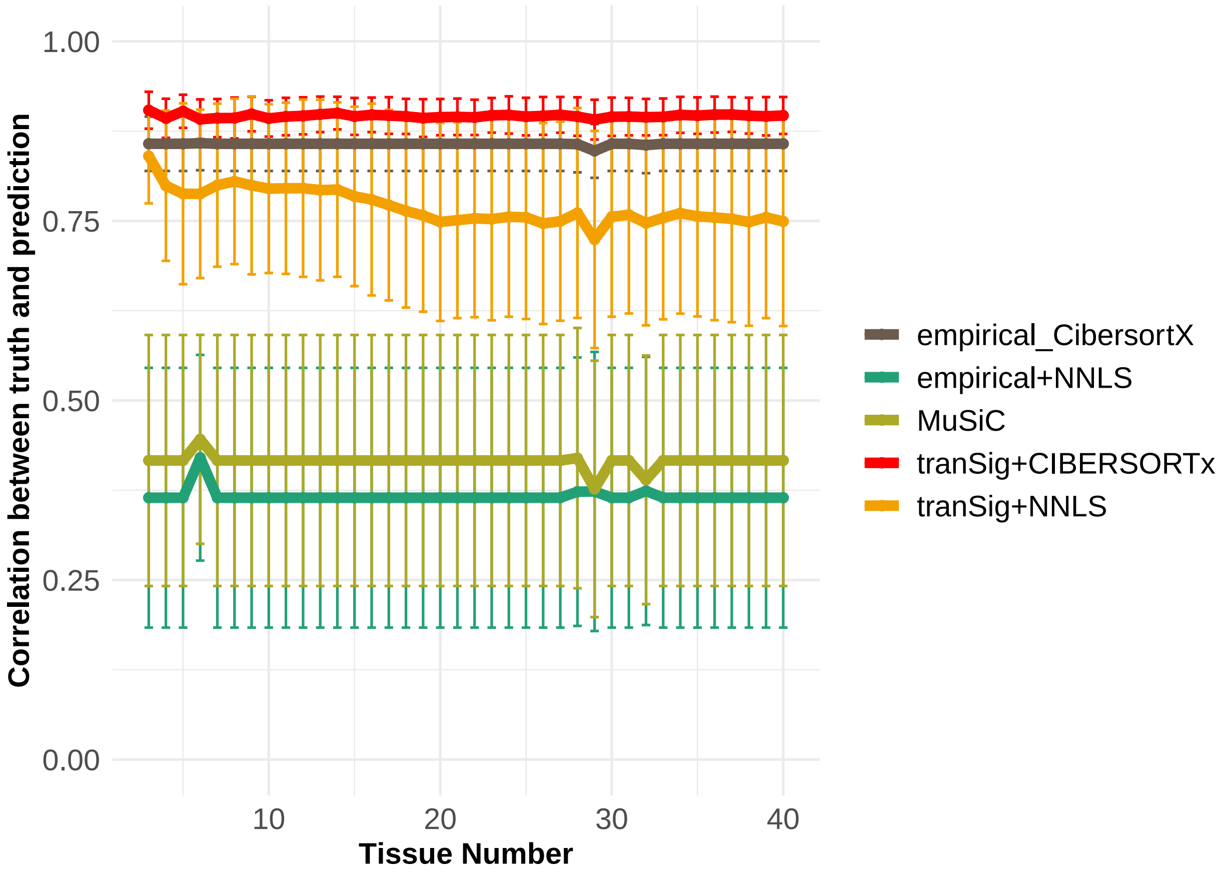


Figure A.6: The curves to evaluate the robustness of tranSig signature matrix against tissue number used in tranSig model. The colors code the combinations of signature matrix and deconvolution tools. The vertical lines are error bar of mean $\pm 0.5$*s.d.

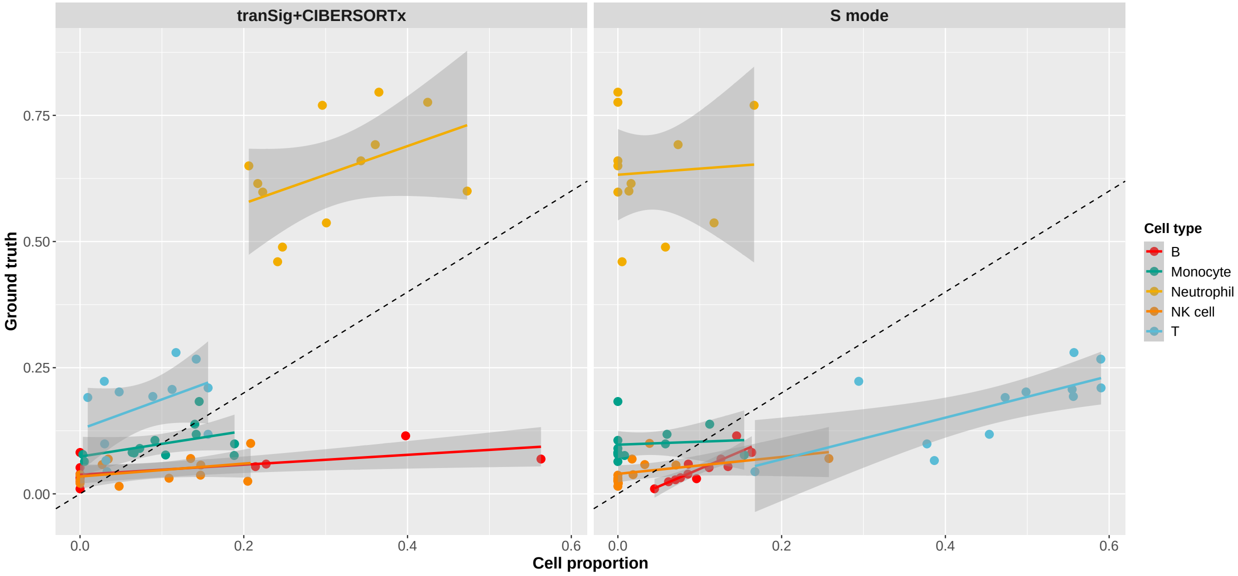


Figure A.7: The ground truth and cell type proportions estimated by tranSig and CIBERSORTx S mode, colored by cell types.

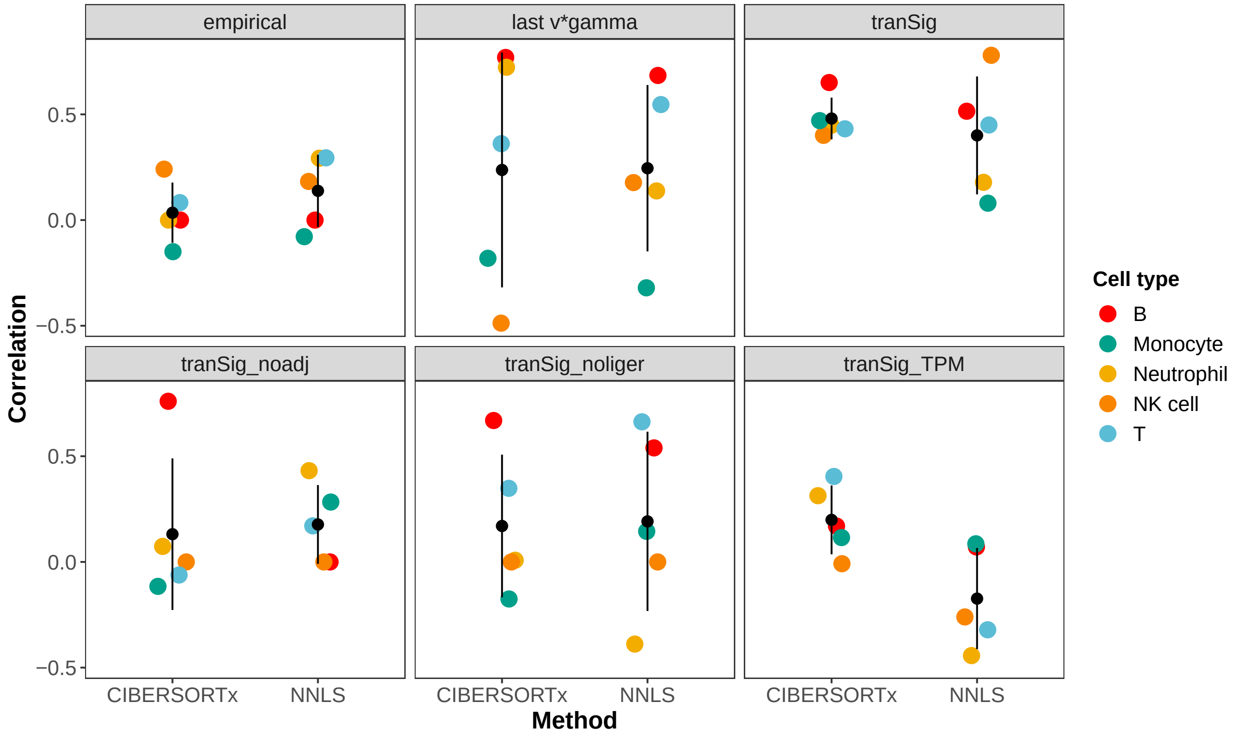


Figure A.8: Jitter plots of the correlations between estimated cell type proportions and the ground truth for Newman et al. blood samples ($n = 12$), with color coded by cell types. Last $v * \gamma$ is denoted as the estimates of the last iteration in the tranSig model. tranSig_noadj is denoted as the deconvolution on the bulk mixture without adjustment to single cell space and the signature matrix constructed by the tranSig model. tranSig_noliger represents that the tranSig framework is implemented on the single cell datasets without LIGER for batch correction. tranSig_TPM represents that single cell expression profiles are normalized in TPM space as the input of the tranSig framework. Data are presented as means ± s.d.

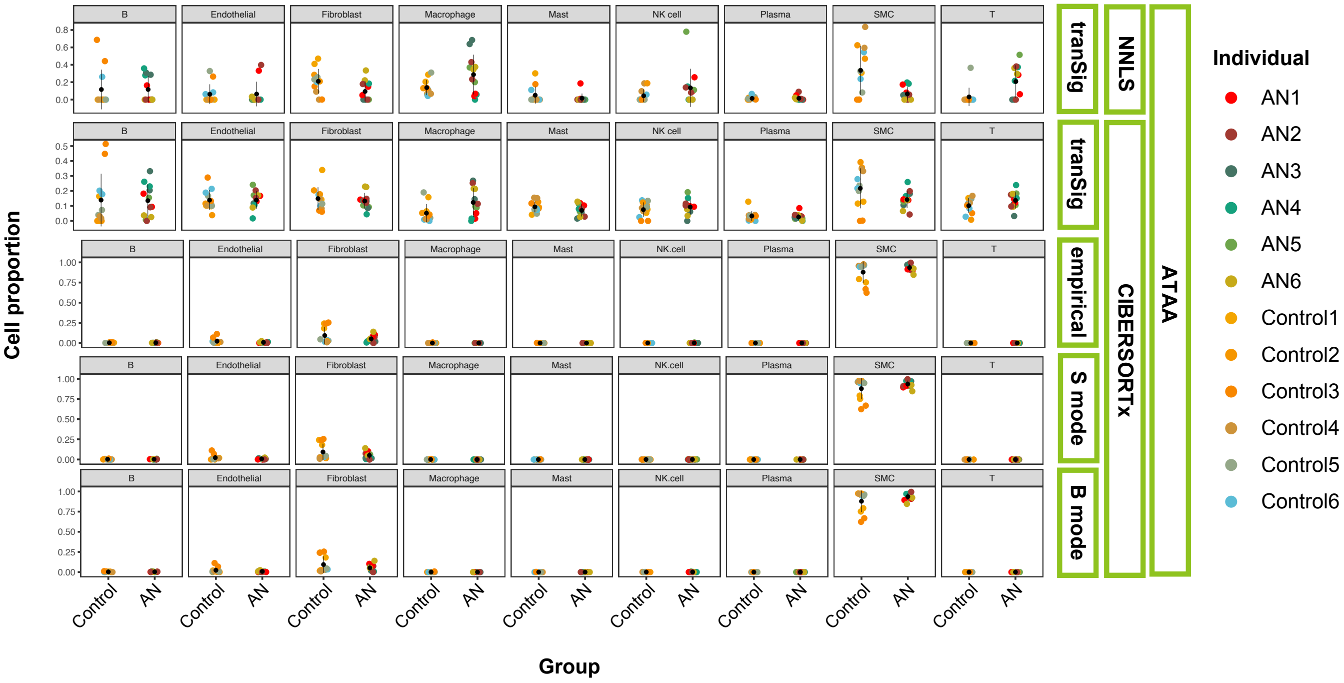


Figure A.9: Jitter plots of estimated cell type proportions for Chen et al. subjects (aortic tissues from 6 health donors as control and aneurysm tissues from patients with ascending aortic aneurysm as AN), color coded by individuals. Color bars on the right annotate the deconvolution methods with the ATAA dataset as the single cell reference. Data are expressed as means ± s.d.

# Appendix B

# ProtAnno Theoretical Proofs

## B.1 Parameter Tuning

To get the optimal penalty parameters $\lambda_1$,$\lambda_2$,$\mu$ and $\eta$, we considered both the KarushKuhn-Tucker (KKT) condition combined with empirical screening. We first obtained the initial $W$ and $H$ by setting the arbitrary penalization: $\lambda_1 = 1, \lambda_2 = 10, \mu = 50$ and $\eta = 50$. This setting can give an acceptable result in most cases based on our empirical experiments. The first order derivatives of loss function are

$$\begin{aligned} \nabla_W L(W,H) &= W(HH^T + \mu I) - XH^T - \lambda_1 AS - \lambda_2 A_0 \\ \nabla_H L(W,H) &= -W^T X + W^T WH + \eta 1_K 1_K 1_K^T H - \eta 1_K 1_N^T \end{aligned} \tag{B.1}$$

We first initialized $\eta$ by satisfying KKT conditions.

$$\begin{aligned} (W(HH^T + \mu I) - XH^T - \lambda_1 AS - \lambda_2 A_0) \odot W &= 0 \\ (-W^T X + W^T WH + \eta 1_K 1_K 1_K^T H - \eta 1_K 1_N^T) \odot H &= 0 \\ W(HH^T + \mu I) - XH^T - \lambda_1 AS - \lambda_2 A_0 &\geq 0 \\ -W^T X + W^T WH + \eta 1_K 1_K 1_K^T H - \eta 1_K 1_N^T &\geq 0 \end{aligned} \tag{B.2}$$

by the second sufficient inequality, we set

$$\eta := \| (W^T WH - W^T X)/(1_K 1_K^T H - 1_K 1_N^T) \|_{\text{median}} \tag{B.3}$$

After getting the initial $\eta$, we initialized $\lambda 2$ and $\lambda_1$ in order by minimizing Adjusted Rank Index (ARI) with Louvain clustering. We considered the tuning parameters $\lambda 2$ and $\lambda_1$ from $0.1, 1, 10$, and $100$. The parameter $\mu$ is charging of the scale and penalization power on signature matrix $W$. Thus, a smaller $\mu$ can result in a larger norm of $W$. Therefore, we developed a new metric to evaluate the reliability by the difference between the mean value of expression profile $X$ and the mean value of signature matrix $W$. To eliminate the non-Gaussian effects, we also considered the difference between the median values of $X$ and $W$. Thus, the new metric is formulated as

$$D(\mu) := (X_{\text{mean}} - W_{\text{mean}}) + (X_{\text{median}} - W_{\text{median}}) \tag{B.4}$$

## B.2 Theoretical Proofs

### Convergence of Algorithms

**Lemma B.2.1.** *For any symmetric nonnegative matrix $Q \in \mathbb{R}^{K\times K}$ and row vector $w \in \mathbb{R}_+^K$ and $a \in \mathbb{R}_{\geq 0}^K$, the following matrix*

$$F = diag\{\frac{(wQ+a)_1}{w_1}, \frac{(wQ+a)_2}{w_2}, \ldots, \frac{(wQ+a)_K}{w_K}\} - Q \in \mathbb{R}^{K\times K} \tag{B.5}$$

*is always semi-positive definite.*

*Proof.* We construct a new matrix $S \in \mathbb{R}^{K\times K}$ by $S_{ij} = w_i F_{ij} w_j$, where $S_{ij}$ is the element of S whose row is $i$ and column is $j$. And we reformulate $F$ to be

$$F_{ij} = \begin{cases} -Q_{ij}, & \text{if } i \neq j \\ \frac{\Sigma_k w_k Q_{ik} + a_i}{w_i} - Q_{ii}, & \text{if } i = j \end{cases} \tag{B.6}$$

For any nonnegative row vector $v \in \mathbb{R}^{K \times K}$, we have

$$\begin{aligned} vSv^T &= \Sigma_{i,j} v_i S_{ij} v_j \\ &= \Sigma_{i,j} v_i w_i F_{ij} w_j v_j \end{aligned} \tag{B.7}$$

Since row vectors $w$ and $v$ are nonnegative, $F$ is semi-positive definite when $S$ is semi-positive definite. Therefore, it is sufficient if we can prove $S$ is semi-positive definite. In the following, we follow equation B.7 and prove that the product is always nonnegative.

$$\begin{aligned} vSv^T &= \Sigma_{i,j} v_i w_i F_{ij} w_j v_j \\ &= \Sigma_i v_i w_i \frac{\Sigma_k w_k Q_{ik} + a_i}{w_i} w_i v_i - \Sigma_{i,j} v_i w_i Q_{ij} w_j v_j \\ &= \Sigma_i (\Sigma_k w_k Q_{ik}) w_i v_i^2 + \Sigma_i a_i w_i v_i^2 - \Sigma_{i,j} v_i w_i Q_{ij} w_j v_j \\ &= \Sigma_{i,j} w_i w_j Q_{ij} v_i^2 + \Sigma_i a_i w_i v_i^2 - \Sigma_{i,j} w_i w_j Q_{ij} v_i v_j \\ &= \Sigma_{i,j} w_i w_j Q_{ij} (v_i^2 - v_i v_j) + \Sigma_i a_i w_i v_i^2 \\ &= \frac{1}{2} \Sigma_{i,j} w_i w_j Q_{ij} (v_i^2 + v_j^2 - 2 v_i v_j) + \Sigma_i a_i w_i v_i^2 \\ &= \frac{1}{2} \Sigma_{i,j} |v_i - v_j| w_i Q_{ij} w_j |v_i - v_j| + \Sigma_i a_i w_i v_i^2 \geq 0 \end{aligned} \tag{B.8}$$

The first term in the above formula is nonnegative since $w$ is nonnegative and $Q$ is semi-positive definite. The second term is always greater than or equal to 0 due the nonnegativity of $a$ and $w$. Thus, $S$ is a semi-positive definite matrix.

□

**Theorem B.2.2.** *Consider the following quadratic optimization problem,*

$$\min_{w\geq 0} E(w) = \frac{1}{2} \parallel x - wH \parallel_F^2 - \lambda_1 tr(w^T a_1) - \lambda_2 tr(w^T a_2) + \frac{\mu}{2} \parallel w \parallel^2 \tag{B.9}$$

*which is the loss function for a row vector $w \in \mathbb{R}^K$. In this optimization problem, $H \in \mathbb{R}^{K\times N}$ is a constant non-negative matrix, and $x, a_1, a_2 \in \mathbb{R}^K$ are constant row vectors. The penalty parameters, $\lambda_1, \lambda_2$, and $\mu$ are positive numbers. The the following update rule*

$$w_j^{t+1} = w_j^t \frac{[xH^T]_j^+ + \lambda_1[(AS)_i]_j^+ + \lambda_2[(A_0)_i]_j^+}{[w_i(HH^T + \mu I)]_j + [xH^T]_j^- + \lambda_1[a_1]_j^- + \lambda_2[a_2]_j^-} \tag{B.10}$$

*where we denote*

$$\begin{aligned} x^+ &= \begin{cases} x, & \textit{if}\, x \geq 0 \\ 0, & \textit{if}\, x < 0 \end{cases} \\ x^- &= \begin{cases} 0, & \textit{if}\, x > 0 \\ -x, & \textit{if}\, x \leq 0 \end{cases} \end{aligned} \tag{B.11}$$

*converges to its optimal solution with the convergence rate*

$$\min_{t\in[1,K]} \parallel w^{t+1} - w^t \parallel \leq \frac{2}{\mu(K-1)} E(w^0) \tag{B.12}$$

*Proof.* To prove that there exists a $w^{t+1}$ that $E(w^{t+1}) \geq E(w^t)$, we would like to construct an auxiliary function $F(w, w^t)$, s.t.

$$\begin{aligned} F(w^t, w^t) &= E(w^t) \\ F(w, w^t) &\geq E(w) \end{aligned} \tag{B.13}$$

Then we have

$$E(w^t) = F(w^t, w^t) \geq F(w^{t+1}, w^t) \geq E(w^{t+1}) \tag{B.14}$$

where $w^{t+1} = \arg\min_w F(w, w^t)$. Thus, the loss function $E(w^t)$ is monotonously non-increasing w.r.t the iteration $t$. We define the auxiliary function as the following.

$$F(w, w^t) = E(w^t) + \nabla E(w^t)(w - w^t)^T + \frac{1}{2}(w - w^t)J(w^t)(w - w^t)^T \tag{B.15}$$

where

$$\begin{aligned} J(w^t) = \quad & \text{diag}(\frac{w_1^t[HH^T + \mu I]_1 + [xH^T]_1^- + [\lambda_1 a_1]_1^- + [\lambda_2 a_2]_1^-}{w_1^t}, \\ & \frac{w_2^t[HH^T + \mu I]_2 + [xH^T]_2^- + [\lambda_1 a_1]_2^- + [\lambda_2 a_2]_2^-}{w_2^t}, \\ & \ldots, \frac{w_K^t[HH^T + \mu I]_K + [xH^T]_K^- + [\lambda_1 a_1]_K^- + [\lambda_2 a_2]_K^-}{w_K^t}) \end{aligned} \tag{B.16}$$

We can approximate $E(w^t)$ based on Taylor expansion.

$$E(w) = E(w^t) + \nabla E(w^t)(w - w^t)^T + \frac{1}{2}(w - w^t)\nabla^2 E(w^t)(w - w^t)^T \tag{B.17}$$

where, by simple derivation,

$$\begin{aligned} \nabla E(w^t) &= w^t[HH^T + \mu I] + [xH^T]^- + [\lambda_1 a_1]^- + [\lambda_2 a_2]^- - [xH^T]^+ - [\lambda_1 a_1]^+ - [\lambda_2 a_2]^+ \\ \nabla^2 E(w^t) &= HH^T + \mu I \end{aligned} \tag{B.18}$$

Now we can have B.14 satisfied if $F(w^{t+1}, w^t) \geq E(w^{t+1})$. To get it, we have

$$F(w^{t+1}, w^t) - E(w^{t+1}) = \frac{1}{2}(w - w^t)[J(w^t) - \nabla^2 E(w^t)](w - w^t)^T \tag{B.19}$$

Obviously, $\nabla^2 E(w^t) = HH^T + \mu I$ is a positive definite matrix since $\mu$ is positive. Furthermore, the row vector

$$(xH^T)^- + [\lambda_1 a_1]^- + [\lambda_2 a_2]^- \tag{B.20}$$

is always nonnegative. Due to the B.2.1, $J(w^t) - \nabla^2 E(w^t)$ is a semi-positive definite matrix. So that $F(w^{t+1}, w^t) \geq E(w^{t+1})$ is always satisfied and $E(w^t)$ is nonincreasing w.r.t to $t$. Next, we would like to find the optimizer of $F(w, w^t)$. To achieve this, we set $\frac{\partial F(w,w^t)}{\partial w}|_{w=w^{t+1}} = 0$, which is equivalent to equation $\nabla E(w^t) + (w^{t+1} - w^t)J(w^t) = 0$. Therefore, we have

$$w^{t+1} = w^t - J^{-1}(w^t)\nabla E(w^t) \tag{B.21}$$

For each element of $w^{t+1}$, the above formula can be reformulated as

$$\begin{aligned}
w_j^{t+1} &= w_j^t - J^{-1}(w^t)_j \nabla E(w^t)_j \\
&= w_j^t - \frac{w_j^t}{w_j^t[HH^T + \mu I]_j + [xH^T]_j^- + [\lambda_1 a_1]_j^- + [\lambda_2 a_2]_j^-} \\
&\cdot [w^t[HH^T + \mu I] + [xH^T]^- + [\lambda_1 a_1]^- + [\lambda_2 a_2]^- - [xH^T]^+ - [\lambda_1 a_1]^+ - [\lambda_2 a_2]^+] \\
&= w_j^t \frac{[xH^T]_j^+ + \lambda_1[(AS)_i]_j^+ + \lambda_2[(A_0)_i]_j^+}{[w_i(HH^T + \mu I)]_j + [xH^T]_j^- + \lambda_1[a_1]_j^- + \lambda_2[a_2]_j^-}
\end{aligned} \tag{B.22}$$

So far we have proved that the above updating rule can make $E(w^t)$ nonincreasing w.r.t to the $t$. Specifically, we have the lower bound for every element on the diagonal of $J$ matrix.

$$J(w^t)_{jj} = \frac{w_j^t[HH^T + \mu I]_j + [xH^T]_j^- + [\lambda_1 a_1]_j^- + [\lambda_2 a_2]_j^-}{w_j^t} \geq \mu \tag{B.23}$$

Based on B.14 and B.23, we can derive the lower bound of the difference between two loss function values after one update.

$$\begin{aligned}
E(w^t) - E(w^{t+1}) &\geq F(w^t, w^t) - F(w^{t+1}, w^t) \\
&= \nabla F(w^{t+1}, w^t)(w^{t+1} - w^t) + \frac{1}{2}(w^t - w^{t+1})\nabla^2 F(w^{t+1}, w^t)(w^t - w^{t+1})^T \\
&= \frac{1}{2}(w^t - w^{t+1})J(w^t)(w^t - w^{t+1})^T \\
&\geq \frac{\mu}{2} \parallel w^t - w^{t+1} \parallel^2
\end{aligned} \tag{B.24}$$

since $\nabla F(w^{t+1}, w^t) = 0$ and $\nabla^2 F(w^{t+1}, w^t) = J(w^t)$. And by accumulating $\parallel w^t - w^{t+1} \parallel^2$ from 0 to $T-1$, we then can have

$$E(w^0) - E(w^K) \geq \frac{\mu}{2}\Sigma_{t=0}^{T-1} \parallel w^t - w^{t+1} \parallel^2 \tag{B.25}$$

Thus, for any given fixed iteration number $T$, we can have

$$\min_{0\leq t\leq T} \parallel w^t - w^{t+1} \parallel^2 \leq \frac{2}{\mu(T-1)}(E(w^0) - E(w^T)) \leq \frac{2}{\mu(T-1)}E(w^0) \tag{B.26}$$

□

**Theorem B.2.3.** *Consider the following quadratic optimization problem,*

$$\min_{w\geq 0} E(h) = \frac{1}{2} \parallel x - Wh \parallel_F^2 + \frac{\eta}{2} \parallel 1_k^T h - 1_N^T \parallel_2^2 \tag{B.27}$$

*which is the loss function for a column vector $h \in \mathbb{R}^K$. In this optimization problem, $W \in \mathbb{R}^{D\times K}$ is the constant non-negative matrix, and $x \in \mathbb{R}^K$ is a constant column vectors. The penalty parameter, $\eta$ is a positive number. The the following update rule*

$$h_j^{t+1} = h_j^t \frac{[W^T x]_j^+ + \eta 1_k}{[(W^T W + \eta 1_k 1_k^T)h^t]_j + [W^T x]_j^-} \tag{B.28}$$

*converges to its optimal solution with the convergence rate*

$$\min_{t\in[1,K]} \parallel h^{t+1} - h^t \parallel \leq \frac{2}{\eta(K-1)}E(h^0) \tag{B.29}$$

*Proof.* To prove that there exists an $h^{t+1}$ that $E(h^{t+1}) \geq E(h^t)$, we can construct an auxiliary function $F(h, h^t)$, s.t.

$$\begin{gathered} F(h^t, h^t) = E(h^t) \\ F(h, h^t) \geq E(h) \end{gathered} \tag{B.30}$$

Then we have

$$E(h^t) = F(h^t, h^t) \geq F(h^{t+1}, h^t) \geq E(h^{t+1}) \tag{B.31}$$

where $h^{t+1} = \arg\min_h F(h, h^t)$. Thus, the loss function $E(h^t)$ is monotonously non-increasing w.r.t the iteration $t$. We define the auxiliary function as the following.

$$F(h, h^t) = E(h^t) + (h - h^t)^T \nabla E(h^t) + \frac{1}{2}(h - h^t)^T J(h^t)(h - h^t) \tag{B.32}$$

where

$$J(h^t) = \quad \mathrm{diag}(\frac{[(W^TW + \eta 1_K 1_K^T)h^t]_1 + [W^Tx]_1^-}{h_1^t}, \frac{[(W^TW + \eta 1_K 1_K^T)h^t]_2 + [W^Tx]_2^-}{h_2^t}, \dots, \frac{[(W^TW + \eta 1_K 1_K^T)h^t]_K + [W^Tx]_K^-}{h_K^t}) \tag{B.33}$$

We can approximate $E(h^t)$ based on Taylor expansion.

$$E(h) = E(h^t) + (h - h^t)^T \nabla E(h^t) + \frac{1}{2}(h - h^t)^T \nabla^2 E(h^t)(h - h^t) \tag{B.34}$$

where, by simple derivation,

$$\begin{aligned} \nabla E(h^t; W) &= (W^TW + \eta 1_k 1_k^T)h^t + [W^Tx]^- - [W^Tx]^+ - \eta 1_K \\ \nabla^2 E(h^t) &= W^TW + \eta 1_k 1_k^T \end{aligned} \tag{B.35}$$

Now we can have B.31 satisfied if $F(h^{t+1}, h^t) \geq E(h^{t+1})$. To get it, we have

$$F(h^{t+1}, h^t) - E(h^{t+1}) = \frac{1}{2}(h - h^t)^T [J(w^t) - \nabla^2 E(h^t)](h - h^t) \tag{B.36}$$

Obviously, $\nabla^2 E(h^t) = W^TW + \eta 1_k 1_k^T$ is a positive definite matrix since $\eta$ is positive. Furthermore, the row vector

$$(W^Tx)^- \tag{B.37}$$

is always nonnegative. Due to the B.2.1, $J(h^t) - \nabla^2 E(h^t)$ is a semi-positive definite matrix. So that $F(h^{t+1}, h^t) \geq E(h^{t+1})$ is always satisfied and $E(h^t)$ is nonincreasing w.r.t to $t$. Nextly, we would like to find the optimizer of $F(h, h^t)$. To achieve this, we set $\frac{\partial F(h,h^t)}{\partial h}|_{h=h^{t+1}} = 0$, which is equivalent to equation $\nabla E(h^t) + (h^{t+1} - h^t)J(h^t) = 0$. Therefore, we have

$$h^{t+1} = h^t - J^{-1}(h^t)\nabla E(h^t) \tag{B.38}$$

For each element of $h^{t+1}$, the above formula can be reformulated as

$$\begin{aligned} h_j^{t+1} &= h_j^t - J^{-1}(h^t)_j \nabla E(h^t)_j \\ &= h_j^t - \frac{h_j^t}{[(W^TW + \eta 1_K 1_K^T)h^t]_j + [W^Tx]_j^-} \\ &\cdot [(W^TW + \eta 1_k 1_k^T)h^t + [W^Tx]^- - [W^Tx]^+ - \eta 1_K] \\ &= h_j^t \frac{[W^Tx]_j^+ + \eta 1_k}{[(W^TW + \eta 1_k 1_k^T)h^t]_j + [W^Tx]_j^-} \end{aligned} \tag{B.39}$$

So far we have proved that the above updating rule can make $E(h^t)$ nonincreasing w.r.t to the $t$. Specifically, we have the lower bound for every element on the diagonal of $J$ matrix.

$$J(h^t)_{jj} = \frac{[(W^TW + \eta 1_K 1_K^T)h^t]_j + [W^Tx]_j^-}{h_j^t} \geq \eta \tag{B.40}$$

Based on B.31 and B.40, we can derive the lower bound of the difference between two lose function values after one update.

$$\begin{aligned} E(h^t) - E(h^{t+1}) &\geq F(h^t, h^t) - F(h^{t+1}, h^t) \\ &= (h^{t+1} - h^t)^T \nabla F(h^{t+1}, h^t) + \frac{1}{2}(h^t - h^{t+1})^T \nabla^2 F(h^{t+1}, h^t)(h^t - h^{t+1}) \\ &= \frac{1}{2}(h^t - h^{t+1})^T J(h^t)(h^t - h^{t+1}) \\ &\geq \frac{\eta}{2} \parallel h^t - h^{t+1} \parallel^2 \end{aligned} \tag{B.41}$$

since $\nabla F(h^{t+1}, h^t) = 0$ and $\nabla^2 F(h^{t+1}, h^t) = J(h^t)$. And by accumulating $\parallel h^t - h^{t+1} \parallel^2$ from 0 to $T-1$, we have

$$E(h^0) - E(h^K) \geq \frac{\eta}{2} \Sigma_{t=0}^{T-1} \parallel h^t - h^{t+1} \parallel^2 \tag{B.42}$$

Thus, for any given fixed iteration number $T$, we have

$$\min_{0 \leq t \leq T} \parallel h^t - h^{t+1} \parallel^2 \leq \frac{2}{\eta(T-1)}(E(h^0) - E(h^T)) \leq \frac{2}{\eta(T-1)} E(h^0) \tag{B.43}$$

□

## B.3 Rewriting the Loss Function

To conduct the parameter analysis and validate the optimization convergence, we rewrote the original loss function to a equivalent formula as following.

$$
\begin{aligned}
\min_{W\geq 0, H\geq 0} L(W,H) &= \frac{1}{2} \parallel X - WH \parallel_F^2 - \lambda_1 tr(W^T AS) - \lambda_2 tr(W^T A_0) \\
&+ \frac{\mu}{2} \parallel W \parallel_F^2 + \frac{\eta}{2} \parallel 1_k^T H - 1_N^T \parallel_2^2 \\
&= \frac{1}{2} \parallel X - WH \parallel_F^2 + \frac{\mu}{4} tr(WW^T) - \lambda_1 tr(W^T AS) \\
&+ \frac{\mu}{4} tr(WW^T) - \lambda_2 tr(W^T A_0) + \frac{\eta}{2} \parallel 1_k^T H - 1_N^T \parallel_2^2 \\
&= \frac{1}{2} \parallel X - WH \parallel_F^2 + \frac{\mu}{4} tr(WW^T - \frac{4\lambda_1}{\mu}(W^T AS) + \frac{4\lambda_1^2}{\mu^2}(AS)(AS)^T) \\
&+ \frac{\mu}{4} tr(WW^T - \frac{4\lambda_1}{\mu}(W^T A_0) + \frac{4\lambda_1^2}{\mu^2} A_0 A_0^T) + \frac{\eta}{2} \parallel 1_k^T H - 1_N^T \parallel_2^2 + \text{constant} \\
&= \frac{1}{2} \parallel X - WH \parallel_F^2 + \frac{\mu}{4} \parallel W - \frac{2\lambda_1}{\mu}(AS) \parallel_F^2 \\
&+ \frac{\mu}{4} \parallel W - \frac{2\lambda_2}{\mu} A_0 \parallel_F^2 + + \frac{\eta}{2} \parallel 1_k^T H - 1_N^T \parallel_2^2 + \text{constant}
\end{aligned}
\tag{B.44}
$$

# Appendix C

# tranSig optimization algorithm based on SAME

Let $\theta_1 = \{\hat{W}^{(1)}, \ldots, \hat{W}^{(T)}\}$ and $\theta_2 = \{\frac{1}{\tau_e}, \frac{1}{\tau_x}, \frac{1}{\tau_w}, \pi, \alpha, V, \gamma\}$. Due to the reason that it is not always easy to have an accurate estimation on the binary variables in the spike-and-slab model, we grouped our primary interested parameters $v$ and $\gamma$ in $\theta_2$ for multiple estimations in each iteration. The full model can formulated as

$$\begin{aligned} x_{dkc_k}^{(t)} &\sim N(\hat{w}_{dk}^{(t)}, \frac{1}{\tau_{x_d}}) \\ \hat{w}_{dk}^{(t)} &\sim N(v_{dk} \cdot \gamma_{dk}, \frac{1}{\tau_w}) \\ \gamma_{dk} &\sim Ber(\pi) \\ v_{dk} &\sim N(0, \frac{1}{\tau_v}) \end{aligned} \tag{C.1}$$

where $v$ is a continuous variable and $\gamma$ is a bernoulli variable.

To infer SAME, we added prior distributions for $\theta_2$ as:

$$\begin{aligned}\tau_{x_d} &= \text{Gamma}(\alpha_{x_d}, \beta_{x_d}) \\ \tau_w &= \text{Gamma}(\alpha_w, \beta_w) \\ \pi &= \text{Beta}(\alpha_\pi, \beta_\pi)\end{aligned} \tag{C.2}$$

To make the above prior distributions to be noninformative, here we set all the parameters $\alpha = 0.001$ and $\beta = 0.001$.

Then the joint distribution is

$$\begin{aligned}\text{Pr}(Y^{(t_0)}, X, \theta_1, \theta_2) = &\cdot \Pi_t \Pi_d \Pi_k \Pi_{c_k} \text{N}(x_{dkc_k}^{(t)} | \hat{w}_{dk}^{(t)}, \frac{1}{\tau_{x_d}}) \cdot \text{Gamma}(\tau_{x_d} | \alpha_{x_d}, \beta_{x_d}) \\ &\cdot \Pi_t \Pi_d \Pi_k \text{N}(\hat{w}_{dk}^{(t)} | v_{dk}\gamma_{dk}, \frac{1}{\tau_w}) \cdot \text{Gamma}(\tau_w | \alpha_w, \beta_w) \\ &\cdot \Pi_d \Pi_k \text{N}(v_{dk} | 0, \frac{1}{\tau_v}) \cdot \cdot \Pi_k \Pi_n \text{N}(z_{kn} | 0, 1) \\ &\cdot \Pi_d \Pi_k \text{Ber}(\gamma_{dk} | \pi) \cdot \text{Beta}(\pi | \alpha_\pi, \beta_\pi)\end{aligned} \tag{C.3}$$

The log likelihood function can be written as:

$$\begin{aligned}\log p(Y^{(t_0)}, X, \theta_1, \theta_2) = &- \Sigma_t \Sigma_d \Sigma_k \Sigma_{c_k} \frac{\tau_{x_d}}{2} (x_{dkc_k}^{(t)} - \hat{w}_{dk}^{(t)})^2 \\ &- \frac{TDK}{2} \log \frac{2\pi}{\tau_w} - \Sigma_t \Sigma_d \Sigma_k \frac{\tau_w}{2} (\hat{w}_{dk}^{(t)} - v_{dk}\gamma_{dk})^2 \\ &- \frac{DK}{2} \log(\frac{2\pi}{\tau_v}) - \Sigma_d \Sigma_k \frac{\tau_v v_{dk}^2}{2} \\ &- \frac{KN}{2} \log(2\pi) - \Sigma_k \Sigma_n \frac{z_{kn}^2}{2} + \Sigma_d \Sigma_k [\gamma_{dk} \log \pi + (1 - \gamma_{dk}) \log(1 - \pi)]\end{aligned} \tag{C.4}$$

**Sample $\theta_2$ for $\Lambda(i)$**

The first step is $j = 1, \ldots, \Lambda(i)$, $\theta_2(i,j)^{new} \sim p(\theta_2 | y, \theta_1^{prev}(i-1))$.

**Update $\gamma_{\mathbf{dk}}(i,j)$**

We notice that

$$\begin{aligned} \frac{p(\gamma_{dk}=1)(i,j)}{p(\gamma_{dk}=0)(i,j)} &= \frac{\text{Pr}(\hat{w}_{dk}^{(1)}, \ldots, \hat{w}_{dk}^{(T)} | v_{dk}, \gamma_{dk}=1, \frac{1}{\tau_w})\text{Pr}(v_{dk}|\gamma_{dk}=1)\text{Pr}(\gamma_{dk}=1)}{\text{Pr}(\hat{w}_{dk}^{(1)}, \ldots, \hat{w}_{dk}^{(T)} | v_{dk}, \gamma_{dk}=0, \frac{1}{\tau_w})\text{Pr}(v_{dk}|\gamma_{dk}=0)\text{Pr}(\gamma_{dk}=0)} \\ &= \frac{\pi(j) N(\hat{w}_{dk}^{(1)}, \ldots, \hat{w}_{dk}^{(T)} | v_{dk}, \frac{1}{\tau_w(j)})}{(1-\pi(j)) N(\hat{w}_{dk}^{(1)}, \ldots, \hat{w}_{dk}^{(T)} | 0, \frac{1}{\tau_w(j)})} \end{aligned} \tag{C.5}$$

Then we have

$$\begin{aligned} \lambda_{dk}(i,j) =& \log \frac{p(\gamma_{dk}=1)(i,j)}{p(\gamma_{dk}=0)(i,j)} \\ =& \log \frac{\pi(i,j-1)}{1-\pi(i,j-1)} - \frac{\tau_w(i,j-1)}{2} \Sigma_t (\hat{w}_{dk}^{(t)}(i-1) - v_{dk}(i,j-1))^2 \\ &+ \frac{\tau_w(i,j)}{2} \Sigma_t \hat{w}_{dk}^{(t)2}(i-1) \end{aligned} \tag{C.6}$$

Then the Bernoulli distribution $\phi_{dk}(i,j) = p(\gamma_{dk}=1)(i,j) = \frac{1}{1+\exp(-\lambda_{dk}(i,j))}$.

**Update $\mathbf{v}_{\mathbf{dk}}(i,j)$**

The variational distribution of $v_{dk}$ is a mixture Gaussian model depending on $\gamma_{dk}$. When $\gamma_{dk} = 1$, the log likelihood of $q$ distribution is

$$\log p(v_{dk}|\gamma_{dk}=1)(i,j) = -\Sigma_t \frac{\tau_w(i,j-1)}{2} (\hat{w}_{dk}^{(t)}(i-1) - v_{dk}\gamma_{dk}(i,j))^2 - \frac{\tau_v(i,j)v_{dk}^2}{2} + \text{const} \tag{C.7}$$

It is equivalent with

$$\begin{aligned} p(v_{dk}|\gamma_{dk}=1)(i,j) &= N(\mu_{v_{dk}}(i,j), \sigma_{v_{dk}}^2(i,j)) \\ p(v_{dk}|\gamma_{dk}=0)(i,j) &= N(0, \frac{1}{\tau_v}) \end{aligned} \tag{C.8}$$

where

$$\begin{aligned} \sigma^2_{v_{dk}}(i,j) &= (T\tau_w(i,j-1)+\tau_v)^{-1} \\ \mu_{v_{dk}}(i,j) &= \sigma^2_{v_{dk}}(i,j)\tau_w(i,j-1)\Sigma_t \hat{w}^{(t)}_{dk}(i-1) \\ \tau_v^{-1} &= 100 \end{aligned} \tag{C.9}$$

**Update** $\pi(i,j)$

$$\pi(i,j) \sim \text{Beta}(\alpha_\pi + \Pi_d \Pi_k \gamma_{dk}(i,j), \beta_\pi + DK - \Pi_d \Pi_k \gamma_{dk}(i,j)) \tag{C.10}$$

**Update** $\tau_{\mathbf{x_d}}(i,j)$

$$\begin{aligned} p(\tau_{x_d}|Y^{(t_0)}, X, \theta_1^{-\tau_{x_d}}, \theta_2) &\propto p(X,W|\tau_{x_d})p(\tau_{x_d}) \\ &= \Pi_t \Pi_k \Pi_{c_k} \mathbf{N}(x^{(t)}_{dkc_k}|\hat{w}^{(t)}_{dk}, \frac{1}{\tau_{x_d}}) \cdot \text{Gamma}(\tau_{x_d}|\alpha_{x_d}, \beta_{x_d}) \end{aligned} \tag{C.11}$$

$$\tau_{x_d}(i,j) \sim \text{Gamma}(\frac{C_0}{2} + \alpha_{x_d}, \frac{1}{2}\Sigma_t \Sigma_k \Sigma_{c_k}(x^{(t)}_{dkc_k} - \hat{w}^{(t)}_{dk}(i-1))^2 + \beta_{x_d}) \tag{C.12}$$

**Update** $\tau_{\mathbf{w}}(i,j)$

$$\begin{aligned} p(\tau_w|Y^{(t_0)}, X, \theta_1^{-\tau_w}, \theta_2) &\propto p(W,V,\gamma|\tau_w)p(\tau_w) \\ &= \Pi_t \Pi_d \Pi_k \mathbf{N}(\hat{w}^{(t)}_{dk}|v_{dk}\gamma_{dk}, \frac{1}{\tau_w}) \cdot \text{Gamma}(\tau_w|\alpha_w, \beta_w) \end{aligned} \tag{C.13}$$

$$\tau_w(i,j) \sim \text{Gamma}(\frac{TDK}{2} + \alpha_w, \frac{1}{2}\Sigma_t \Sigma_d \Sigma_k(\hat{w}^{(t)}_{dk}(i-1) - v_{dk}(i,j)\gamma_{dk}(i,j))^2 + \beta_w) \tag{C.14}$$

**Sample $\theta_1$**

In the second step is

$$\begin{aligned} \theta_1(i) &\sim q_\gamma(\theta_1|y_1, \theta_2(i,1), \ldots, \theta_2(i,\Lambda(i)))) \\ &\propto \Pi_{j=1}^{\Lambda(i)} p(\theta_1|y, \theta_2(i,j)) \end{aligned} \tag{C.15}$$

We defined $q^*(\theta_1)(i,j)$ to be the $j$th marginal posterior distribution of $\theta_1$ in the $i$th iteration.

**Update $\mathbf{w}_{\mathbf{dk}}^{(\mathbf{t})}(i)$**

$$\log q^*(\hat{w}_{dk}^{(t)})(i,j) = -\frac{\tau_{x_d}(i,j)}{2}\Sigma_{c_k}(x_{dkc_k}^{(t)} - w_{dk}^{(t)})^2 - \frac{\tau_w(i,j)}{2}\left(\hat{w}_{dk}^{(t)} - v_{dk}(i,j)\gamma_{dk}(i,j)\right)^2 + \text{const} \tag{C.16}$$

Thus we have $\hat{w}_{dk}^{(t)}(i) \sim N(\mu_{\hat{w}_{dk}^{(t)}}(i), \tau^{-1}_{\hat{w}_{dk}^{(t)}}(i))$

where

$$\begin{aligned} \tau_{\hat{w}_{dk}^{(t)}}(i) =& \sum_{\Lambda(j)} (C_k \tau_{x_d}(i,j) + \tau_w(i,j)) \\ \mu_{\hat{w}_{dk}^{(t)}}(i) =& \tau^{-1}_{\hat{w}_{dk}^{(t)}}(i) \sum_{\Lambda(j)} \left(\tau_{x_d}(i,j)\Sigma_{c_k} x_{dkc_k}^{(t_0)} + \tau_w(i,j) v_{dk}(i,j)\gamma_{dk}(i,j)\right) \end{aligned} \tag{C.17}$$

# Appendix D

# tranSig Bulk and Single Cell Data Batch Correction

To generate the pseudo-bulk expression profiles for batch-effect correction, the single cell expression profiles of the target tissue are normalized to the TPM space. The compositions of pseudo-bulk profiles are set according to a Normal distribution $N(\mu, \sigma)$ for each cell type, whereas $\mu$ is set to the empirical cell proportion of single cell datasets and $\sigma$ is set to $2\mu$. After randomly drawing the fractions of pseudo-bulk profiles from $N(\mu, \sigma)$, the negative values are set to $0$, and the fractions are normalized across cell types so that the summation of fractions is equal to $1$. Then, according to the normalized fractions, the cells are sampled from the single cell profiles with replacement for $10,000$ times in total. The single cell expression profiles of these cells are averaged as one sample of pseudo-bulk expression profiles. Next, the process is repeated for N times same with the sample size of bulk-seq. Combat is performed on the pseudo-bulk $Y_{sc}$ and bulk RNA-seq $Y_0$ in the subsequent batch correction.

Based on the assumption of Combat, suppose the data contain two batches including $n_i$ samples for each batch $i = 1, 2$, with signature genes $g = 1, 2, \cdots, G$, in which we denote bulk RNA-seq to be $Y^1$ as be the first batch and denote a pseudo mixture to be $Y^2$ as the second batch. The model is assumed

$$Y_{gn}^i = \alpha_g + B_g^i + \delta_g^i \epsilon_{gn}^i$$

where $\alpha_g$ is the overall gene expression, $B_g^i$ and $\delta_j^i$ are the batch- and gene-specific random effects. $\epsilon_{gn}^i$ is normally distributed with mean $0$ and variance $\sigma_g^2$.

First, the data are standardized so that genes have similar overall mean and variance. The standardized data, $Z_{gn}^i$ is calculated by

$$Z_{gn}^i = \frac{Y_{gn}^i - \widehat{\alpha_g}}{\widehat{\sigma}_g}$$

where $\widehat{\alpha}_g$ is the estimation of $\alpha_g$, and $\widehat{\sigma}_g$ is estimated by $\widehat{\sigma}_g^2 = \frac{1}{N}\Sigma_{in}(Y_{gn}^i - \widehat{\alpha}_g - B_g^i)^2$ ($N$ is the total number of samples).

Then the EB parameters are estimated by parametric empirical priors. We assume $Z_{gn}^i \sim N(B_g^i, {\delta_g^i}^2)$, and the prior distributions of the batch effect parameters are assumed as $B_g^i \sim N(Y^i, \tau_i^2)$ and ${\delta_g^i}^2 \sim \mathrm{InverseGamma}(\lambda^i, \theta^i)$.

Based on the assumptions, EB estimates the batch effect parameters by conditional posterior distribution means

$$B_g^{i*} = \frac{n_i \tau_i^2 \widehat{B_g^i} + \delta_g^{i*2} \overline{Y}^i}{n_i \tau_i^2 + \delta_g^{i*2}}$$

and

$${\delta_g^i}^2 = \frac{\overline{\theta}_i + \frac{1}{2}\Sigma_n (Z_{gn}^i - B_g^{i*})^2}{\frac{1}{2}n_i + \lambda_i - 1}$$

After calculating the batch effect estimations, different from the original Combat model, we only adjust bulk RNA-seq $Y^1$ to single cell space$Y^2$. The adjusted mixture can be calculated by

$$Y_{gn}^{1*} = \widehat{\alpha}_g + \widehat{B}_g^{2*} + \widehat{\sigma}_g \frac{\widehat{\delta}_g^{2*}}{\widehat{\delta}_g^{1*}} (Z_{gn}^1 - \widehat{B}_g^{1*})$$

Notably, we use an equal number of samples from bulk RNA-seq and pseudo mixture in real applications.

# Appendix E

# tranSig Gene List

## E.1 tranSig signature gene list

ACTB, ACTG1, AGTRAP, AIF1, ANXA1, ANXA3, ANXA5, AP1S2, APOBEC3A, ARHGDIB, ARL4C, ARPC1B, ARPC2, ARPC3, ASAH1, ATP6V1F, B2M, BANK1, BCL2A1, BIRC3, BLVRB, BTG1, C12orf75, C1orf162, C4orf3, C5AR1, CALM1, CALM2, CALR, CAMP, CAPG, CARD16, CD1C, CD24, CD27, CD36, CD37, CD38, CD3D, CD48, CD53, CD59, CD63, CD74, CD79A, CD79B, CD99, CDA, CDKN1C, CFD, CFL1, CHCHD2, CLEC10A, CLEC12A, CLEC4E, CLEC7A, CMC1, COPE, CORO1A, COTL1, COX4I1, COX5B, COX6A1, COX6C, COX7A2, COX7B, CPVL, CSF3R, CST3, CSTA, CSTB, CTSA, CTSB, CTSC, CTSD, CTSS, CUTA, CUX1, CYBA, CYBB, CYCS, DAD1, DEFA3, DEFA4, DERL3, DNAJA1, DUSP1, DUSP11, DYNLRB1, DYNLT1, EAF2, EDF1, EEF1B2, EEF1D, EEF2, EIF1, EMB, ERH, ERP29, EVI2B, FCER1A, FCER1G, FCGR3A, FCGRT, FCMR, FCN1, FGFBP2, FGL2, FGR, FKBP11, FKBP2, FOLR3, FOS, FPR1, FTL, GAPDH, GCA, GLIPR1, GLRX, GMFG, GNAS, GNG7, GRN, GSTP1, GTF3A, GZMA, GZMB, GZMK, HCK, HERPUD1, HINT1, HLA-A, HLA-DMA, HLA-DMB, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DRA, HLA-DRB1, HLA-DRB5, HM13, HMGN1, HMOX1, HSBP1, HSP90AA1, HSP90B1, HSPA5, IFI44L, IFITM1, IFITM2, IFITM3, IGLL5, IGSF6, ILF2, IRF7, IRF8, ISG15, ISG20, ITM2B, ITM2C, JAML, JCHAIN, KCTD12, KDELR2, KLRB1, KLRD1, LCN2, LCP1, LDHB, LGALS2, LILRB2, LIMD2, LMAN2, LSP1, LST1, LTA4H, LTB, LY6E, LY86, LY96, LYZ, MANF, MCL1, MIF, MNDA, MPEG1, MRPL33, MRPL52, MS4A1, MS4A6A, MS4A7, MT-ATP6, MT-ATP8, MT-CO2, MT-CO3, MT-CYB, MT-ND1, MT-ND2, MYDGF, MYL6, MZB1, NAAA, NACA, NAP1L1, NCF2, NDUFA1, NDUFA11, NDUFA4, NDUFB1, NDUFB4, NDUFB6, NME1, NPC2, NPM1, NUP214, OAZ1, OST4, OSTC, P4HB, PARK7, PCBP1, PDIA4, PDIA6, PEBP1, PECAM1, PFDN5, PFN1, PGK1, PILRA, PIM2, PLAC8, PLBD1, PLD4, PLPP5, POMP, POU2F2, PPIA, PPIB, PPT1, PRDX1, PRDX4, PSAP, PSMA2, PSMA3, PSMB2, PSME2, PYCARD, RABAC1, RACK1, RAN, RETN, RGS2, RHOA, RHOC, RNASE2, RNASE3, RNASE6, RNASET2, RNF130, ROMO1, RPL10, RPL10A, RPL11, RPL12, RPL13, RPL13A, RPL14, RPL15, RPL18, RPL18A, RPL19, RPL21, RPL22, RPL22L1, RPL23, RPL23A, RPL24, RPL27, RPL27A, RPL29, RPL3, RPL30, RPL31, RPL32, RPL34, RPL35, RPL35A, RPL36, RPL36AL, RPL37, RPL37A, RPL39, RPL4, RPL41, RPL5, RPL7, RPL7A, RPL8, RPL9, RPLP0, RPLP1, RPLP2, RPN1, RPN2, RPS10, RPS11, RPS12, RPS14, RPS15, RPS15A, RPS16, RPS17, RPS18, RPS19, RPS2, RPS20, RPS21, RPS23, RPS25, RPS26, RPS27, RPS27A, RPS28, RPS29, RPS3, RPS3A, RPS4X, RPS5, RPS6, RPS7, RPS8, RPS9, RPSA, S100A11, S100A12, S100A6, S100A8, S100A9, S100P, SAMHD1, SAT1, SDCBP, SDF2L1, SEC11C, SEC61B, SEC61G, SEC62, SELL, SERF2, SERPINA1, SERPINF1, SH3BGRL, SH3BGRL3, SLC25A6, SLIRP, SMDT1, SNRPD2, SNRPG, SNU13, SNX3, SOD2, SPCS1, SPCS2, SPCS3, SPI1, SPIB, SRP14, SSR2, SSR3, SSR4, STXBP2, SUB1, SYNGR2, TAGLN2, TALDO1, TIMP1, TKT, TMBIM6, TMED9, TMEM156, TMEM176B, TMEM258, TMSB10, TMSB4X, TNFRSF17, TNFSF10, TNFSF13B, TPI1, TRMT112, TSPO, TUBA1B, TXN, TYMP, UBA52, UBE2J1, UBE2N, UCP2, UFM1, UQCR11, UQCRH, UQCRQ, VAMP8, VCAN, VPREB3, XBP1, YWHAB, ZFP36L2, ZNF706, ABCB4, ABCB9, ACAP1, ACHE, ACP5, ADAM28, ADAMDEC1, ADAMTS3, ADRB2, AIM2, ALOX15, ALOX5, AMPD1, ANGPT4, ANKRD55, APOBEC3G, APOL3, APOL6, AQP9, ARHGAP22, ARRB1, ASGR1, ASGR2, ATP8B4, AZU1, BACH2, BARX2, BCL11B, BCL7A, BEND5, BFSP1, BHLHE41, BLK, BMP2K, BPI, BRAF, BRSK2, BST1, BTNL8, C11orf80, C1orf54, C3AR1, C5AR2, CA8, CASP5, CCDC102B, CCL1, CCL13, CCL14, CCL17, CCL18, CCL19, CCL20, CCL22, CCL23, CCL4, CCL5, CCL7, CCL8, CCND2, CCR10, CCR2, CCR3, CCR5, CCR6, CCR7, CD160, CD180, CD19, CD1A, CD1B, CD1D, CD1E, CD2, CD209, CD22, CD244, CD247, CD28, CD300A, CD33, CD3E, CD3G, CD4, CD40, CD40LG, CD5, CD6, CD68, CD69, CD7, CD70, CD72, CD80, CD86, CD8A, CD8B, CD96, CDC25A, CDH12, CDHR1, CDK6, CEACAM3, CEACAM8, CEMP1, CFP, CHI3L1, CHI3L2, CHST15, CHST7, CLC, CLEC2D, CLEC4A, CLIC2, CMA1, COL8A2, COLQ, CPA3, CR2, CREB5, CRISP3, CRTAM, CRYBB1, CSF1, CSF2, CST7, CTLA4, CTSG, CTSW, CXCL10, CXCL11, CXCL13, CXCL3, CXCL5, CXCL9, CXCR1, CXCR2, CXCR5, CXCR6, CYP27A1, CYP27B1, DACH1, DAPK2, DCSTAMP, DENND5B, DEPDC5, DGKA, DHRS11, DHX58, DPEP2, DPP4, DSC1, DUSP2, EBI3, EFNA5, EGR2, ELANE, EPB41, EPHA1, EPN2, ETS1, ETV3, FAM124B, FAM174B, FASLG, FBXL8, FCER2, FCGR2B, FCGR3B, FCRL2, FES, FFAR2, FLT3LG, FLVCR2, FOSB, FOXP3, FPR2, FPR3, FRK, FRMD4A, FRMD8, FZD2, FZD3, GAL3ST4, GFI1, GGT5, GIPR, GNLY, GPC4, GPR1, GPR171, GPR18, GPR183, GPR19, GPR25, GPR65, GRAP2, GYPE, GZMH, GZMM, HAL, HDC, HESX1, HHEX, HIC1, HK3, HLA-DOB, HNMT, HOXA1, HPGDS, HPSE, HRH1, HSPA6, HTR2B, ICA1, ICOS, IDO1, IFNG, IL12B, IL12RB2, IL17A, IL18R1, IL18RAP, IL1A, IL1B, IL1RL1, IL21, IL26, IL2RA, IL2RB, IL3, IL4, IL4R, IL5, IL5RA, IL7, IL7R, ITK, KCNA3, KCNG2, KIR2DL1, KIR2DL4, KIR2DS4, KIR3DL2, KLRC3, KLRC4, KLRF1, KLRG1, KLRK1, KYNU, LAG3, LAIR2, LAMP3, LAT, LCK, LEF1, LHCGR, LILRA2, LILRA4, LIME1, LRMP, LTA, LY9, MAK, MAN1A1, MANEA, MAP3K13, MAP4K1, MAP4K2, MAP9, MARCO, MAST1, MEFV, MEP1A, MGAM, MICAL3, MMP12, MMP25, MMP9, MROH7, MS4A2, MS4A3, MSC, MXD1, MYB, NAALADL1, NCR3, NFE2, NIPSNAP3B, NKG7, NLRP3, NME8, NOD2, NOX3, NPAS1, NPIPB15, NPL, NR4A3, NTRK1, ORC1, OSM, P2RX1, P2RX5, P2RY10, P2RY13, P2RY14, P2RY2, PADI4, PAQR5, PASK, PBXIP1, PCDHA5, PDCD1, PDCD1LG2, PDE6C, PDK1, PGLYRP1, PIK3IP1, PKD2L2, PLA1A, PLA2G7, PLCH2, PLEKHF1, PLEKHG3, PMCH, PNOC, PPBP, PPFIBP1, PRF1, PRG2, PRR5L, PSG2, PTGDR, PTGER2, PTGIR, PTPRG, QPCT, RAB27B, RALGPS2, RASA3, RASGRP2, RASGRP3, RASSF4, RCAN3, REN, RENBP, REPS2, RGS1, RGS13, RRP12, RRP9, RSAD2, RYR1, S1PR5, SAMSN1, SCN9A, SEC31B, SERGEF, SH2D1A, SIGLEC1, SIK1, SIRPG, SIT1, SKA1, SKAP1, SLAMF1, SLAMF8, SLC12A8, SLC15A3, SLC2A6, SLC7A10, SLCO5A1, SMPD3, SMPDL3B, SOCS1, SP140, SPAG4, SPOCK2, ST3GAL6, ST6GALNAC4, ST8SIA1, STAP1, STEAP4, STXBP6, TBX21, TCF7, TCL1A, TEC, TEP1, TGM5, TLR2, TLR7, TLR8, TMEM255A, TNFAIP6, TNFRSF10C, TNFRSF11A, TNFRSF13B, TNFRSF4, TNFSF14, TNIP3, TPSAB1, TRAF4, TRAT1, TREM1, TREM2, TREML2, TRIB2, TRPM4, TRPM6, TSHR, TTC38, TXK, UBASH3A, UPK3A, VILL, VNN1, VNN2, VNN3, WNT5B, WNT7A, ZAP70, ZBP1, ZBTB10, ZBTB32, ZNF135, ZNF165, ZNF222, ZNF286A, ZNF324, ZNF442, ABCA5, ABCB1, ABHD5, ACSM3, ADAM19, ADAMTS5, ADI1, AGPAT5, ALAS1, ALPL, ANK3, AOC2, APOE, ARID4A, ARNT2, ASRGL1, ATP2A1, ATP2B1, BEX1, BMX, C1QA, C1QB, CA4, CACNA2D3, CALB2, CALML4, CAMK4, CASP1, CD14, CD163, CD207, CDC14B, CDC42EP4, CDK2AP2, CDR2L, CEACAM1, CES1, CFB, CHMP7, CIB2, CIITA, CLCC1, CLCF1, CLEC5A, CNNM1, CNOT1, COL4A3, CP, CRISPLD2, CRLF2, CYP4F3, CYSLTR1, DEFB1, DENND3, DPYD, DUSP4, DYSF, EDN1, EMILIN2, ESPL1, EVL, FARS2, FBLN1, FCHO1, FGFR3, FLT4, FMO5, FST, FSTL1, FUT3, FXYD6, GAS7, GATA2, GBP1, GCH1, GFOD1, GIMAP4, GJB1, GLRX2, GP5, GPNMB, HAGH, HAVCR1, HBD, HGD, HOMER2, HOXA2, HTRA1, IFI27, IFNB1, IFT20, IGFBP2, IL15RA, IL1R2, IL1RAP, IL32, IL6ST, ING2, IRS1, JRKL, KCNJ15, KIF22, KIR3DL1, KLHL18, KRT19, KRT5, KSR1, LAIR1, LILRA5, LILRB1, LIMA1, LIMK2, LRP5L, LRRC8D, LSM4, MAG, MAL, MAOA, MAPK7, MAT2B, MEST, MME, MMP8, MOCS3, MPO, MPPED2, MRPL3, MRPL4, MT1X, MTMR11, MTSS1, MUC1, MYLIP, NAGA, NBN, NBR1, NDRG2, NEFL, NOTCH4, NPEPPS, NR2E3, NR4A2, NRG1, NRGN, NUDT1, NUDT18, NXT1, NXT2, OLFM1, ORM1, OSBPL10, PALLD, PANX1, PAX5, PCGF2, PDGFB, PDK4, PHEX, PI3, PIK3CG, PLAT, POU2AF1, PPA1, PROM1, PRR5, PSAT1, PTPN13, PTPRK, PTPRS, PTTG2, QPRT, RAB9A, RAMP1, RARRES2, RNASE1, RNASE4, RNF122, RRAS, S100B, SCRN1, SDC1, SEC63, SERPINF2, SETBP1, SF3A3, SFTPD, SFXN3, SH3BP2, SIDT1, SIGLEC6, SLC12A3, SLC17A5, SLC1A4, SLC38A1, SLC4A1AP, SLC6A13, SLC7A7, SLC9A3R1, SLCO2B1, SMARCD3, SOCS2, STAB2, SYNE1, TAGLN, TBC1D8, TFEC, TGM3, TLL1, TLR5, TMC6, TMEM9B, TMF1, TNFRSF25, TNNI2, TOMM22, TOMM34, TRAF3IP2, TRAK1, TRIB1, TSPAN7, TUBB6, TULP2, TYRO3, ULK2, WEE1, YTHDF3, ZC3H12A, ZDHHC13, ZNF180, ZNF189, ZNF34, ZNF552, ZNF593

## E.2 AM differentially expressed genes

.

MARCKS, C15orf48, MCEMP1, PLA2G7, SOD2, GPR183, HP, BASP1, EMP1, CTSL, CCL3, FOLR3, SDS, TMEM176B, VCAN, TNFAIP6, ZFP36L1, CCL2, CCL20, NEAT1, IFITM3, CD36, CCL4L2, RNASE1, FNIP2, CCL4, LGMN, TIMP1, HIF1A, CXCL8, CXCL10, MAFB, IER3, SPP1, G0S2, CCL3L1, SGK1, MT1G, AREG, NFKBIA, CXCL2, CCL18, MT1X, MT2A, CXCL3, IL1B, HSPA1B, HSPA1A

# Bibliography

[1] J. Eberwine, J.Y. Sul, T. Bartfai, and J. Kim. The promise of single-cell sequencing. *Nature Methods*, 11(1):25, 2014. 2

[2] X. Dong, L. Zhang, B. Milholland, M. Lee, A.Y. Maslov, T. Wang, and J. Vijg. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nature Methods*, 14(5):491–493, 2017. 2

[3] T. Nagano, Y. Lubling, T.J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E.D. Laue, A. Tanay, and P. Fraser. Single-Cell Hi-C Reveals Cell-to-Cell Variability in Chromosome Structure. *Nature*, 502(7469):59–64, 2013. 2

[4] B. Hwang, J.H. Lee, and D. Bang. Single-cell RNA Sequencing Technologies and Bioinformatics Pipelines. *Experimental & Molecular Medicine*, 50(8):1–14, 2018. 2

[5] M. Labib and S.O. Kelley. Single-cell Analysis Targeting the Proteome. *Nature Reviews Chemistry*, 4(3):143–58, 2020. 2

[6] R. Fang, S. Preissl, Y. Li, X. Hou, J. Lucero, X. Wang, A. Motamedi, and et al. Comprehensive Analysis of Single Cell ATAC-Seq Data with SnapATAC. *Nature Communications*, 12(1):1337, 2021. 2

[7] D.R. Bandura, V.I. Baranov, O.I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J.E. Dick, and S.D. Tanner. Mass Cytometry: Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-of-Flight Mass Spectrometry. *Analytical Chemistry*, 81(16):6813–22, 2009. 2

[8] M.H. Spitzer and G.P. Nolan. Mass Cytometry: Single Cells, Many Features. *Cell*, 165(4):780–91, 2016. 2

[9] O. Ornatsky, D. Bandura, V. Baranov, M. Nitz, M.A. Winnik, and S. Tanner. Highly Multiparametric Analysis by Mass Cytometry. *Journal of Immunological Methods*, 361(1-2):1–20, 2010. 2

[10] A. Regev, S.A. Teichmann, E.S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, and et al. The Human Cell Atlas. *eLife 6*, (December). 2, 31

[11] R.G.H. Lindeboom, A. Regev, and S.A. Teichmann. Towards a Human Cell Atlas: Taking Notes from the Past. *Trends in Genetics: TIG*, (April). 2

[12] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P.K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. Simultaneous Epitope and Transcriptome Measurement in Single Cells. *Nature Methods*, 14(9):865–68, 2017. 2

[13] Z. Zhou, C. Ye, J. Wang, and N.R. Zhang. Surface Protein Imputation from Single Cell Transcriptomes by Deep Neural Networks. *Nature Communications*, 11(1):651, 2020. 2

[14] H. Li, U. Shaham, Stanton K.P., Y. Yao, R.R. Montgomery, and Y. Kluger. Gating Mass Cytometry Data by Deep Learning. *Bioinformatics*, 33(21):3423–30, 2017. 3

[15] S. van Gassen, B. Callebaut, M.J. van Helden, B.N. Lambrecht, P. Demeester, T. Dhaene, and Y. Saeys. FlowSOM: Using Self-Organizing Maps for Visualization and Interpretation of Cytometry Data. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology*, 87(7):636–45, 2015. 3, 12, 18

[16] J.H. Levine, E.F. Simonds, S.C. Bendall, K.L. Davis, E.D. Amir, M.D. Tadmor, O. Litvin, and et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells That Correlate with Prognosis. *Cell*, 162(1):184–97, 2015. 3

[17] D. Aran, A.P. Looney, L. Liu, E. Wu, V. Fong, A. Hsu, S. Chak, and et al. Reference-Based Analysis of Lung Single-Cell Sequencing Reveals a Transitional Profibrotic Macrophage. *Nature Immunology*, 20(2):163–72, 2019. 3, 22

[18] Y. Chen, T. Lakshmikanth, J. Mikes, and P. Brodin. Single-Cell Classification Using Learned Cell Phenotypes. bioRxiv, 2020.07.22.216002, 2020. 3, 14

[19] A. Unterman, T.S. Sumida, N. Nouri, X. Yan, A.Y. Zhao, V. Gasque, J.C. Schupp, and et al. Single-Cell Omics Reveals Dyssynchrony of the Innate and Adaptive Immune System in Progressive COVID-19. medRxiv, 2020.07.16.20153437, 2020. 4, 14, 22

[20] A. Ramaswamy, N.N. Brodsky, T.S. Sumida, M. Comi, H. Asashima, K.B. Hoehn, N. Li, and et al. Immune Dysregulation and Autoreactivity Correlate with Disease Severity in SARS-CoV-2-Associated Multisystem Inflammatory Syndrome in Children. *Immunity*, 54(5):1083–95.e7, 2021. 4, 13, 14, 22

[21] J. Wang, D. Agarwal, M. Huang, G. Hu, Z. Zhou, C. Ye, and N.R. Zhang. Data Denoising with Transfer Learning in Single-Cell Transcriptomics. *Nature Methods*, 16(9):875–78, 2019. 4, 22

[22] J. Zhang, L. Wei, X. Feng, Z. Ma, and Y. Wang. Pattern Expression Nonnegative Matrix Factorization: Algorithm and Applications to Blind Source Separation. *Computational Intelligence and Neuroscience*, (168769):1687–5265, 2008. 5, 8

[23] S. Wu and J. Wang. Nonnegative Matrix Factorization: When Data Is Not Nonnegative. *7th International Conference on Biomedical Engineering and Informatics. IEEE.*, 2014. 5

[24] B. Bodenmiller, E.R. Zunder, R. Finck, T.J. Chen, E.S. Savig, R.V. Bruggner, E.F. Simonds, and et al. Multiplexed Mass Cytometry Profiling of Cellular States Perturbed by Small-Molecule Regulators. *Nature Biotechnology*, 30(9):858–67, 2012. 12

[25] M. Nowicka, C. Krieg, H.L. Crowell, L.M. Weber, F.J. Hartmann, S. Guglietta, B. Becher, M.P. Levesque, and M.D. Robinson. CyTOF Workflow: Differential Discovery in High-Throughput High-Dimensional Cytometry Datasets. *F1000Research*, 6(May):748, 2017. 12, 22

[26] L.M. Weber, M. Nowicka, C. Soneson, and M.D. Robinson. Diffcyt: Differential Discovery in High-Dimensional Cytometry via High-Resolution Clustering. *Communications Biology*, 2(May):183, 2019. 13

[27] X. Wang, Z. Sun, Y. Zhang, Z. Xu, H. Xin, H. Huang, R.H. Duerr, K. Chen, Y. Ding, and W. Chen. BREM-SC: A Bayesian Random Effects Mixture Model for Joint Clustering Single Cell Multi-Omics Data. *Nucleic Acids Research*, 48(11):5814–24, 2020. 14, 22

[28] L. Rodriguez, P.T. Pekkarinen, T. Lakshmikanth, Z. Tan, C.R. Consiglio, C. Pou, Y. Chen, and et al. Systems-Level Immunomonitoring from Acute to Recovery Phase of Severe COVID-19. *Cell Reports. Medicine*, 1(5):100078, 2020. 14

[29] A.C. Aschenbrenner, M. Mouktaroudi, B. Krämer, M. Oestreich, N. Antonakos, M. Nuesch-Germano, K. Gkizeli, and et al. Disease Severity-Specific Neutrophil Signatures in Blood Transcriptomes Stratify COVID-19 Patients. *Genome Medicine*, 13(1):7, 2021. 15

[30] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, and et al. Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China. *The Lancet*, 395(10223):497–506, 2020. 15

[31] E.Z. Ong, Y.F.Z. Chan, W.Y. Leong, N.M.Y. Lee, S. Kalimuddin, S.M.H. Mohideen, K.S. Chan, and et al. A Dynamic Immune Response Shapes COVID-19 Progression. *Cell Host & Microbe*, 27(6):879–82.e2, 2020. 15

[32] Y. Su, D. Chen, D. Yuan, C. Lausted, J. Choi, C.L. Dai, V. Voillet, and et al. Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19. *Cell*, 183(6):1479–95.e20, 2020. 15

[33] J. Hadjadj, N. Yatim, L. Barnabei, A. Corneau, J. Boussier, N. Smith, H. Péré, and et al. Impaired Type I Interferon Activity and Inflammatory Responses in Severe COVID-19 Patients. *Science*, 369(6504):718–24, 2020. 15

[34] N. Fathi and N. Rezaei. Lymphopenia in COVID-19: Therapeutic Opportunities. *Cell Biology International*, 44(9):1792–97, 2020. 16

[35] S. Tavakolpour, T. Rakhshandehroo, E.X. Wei, and M. Rashidian. Lymphopenia during the COVID-19 Infection: What It Shows and What Can Be Learned. *Immunology Letters*, 225(4), 2020. 16

[36] I. Huang and R. Pranata. Lymphopenia in Severe Coronavirus Disease-2019 (COVID-19): Systematic Review and Meta-Analysis. *Journal of Intensive Care Medicine*, 8(May):36, 2020. 16

[37] P. Qiu, E.F. Simonds, S.C. Bendall, Bruggner R.V. Gibbs, K., M.D. Linderman, K. Sachs, G.P. Nolan, and S.K. Plevritis. Extracting a Cellular Hierarchy from High-Dimensional Cytometry Data with SPADE. *Nature Biotechnology*, 29(10):886–91. 18

[38] L. Scrucca, M. Fop, T.B. Murphy, and A.E. Raftery. Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, 8(1):289–317, 2016. 22

[39] N.X. Vinh, J. Epps, and J. Bailey. Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary? *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009. 22

[40] M. Büttner, Z. Miao, F.A. Wolf, S.A. Teichmann, and F.J. Theis. A Test Metric for Assessing Single-Cell RNA-Seq Batch Correction. *Nature Methods*, 16(1):43–49, 2018. 22

[41] L.M. Weber and C. Soneson. HDCytoData: Collection of High-Dimensional Cytometry Benchmark Datasets in Bioconductor Object Formats. *F1000Research*, 8(August):1459, 2019. 22

[42] Y. Hao, S. Hao, E. Andersen-Nissen, W.M. Mauck, S. Zheng, A. Butler, M.J. Lee, and et al. Integrated Analysis of Multimodal Single-Cell Data. bioRxiv, 2020.10.12.335331, 2020. 22

[43] K. O'Neill, N. Aghaeepour, J. Spidlen, and R. Brinkman. Flow Cytometry Bioinformatics. *PLoS Computational Biology*, 9(12):e1003365, 2013. 30

[44] E. Lugli, M. Roederer, and A. Cossarizza. Data Analysis in Flow Cytometry: The Future Just Started. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology*, 77(7):705–13, 2010. 30

[45] J.V. Watson. *Introduction to Flow Cytometry*. Cambridge University Press, 2004. 30

[46] J.A. Ramos-Vara and M.A. Miller. When Tissue Antigens and Antibodies Get Along: Revisiting the Technical Aspects of Immunohistochemistry—The Red, Brown, and Blue Technique. *Veterinary Pathology*, 51(1):42–87, 2014. 30

[47] I.B. Buchwalow and W. Bocker. *Immunohistochemistry: Basics and Methods*. Berlin, Germany: Springer, 2010th edition, 2010. 30

[48] E.A. Madissoon, A. Wilbrey-Clark, R.J. Miragaia, K. Saeb-Parsy, K.T. Mahbubani, N. Georgakopoulos, P. Harding, and et al. scRNA-Seq Assessment of the Human Lung, Spleen, and Esophagus Tissue Stability after Cold Preservation. *Genome Biology*, 21(1):1, 2019. 30

[49] F. A. Cobos, J. Alquicira-Hernandez, J.E. Powell, P. Mestdagh, and K. De Preter. Benchmarking of Cell Type Deconvolution Pipelines for Transcriptomics Data. *Nature Communications*, 11(1):1–14, 2020. 30

[50] A. M. Newman, C.L. Liu, M.R. Green, A.J. Gentles, W. Feng, Y. Xu, C.D. Hoang, M. Diehn, and A.A. Alizadeh. Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nature Methods*, 12(5):453–57, 2015. 30, 31, 34

[51] F. Vallania, A. Tam, S. Lofgren, S. Schaffert, T.J. Azad, E. Bongen, W. Haynes, and et al. Leveraging Heterogeneity across Multiple Datasets Increases Cell-Mixture Deconvolution Accuracy and Reduces Biological and Technical Biases. *Nature Communications*, 9(1):4735, 2018. 30

[52] A. Bezginov, G.W. Clark, R.L. Charlebois, V. Dar, and Tillier E.R.M. Coevolution Reveals a Network of Human Proteins Originating with Multicellularity. *Molecular Biology and Evolutio*, 30(2):332–46, 2013. 30

[53] M. Lukk, M. Kapushesky, J. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, and A. Brazma. A Global Map of Human Gene Expression. *Nature Biotechnology*, 28(4):322–24, 2010. 30

[54] L. Lahti, A. Torrente, L.L. Elo, A. Brazma, and J. Rung. A Fully Scalable Online Pre-Processing Algorithm for Short Oligonucleotide Microarray Atlases. *Nucleic Acids Research*, 41(10):e110, 2013. 30

[55] L. Zappia, B. Phipson, and A. Oshlack. Exploring the Single-Cell RNA-Seq Analysis Landscape with the scRNA-Tools Database. *PLoS Computational Biology*, 14(6):e1006245, 2018. 30

[56] A.S. Venteicher, I. Tirosh, C. Hebert, K. Yizhak, C. Neftel, M.G. Filbin, V. Hovestadt, and et al. Decoupling Genetics, Lineages, and Microenvironment in IDH-Mutant Gliomas by Single-Cell RNA-Seq. *Science*, 355(6332), 2017. 30

[57] M. Slyper, C.B.M. Porter, O. Ashenberg, J. Waldman, E. Drokhlyansky, I. Wakiro, C. Smillie, and et al. Author Correction: A Single-Cell and Single-Nucleus RNA-Seq Toolbox for Fresh and Frozen Human Tumors. *Nature Medicine*, 26(8):1307, 2020. 30

[58] L. Chu, N. Leng, J. Zhang, Z. Hou, D. Mamott, D.T. Vereide, J. Choi, C. Kendziorski, R. Stewart, and J.A. Thomson. Single-Cell RNA-Seq Reveals Novel Regulators of Human Embryonic Stem Cell Differentiation to Definitive Endoderm. *Genome Biology*, 17(1):173, 2016. 30

[59] R.M. Harland. A New View of Embryo Development and Regeneration. *Science*, 2018. 30

[60] A.C. Boroughs, R.C. Larson, N.D. Marjanovic, K. Gosik, A.P. Castano, C.B.M. Porter, S.J. Lorrey, and et al. A Distinct Transcriptional Program in Human CAR T Cells Bearing the 4-1BB Signaling Domain Revealed by scRNA-Seq. *Molecular Therapy: The Journal of the American Society of Gene Therapy*, 28(12):2577–92, 2020. 30

[61] M. Lavaert, K.L. Liang, N. Vandamme, J. Park, J. Roels, M.S. Kowalczyk, B. Li, and et al. Integrated scRNA-Seq Identifies Human Postnatal Thymus Seeding Progenitors and Regulatory Dynamics of Differentiating Immature Thymocytes. *Immunity*, 52(6):1088–1104.e6, 2020. 30

[62] O. Rozenblatt-Rosen, M.J.T. Stubbington, A. Regev, and S.A. Teichmann. The Human Cell Atlas: From Vision to Reality. *Nature*, 550(7677):451–53, 2017. 31

[63] X. Han, Z. Zhou, L. Fei, H. Sun, R. Wang, Y. Chen, H. Chen, and et al. Construction of a Human Cell Landscape at Single-Cell Level. *Nature*, 581(7808):303–9, 2020. 31, 32, 57, 58

[64] X. Wang, J. Park, K. Susztak, N.R. Zhang, and M. Li. Bulk Tissue Cell Type Deconvolution with Multi-Subject Single-Cell Expression Reference. *Nature Communications*, 10(1):1–9, 2019. 31, 34

[65] P.A. Szabo, M. Miron, and D.L. Farber. Location, Location, Location: Tissue Resident Memory T Cells in Mice and Humans. *Science Immunology*, 4(34), 2019. 31

[66] W. Meng, B. Zhang, G.W. Schwartz, A.M. Rosenfeld, D. Ren, J.J.C. Thome, D.J. Carpenter, and et al. An Atlas of B-Cell Clonal Distribution in the Human Body. *Nature Biotechnology*, 35(9):879–84, 2017. 31

[67] J. Liu, C. Gao, J. Sodicoff, V. Kozareva, E.Z. Macosko, and J.D. Welch. Jointly Defining Cell Types from Multiple Single-Cell Datasets Using LIGER. *Nature Protocols*, 15(11):3632–62, 2020. 32, 39, 42

[68] J.D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, and E.Z. Macosko. Single-Cell Multi-Omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 177(7):1873–87.e17, 2019. 32, 39

[69] A. Doucet, S.J. Godsill, and C.P. Robert. Marginal Maximum a Posteriori Estimation Using Markov Chain Monte Carlo. *Statistics and Computing*, 12(1):77–84, 2002. 33, 39, 44

[70] A.M. Newman, C.B. Steen, C.L. Liu, A.J. gentles, A.A. Chaudhuri, F. Scherer, M.S. Khodadoust, and et al. Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry. *Nature Biotechnology*, 37(7):773–82, 2019. 33, 36, 39, 48

[71] K.D. Hansen, R.A. Irizarry, and Z. Wu. Removing Technical Variability in RNA-Seq Data Using Conditional Quantile Normalization. *Biostatistics*, 13(2):204–16, 2012. 33

[72] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, 65(4):631–43.e4, 2017. 34

[73] Y. Li, P. Ren, A. Dawson, H.G. Vasquez, W. Ageedi, C. Zhang, W. Luo, and et al. Single-Cell Transcriptome Analysis Reveals Dynamic Cell Populations and Differential Gene Expression Patterns in Control and Aneurysmal Human Aortic Tissue. *Circulation*, 142(14):1374–88, 2020. 34, 38, 48

[74] T. Gong, N. Hartmann, I.S. Kohane, V. Brinkmann, F. Staedtler, M. Letzkus, S. Bongiovanni, and J.D. Szustakowski. Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples. *PloS One*, 6(11):e27156, 2011. 36, 40

[75] M. Vukmirovic, X. Yan, K.F. Gibson, M. Gulati, J.C. Schupp, G. DeIuliis, T.S. Adams, and et al. Transcriptomics of Bronchoalveolar Lavage Cells Identifies New Molecular Endotypes of Sarcoidosis. *The European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology*, 2021. 37, 38, 48

[76] V.I. Patel and J.P. Metcalf. Airway Macrophage and Dendritic Cell Subsets in the Resting Human Lung. *Critical Reviews in Immunology*, 38(4):303–31, 2018. 37

[77] G. Hu and J.W. Christman. Editorial: Alveolar Macrophages in Lung Inflammation and Resolution. *Frontiers in Immunology*, 10(September):2275, 2019. 37

[78] P. Chen, L. Qin, G. Li, J. Malagon-Lopez, Z. Wang, S. Bergaya, S. Gujja, and et al. Smooth Muscle Cell Reprogramming in Aortic Aneurysms. *Cell Stem Cell*, 26(4):542–57.e11, 2020. 38, 48

[79] N. Sakalihasan, R. Limet, and O.D. Defawe. Abdominal Aortic Aneurysm. *The Lancet*, 365(9470):1577–89, 2005. 38

[80] C.B. Ernst. Abdominal Aortic Aneurysm. *The New England Journal of Medicine*, 328(16):1167–72, 1993. 38

[81] K. Shimizu, R.N. Mitchell, and P. Libby. Inflammation and Cellular Immune Responses in Abdominal Aortic Aneurysms. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 26(5):987–94, 2006. 38

[82] D.L. Rateri, F.M. Davis, A. Balakrishnan, D.A. Howatt, J.J. Moorleghen, W.N. O'Connor, R. Charnigo, L.A. Cassis, and A. Daugherty. Angiotensin II Induces Region-Specific Medial Disruption during Evolution of Ascending Aortic Aneurysms. *The American Journal of Pathology*, 184(9):2586–95, 2014. 38

[83] R.A. Quintana and W.R. Taylor. Cellular Mechanisms of Aortic Aneurysm Formation. *Circulation Research*, 124(4):607–18, 2019. 38

[84] J.A. Curci, S. Liao, M.D. Huffman, S.D. Shapiro, and R.W. Thompson. Expression and Localization of Macrophage Elastase (matrix Metalloproteinase-12) in Abdominal Aortic Aneurysms. *The Journal of Clinical Investigation*, 102(11):1900–1910, 1998. 38

[85] J. Raffort, F. Lareyre, M. Clément, R. Hassen-Khodja, G. Chinetti, and Z. Mallat. Monocytes and Macrophages in Abdominal Aortic Aneurysm. *Nature Reviews. Cardiology*, 14(8):457–71, 2017. 38

[86] W. Xiong, Y. Zhao, A. Prall, T.C. Greiner, and B.T. Baxter. Key Roles of CD4+ T Cells and IFN-$\gamma$ in the Development of Abdominal Aortic Aneurysms in a Murine Model. *The Journal of Immunology*, 172(4):2607–12, 2004. 38

[87] H. Ait-Oufella, Y. Wang, O. Herbin, S. Bourcier, S. Potteaux, J. Joffre, X. Loyer, and et al. Natural Regulatory T Cells Limit Angiotensin II-Induced Aneurysm Formation and Rupture in Mice. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 33(10):2374–79, 2013. 38

[88] M. Fanjul-Fernández, A.R. Folgueras, S. Cabrera, and C. López-Otín. Matrix Metalloproteinases: Evolution, Gene Regulation and Functional Analysis in Mouse Models. *Biochimica et Biophysica Acta*, 1803(1):3–19, 2010. 38

[89] W.E. Johnson, C. Li, and A. Rabinovic. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics*, 8(1):118–27, 2007. 39

[90] Y. Zhang, D.F. Jenkins, S. Manimaran, and W.E. Johnson. Alternative Empirical Bayes Models for Adjusting for Batch Effects in Genomic Studies. *BMC Bioinformatics*, 19(1):262, 2018. 39

[91] X. Han, R. Wang, Y. Zhou, L. Fei, H. Sun, S. Lai, A. Saadatpour, and et al. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, 173(5):1307, 2018. 40

[92] J. Ding, X. Adiconis, S.K. Simmons, M.S. Kowalczyk, C.C. Hession, N.D. Marjanovic, T.K. Hughes, and et al. Systematic Comparative Analysis of Single Cell RNA-Sequencing Methods. bioRxiv, https://doi.org/10.1101/632216, 2019. 40

[93] G. Chen, B. Ning, and T. Shi. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Frontiers in Genetics*, 10(April):317, 2019. 41

[94] D. Lähnemann, J. Köster, E. Szczurek, D.J. McCarthy, S.C. Hicks, M.D. Robinson, C.A. Vallejos, and et al. Eleven Grand Challenges in Single-Cell Data Science. *Genome Biology*, 21(1):31, 2020. 41

[95] S.C. Hicks, F.W. Townes, M. Teng, and R.A. Irizarry. Missing Data and Technical Variability in Single-Cell RNA-Sequencing Experiments. *Biostatistics*, 19(4):562–78, 2018. 41

[96] R. Bacher and C. Kendziorski. Design and Computational Analysis of Single-Cell RNA-Sequencing Experiments. *Genome Biology*, 17(April):63, 2016. 41

[97] K.D. Korthauer, L. Chu, M.A. Newton, Y. Li, J. Thomson, R. Stewart, and C. Kendziorski. A Statistical Approach for Identifying Differential Distributions in Single-Cell RNA-Seq Experiments. *Genome Biology*, 17(1):222, 2016. 41

[98] D.R. Moller, L.L. Koth, L.A. Maier, A. Morris, W. Drake, M. Rossman, J.K. Leader, and et al. Rationale and Design of the Genomic Research in Alpha-1 Antitrypsin Deficiency and Sarcoidosis (GRADS) Study. Sarcoidosis Protocol. *Annals of the American Thoracic Society*, 12(10):1561–71, 2015. 48

[99] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015. 56

[100] Brian W Kunkle, Benjamin Grenier-Boley, Rebecca Sims, Joshua C Bis, Vincent Damotte, Adam C Naj, Anne Boland, Maria Vronskaya, Sven J Van Der Lee, Alexandre Amlie-Wolf, et al. Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates a$\beta$, tau, immunity and lipid processing. *Nature genetics*, 51(3):414–430, 2019. 56

[101] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, et al. Schizophrenia working group of the psychiatric genomics c, et al. ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*, 47(3):291–5, 2015. 56

[102] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015. 56

[103] Tian Ge, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature communications*, 10(1):1–10, 2019. 56, 60

[104] Yiming Hu, Qiongshi Lu, Ryan Powles, Xinwei Yao, Can Yang, Fang Fang, Xinran Xu, and Hongyu Zhao. Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS computational biology*, 13(6):e1005589, 2017. 56

[105] Esben Agerbo, Patrick F Sullivan, Bjarni J Vilhjalmsson, Carsten B Pedersen, Ole Mors, Anders D Børglum, David M Hougaard, Mads V Hollegaard, Sandra Meier, Manuel Mattheisen, et al. Polygenic risk score, parental socioeconomic status, family history of psychiatric disorders, and the risk for schizophrenia: a danish population-based study and meta-analysis. *JAMA psychiatry*, 72(7):635–641, 2015. 56

[106] Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P Tyrer, Ting-Huei Chen, Qin Wang, Manjeet K Bolla, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics*, 104(1):21–34, 2019. 56

[107] Pradeep Natarajan, Robin Young, Nathan O Stitziel, Sandosh Padmanabhan, Usman Baber, Roxana Mehran, Samantha Sartori, Valentin Fuster, Dermot F Reilly, Adam Butterworth, et al. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation*, 135(22):2091–2101, 2017. 56

[108] Daniel I Chasman, Franco Giulianini, Olga V Demler, and Miriam S Udler. Pleiotropy-based decomposition of genetic risk scores: Association and interaction analysis for type 2 diabetes and cad. *The American Journal of Human Genetics*, 106(5):646–658, 2020. 57

[109] Matthew Aguirre, Yosuke Tanigawa, Guhan Ram Venkataraman, Rob Tibshirani, Trevor Hastie, and Manuel A Rivas. Polygenic risk modeling with latent trait-related genetic components. *European Journal of Human Genetics*, pages 1–11, 2021. 57

[110] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018. 57

[111] Koen Van den Berge, Hector Roux De Bezieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature communications*, 11(1):1–13, 2020. 57

[112] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020. 57

[113] Tianyu Wang, Boyang Li, Craig E Nelson, and Sheida Nabavi. Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data. *BMC bioinformatics*, 20(1):1–16, 2019. 57

[114] Xingjie Hao, Kai Wang, Chengguqiu Dai, Zeyang Ding, Wei Yang, Chaolong Wang, and Shanshan Cheng. Integrative analysis of scrna-seq and gwas data pinpoints periportal hepatocytes as the relevant liver cell types for blood lipids. *Human Molecular Genetics*, 29(18):3145–3153, 2020. 57

[115] Wei Liu, Wenxuan Deng, Ming Chen, Zihan Dong, Biqing Zhu, Zhaolong Yu, Daiwei Tang, Maor Sauler, Louise V Wain, Michael Cho, et al. A statistical framework to identify cell types whose genetically regulated proportions are associated with complex diseases. *medRxiv*, 2021. 57

[116] Arunabha Majumdar, Claudia Giambartolomei, Na Cai, Tanushree Haldar, Tommer Schwarz, Michael Gandal, Jonathan Flint, and Bogdan Pasaniuc. Leveraging eqtls to identify individual-level tissue of interest for a complex trait. *PLoS computational biology*, 17(5):e1008915, 2021. 58

[117] Jason M Torres, Eric R Gamazon, Esteban J Parra, Jennifer E Below, Adan Valladares-Salgado, Niels Wacher, Miguel Cruz, Craig L Hanis, and Nancy J Cox. Cross-tissue and tissue-specific eqtls: partitioning the heritability of a complex trait. *The American Journal of Human Genetics*, 95(5):521–534, 2014. 58

[118] Tune H Pers, Juha M Karjalainen, Yingleong Chan, Harm-Jan Westra, Andrew R Wood, Jian Yang, Julian C Lui, Sailaja Vedantam, Stefan Gustafsson, Tonu Esko, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature communications*, 6(1):1–9, 2015. 58

[119] Ruo-Han Hao, Tie-Lin Yang, Yu Rong, Shi Yao, Shan-Shan Dong, Hao Chen, and Yan Guo. Gene expression profiles indicate tissue-specific obesity regulation changes and strong obesity relevant tissues. *International Journal of Obesity*, 42(3):363–369, 2018. 58

[120] Halit Ongen, Andrew A Brown, Olivier Delaneau, Nikolaos I Panousis, Alexandra C Nica, and Emmanouil T Dermitzakis. Estimating the causal tissues for complex traits and diseases. *Nature genetics*, 49(12):1676–1683, 2017. 58

[121] Diego Calderon, Anand Bhaskar, David A Knowles, David Golan, Towfique Raj, Audrey Q Fu, and Jonathan K Pritchard. Inferring relevant cell types for complex traits by using single-cell gene expression. *The American Journal of Human Genetics*, 101(5):686–699, 2017. 58

[122] Kamil Slowikowski, Xinli Hu, and Soumya Raychaudhuri. Snpsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics*, 30(17):2496–2497, 2014. 58

[123] Y. Ma, S. Sun, X. Shang, E.T. Keller, M. Chen, and X. Zhou. Integrative differential expression and gene set enrichment analysis using summary statistics for scrna-seq studies. *Nature communications*, 11(1):1–13, 2020. 59

[124] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A.K. Shalek, C.K. Slichter, H.W. Miller, M.J. McElrath, M. Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):1–13, 2015. 60

[125] M.I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014. 60

[126] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007. 60

[127] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228–1235, 2015. 61

[128] KL Huang, E Marcora, AA Pimenova, AF Di Narzo, M Kapoor, SC Jin, O Harari, S Bertelsen, BP Fairfax, J Czajkowski, et al. International genomics of alzheimer's project. *Alzheimer's Disease Neuroimaging Initiative, Sims R, Escott-Price V, Mayeux R, Haines JL, Farrer LA, Pericak-Vance MA, Lambert JC, van Duijn C, Launer L, Seshadri S, Williams J, Amouyel P, Schellenberg GD, Zhang B, Borecki I, Kauwe JSK, Cruchaga C, Hao K, Goate AM*, pages 1052–1061, 2017. 63

[129] David Gate, Naresha Saligrama, Olivia Leventhal, Andrew C Yang, Michael S Unger, Jinte Middeldorp, Kelly Chen, Benoit Lehallier, Divya Channappa, B Mark, et al. Clonally expanded cd8 t cells patrol the cerebrospinal fluid in alzheimer's disease. *Nature*, 577(7790):399–404, 2020. 63

[130] Alon Monsonego, Victor Zota, Arnon Karni, Jeffery I Krieger, Amit Bar-Or, Gal Bitan, Andrew E Budson, Reisa Sperling, Dennis J Selkoe, Howard L Weiner, et al. Increased t cell reactivity to amyloid $\beta$ protein in older humans and patients with alzheimer disease. *The Journal of clinical investigation*, 112(3):415–422, 2003. 63

[131] Terrence Town, Jun Tan, Richard A Flavell, and Mike Mullan. T-cells in alzheimer's disease. *Neuromolecular medicine*, 7(3):255–264, 2005. 63

[132] PL McGeer, H Akiyama, S Itagaki, and EG McGeer. Immune system response in alzheimer's disease. *Canadian Journal of Neurological Sciences*, 16(S4):516–527, 1989. 64

[133] Jean C Cruz Hernandez, Oliver Bracko, Calvin J Kersbergen, Victorine Muse, Mohammad Haft-Javaherian, Maxime Berg, Laibaik Park, Lindsay K Vinarcsik, Iryna Ivasyk, Daniel A Rivera, et al. Neutrophil adhesion in brain capillaries reduces cortical blood flow and impairs memory function in alzheimer's disease mouse models. *Nature neuroscience*, 22(3):413–420, 2019. 64

[134] Ki Kim, Xin Wang, Emeline Ragonnaud, Monica Bodogai, Tomer Illouz, Marisa DeLuca, Ross A McDevitt, Fedor Gusev, Eitan Okun, Evgeny Rogaev, et al. Therapeutic b-cell depletion reverses progression of alzheimer's disease. *Nature communications*, 12(1):1–11, 2021. 64

[135] Haichun Yang, Agnes B Fogo, and Valentina Kon. Kidneys: key modulators of hdl levels and function. *Current opinion in nephrology and hypertension*, 25(3):174, 2016. 64, 73

[136] Jacek Rysz, Anna Gluba-Brzózka, Magdalena Rysz-Górzyńska, and Beata Franczyk. The role and function of hdl in patients with chronic kidney disease and the risk of cardiovascular disease. *International journal of molecular sciences*, 21(2):601, 2020. 64, 73

[137] Airlie Cameron, Miles J Schwartz, Richard A Kronmal, and Andrzej S Kosinski. Prevalence and significance of atrial fibrillation in coronary artery disease (cass registry). *The American journal of cardiology*, 61(10):714–717, 1988. 64

[138] Michael S Lauer, Kim A Eagle, Mortimer J Buckley, and Roman W DeSanctis. Atrial fibrillation following coronary artery bypass surgery. *Progress in cardiovascular diseases*, 31(5):367–378, 1989. 64

[139] Umberto Benedetto, Mario F Gaudino, Arnaldo Dimagli, Stephen Gerry, Alastair Gray, Belinda Lees, Marcus Flather, and David P Taggart. Postoperative atrial fibrillation and long-term risk of stroke after isolated coronary artery bypass graft surgery. *Circulation*, 142(14):1320–1329, 2020. 64

[140] Ewelina Michniewicz, Elżbieta Mlodawska, Paulina Lopatowska, Anna Tomaszuk-Kazberuk, and Jolanta Malyszko. Patients with atrial fibrillation and coronary artery disease–double trouble. *Advances in medical sciences*, 63(1):30–35, 2018. 64

[141] Manjari Devidi, Avanija Buddam, Sunil Dacha, and D Sudhaker Rao. Atrial fibrillation and its association with endocrine disorders. *Journal of atrial fibrillation*, 6(5), 2014. 64

[142] Abraham A Embi and Benjamin J Scherlag. An endocrine hypothesis for the genesis of atrial fibrillation: the hypothalamic-pituitary-adrenal axis response to stress and glycogen accumulation in atrial tissues. *North American journal of medical sciences*, 6(11):586, 2014. 64

[143] Raffaele De Caterina, Ana Ruigómez, and Luís Alberto García Rodríguez. Long-term use of anti-inflammatory drugs and risk of atrial fibrillation. *Archives of internal medicine*, 170(16):1450–1455, 2010. 64

[144] Cornelis S Van Der Hooft, Jan Heeringa, Guy G Brusselle, Albert Hofman, Jacqueline CM Witteman, J Herre Kingma, Miriam CJM Sturkenboom, and Bruno H Ch Stricker. Corticosteroids and the risk of atrial fibrillation. *Archives of internal medicine*, 166(9):1016–1020, 2006. 64

[145] Duncan Howie, Annemieke Ten Bokum, Andra Stefania Necula, Stephen Paul Cobbold, and Herman Waldmann. The role of lipid metabolism in t lymphocyte differentiation and survival. *Frontiers in immunology*, 8:1949, 2018. 65

[146] Andreas Bietz, Hengyu Zhu, Manman Xue, and Chenqi Xu. Cholesterol metabolism in t cells. *Frontiers in immunology*, 8:1664, 2017. 65

[147] Feiyang Cai, Shuxin Jin, and Guangjie Chen. The effect of lipid metabolism on cd4+ t cells. *Mediators of Inflammation*, 2021, 2021. 65

[148] Laura C Echeverri Tirado and Lina M Yassin. B cells interactions in lipid immune responses: implications in atherosclerotic disease. *Lipids in health and disease*, 16(1):1–11, 2017. 65

[149] Anneleen Remmerie and Charlotte L Scott. Macrophages and lipid metabolism. *Cellular immunology*, 330:27–42, 2018. 66

[150] MacRae F Linton, Patricia G Yancey, Sean S Davies, W Gray Jerome, Edward F Linton, Wenliang L Song, Amanda C Doran, and Kasey C Vickers. The role of lipids and lipoproteins in atherosclerosis. *Endotext [Internet]*, 2019. 66

[151] S Sini, D Deepa, S Harikrishnan, and N Jayakumari. Adverse effects on macrophage lipid transport and survival by high density lipoprotein from patients with coronary heart disease. *Journal of biochemical and molecular toxicology*, 32(9):e22192, 2018. 66

[152] Dwight E Bergles and William D Richardson. Oligodendrocyte development and plasticity. *Cold Spring Harbor perspectives in biology*, 8(2):a020453, 2016. 66

[153] James L Salzer. Schwann cell myelination. *Cold Spring Harbor perspectives in biology*, 7(8):a020529, 2015. 66

[154] Klaus-Armin Nave and Hauke B Werner. Myelination of the nervous system: mechanisms and functions. *Annual review of cell and developmental biology*, 30:503–533, 2014. 66

[155] Laura Montani. Lipids in regulating oligodendrocyte structure and function. In *Seminars in cell & developmental biology*. Elsevier, 2020. 66

[156] Davide Marangon, Marta Boccazzi, Davide Lecca, and Marta Fumagalli. Regulation of oligodendrocyte functions: targeting lipid metabolism and extracellular matrix for myelin repair. *Journal of clinical medicine*, 9(2):470, 2020. 66

[157] Nutabi Camargo, Andrea Goudriaan, Anne-Lieke F van Deijk, Willem M Otte, Jos F Brouwers, Hans Lodder, David H Gutmann, Klaus-Armin Nave, Rick M Dijkhuizen, Huibert D Mansvelder, et al. Oligodendroglial myelination requires astrocyte-derived lipids. *PLoS biology*, 15(5):e1002605, 2017. 67

[158] Kristina Hofmann, Rosalia Rodriguez-Rodriguez, Anne Gaebler, Núria Casals, Anja Scheller, and Lars Kuerschner. Astrocytes and oligodendrocytes in grey and white matter regions of the brain metabolize fatty acids. *Scientific reports*, 7(1):1–12, 2017. 67

[159] Bailey A Loving and Kimberley D Bruce. Lipid and lipoprotein metabolism in microglia. *Frontiers in physiology*, 11:393, 2020. 67

[160] Theo Vos, Christine Allen, Megha Arora, Ryan M Barber, Zulfiqar A Bhutta, Alexandria Brown, Austin Carter, Daniel C Casey, Fiona J Charlson, Alan Z Chen, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. *The lancet*, 388(10053):1545–1602, 2016. 67

[161] Haidong Wang, Mohsen Naghavi, Christine Allen, Ryan M Barber, Zulfiqar A Bhutta, Austin Carter, Daniel C Casey, Fiona J Charlson, Alan Zian Chen, Matthew M Coates, et al. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015. *The lancet*, 388(10053):1459–1544, 2016. 67

[162] Julien Bensaid. Idiopathic hypertrophic subaortic stenosis and associated coronary artery disease: clinical, hemodynamic, and therapeutic features from a review of 57 cases of the literature. *Angiology*, 30(9):585–593, 1979. 68

[163] Abe Walston II and Victor S Behar. Spectrum of coronary artery disease in idiopathic hypertrophic subaortic stenosis. *The American journal of cardiology*, 38(1):12–16, 1976. 68

[164] Mihoko Kono, Akira Kisanuki, Kunitsugu Takasaki, Kenichi Nakashiki, Toshinori Yuasa, Eiji Kuwahara, Naoko Mizukami, Takeshi Uemura, Kayoko Kubota, Nami Ueya, et al. Left ventricular systolic function is abnormal in diastolic heart failure: re-assessment of systolic function using cardiac time interval analysis. *Journal of cardiology*, 53(3):437–446, 2009. 68

[165] Kunitsugu Takasaki, Masaaki Miyata, Masakazu Imamura, Toshinori Yuasa, Eiji Kuwahara, Kayoko Kubota, Mihoko Kono, Nami Ueya, Yoshihisa Horizoe, Hideto Chaen, et al. Left ventricular dysfunction assessed by cardiac time interval analysis among different geometric patterns in untreated hypertension. *Circulation Journal*, 76(6):1409–1414, 2012. 68

[166] Mohammad Shenasa, Hossein Shenasa, and Nabil El-Sherif. Left ventricular hypertrophy and arrhythmogenesis. *Cardiac electrophysiology clinics*, 7(2):207–220, 2015. 68

[167] Dennis V Cokkinos, Zvonimir Krajcer, and Robert D Leachman. Coronary artery disease in hypertrophic cardiomyopathy. *The American journal of cardiology*, 55(11):1437–1438, 1985. 68

[168] Hector Lardani, Jose A Serrano, and Ramon J Villamil. Hemodynamics and coronary angiography in idiopathic hypertrophic subaortic stenosis. *The American journal of cardiology*, 41(3):476–481, 1978. 68

[169] Ettore Lazzeroni, Angelo Rolli, Enrico Aurier, and Giuseppe Botti. Clinical significance of coronary artery disease in hypertrophic cardiomyopathy. *The American journal of cardiology*, 70(4):499–501, 1992. 68

[170] M Chiong, ZV Wang, Z Pedrozo, DJ Cao, R Troncoso, M Ibacache, A Criollo, A Nemchenko, JA Hill, and S Lavandero. Cardiomyocyte death: mechanisms and translational implications. *Cell death & disease*, 2(12):e244–e244, 2011. 68

[171] Maja-Theresa Dieterlen, Katja John, Hermann Reichenspurner, Friedrich W Mohr, and Markus J Barten. Dendritic cells and their role in cardiovascular diseases: a view on human studies. *Journal of immunology research*, 2016, 2016. 68

[172] Sophie Quick, Jonathan Moss, Rikesh M Rajani, and Anna Williams. A vessel for change: endothelial dysfunction in cerebral small vessel disease. *Trends in Neurosciences*, 2020. 68

[173] Divine C Nwafor, Allison L Brichacek, Ahsan Ali, and Candice M Brown. Tissue-nonspecific alkaline phosphatase in central nervous system health and disease: A focus on brain microvascular endothelial cells. *International Journal of Molecular Sciences*, 22(10):5257, 2021. 68

[174] Sheng Jun An, Pei Liu, Tie Mei Shao, Zhi Jun Wang, Hai Gang Lu, Zhan Jiao, Xue Li, and Jun Qiu Fu. Characterization and functions of vascular adventitial fibroblast subpopulations. *Cellular Physiology and Biochemistry*, 35(3):1137–1150, 2015. 70

[175] Di Wang, Zhiyan Wang, Lili Zhang, and Yi Wang. Roles of cells from the arterial vessel wall in atherosclerosis. *Mediators of inflammation*, 2017, 2017. 70

[176] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Blood vessels and endothelial cells. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002. 70

[177] Nathaniel G dela Paz and Patricia A D'Amore. Arterial versus venous endothelial cells. *Cell and tissue research*, 335(1):5–16, 2009. 70

[178] Ira J Goldberg and Karin E Bornfeldt. Lipids and the endothelium: bidirectional interactions. *Current atherosclerosis reports*, 15(11):365, 2013. 70

[179] Anthony M Dart and Jaye PF Chin-Dusting. Lipids and the endothelium. *Cardiovascular research*, 43(2):308–322, 1999. 70

[180] Tessa J Barrett. Macrophages in atherosclerosis regression. *Arteriosclerosis, thrombosis, and vascular biology*, 40(1):20–33, 2020. 71

[181] Kathryn J Moore, Frederick J Sheedy, and Edward A Fisher. Macrophages in atherosclerosis: a dynamic balance. *Nature Reviews Immunology*, 13(10):709–721, 2013. 71

[182] Marie-Louise M Grønholdt, Børge G Nordestgaard, Jacob Bentzon, Britt M Wiebe, Ji Zhou, Erling Falk, and Henrik Sillesen. Macrophages are associated with lipid-rich carotid artery plaques, echolucency on b-mode imaging, and elevated plasma lipid levels. *Journal of vascular surgery*, 35(1):137–145, 2002. 71

[183] Bernd Hewing, Kathryn J Moore, and Edward A Fisher. Hdl and cardiovascular risk: time to call the plumber? *Circulation research*, 111(9):1117–1120, 2012. 73

[184] J. M. Ramirez Decrescenzo, A. R Munoz, A.S Liss, F. M. Kuehn, F. Adiliaghdam, S. R. Hamarneh, and R. A. Hodin. 80.08 a novel alkaline phosphatase in pancreatic -cells. 73

[185] M Mahmood Hussain. Intestinal lipid absorption and lipoprotein formation. *Current opinion in lipidology*, 25(3):200, 2014. 73

[186] Paolo Zanoni, Sumeet A Khetarpal, Daniel B Larach, William F Hancock-Cerutti, John S Millar, Marina Cuchel, Stephanie DerOhannessian, Anatol Kontush, Praveen Surendran, Danish Saleheen, et al. Rare variant in scavenger receptor bi raises hdl cholesterol and increases risk of coronary heart disease. *Science*, 351(6278):1166–1171, 2016. 73

[187] Ji Hye Park, Seyeon Mun, Dong Phil Choi, Joo Young Lee, and Hyeon Chang Kim. Association between high-density lipoprotein cholesterol level and pulmonary function in healthy korean adolescents: the js high school study. *BMC pulmonary medicine*, 17(1):1–7, 2017. 73

[188] Kyriaki Michailidou, Sara Lindström, Joe Dennis, Jonathan Beesley, Shirley Hui, Siddhartha Kar, Audrey Lemaçon, Penny Soucy, Dylan Glubb, Asha Rostamianfar, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94, 2017. 76

[189] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015. 76