DISCOVERY AND INTERPRETATION OF SUBSPACE STRUCTURES IN OMICS DATA BY LOW-RANK REPRESENTATION

Xiaoyu Lu

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing
Indiana University
October 2022

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Sha Cao, Ph.D., Chair

Chi Zhang, Ph.D.

July 6, 2022

Jingwen Yan, Ph.D.

Yong Zang, Ph.D.

ACKNOWLEDGEMENT

I would like to start with expressing my gratitude to my Ph.D. advisor Prof. Sha Cao and Prof. Chi Zhang. Their care and mentorship are my strongest support during these five years, where I am totally reshaped from a boy to a man with critical thinking and rigorous attitude. I still remember the day Prof. Cao ‘interviewed’ me. She didn’t realize my intention until I gave her my resume. More importantly, Prof. Sha Cao and Prof. Chi Zhang made me feel warm in this strange country as an international student. They are not only wise advisors but also the best friends in my life. Not limited to research, they support me to doing everything which may benefit my career and life. It gives me quite a lot of flexibility and chances to experience everything and figure out what kind of person I want to be in the future.

I would like to thank Prof. Jingwen Yan and Prof. Yong Zang in my research committee for their help. My research cannot be solid without their suggestions and feedback.

I also would like to shout out to other lab members and friends at IU. You guys bring me so much happiness and light up my life. It is hard to imagine how my life would be without you.

Furthermore, I am very thankful for the help from all my colleagues at Bristol Myers Squibb and Merck, especially, Dr. Wei Zhang and Dr. Erika Bongen, my managers at BMS, Dr. Dan Chang, my co-op manager at Merck. You helped me to apply my knowledge in drug development and helping patients which makes me clearer about my future career.

Lastly, I would like to thank my family. From my childhood to now, my parents always support all my decisions and give me as many resources as they can to help me to pursue my dream. My beloved wife, thank you for accompanying me for the last three years as well as the rest of my life. Your support, understanding and tolerance, give me endless power to fight for our future.

Xiaoyu Lu

DISCOVERY AND INTERPRETATION OF SUBSPACE STRUCTURES IN OMICS DATA BY LOW-RANK REPRESENTATION

Biological functions in cells are highly complicated and heterogenous, and can be reflected by omics data, such as gene expression levels. Detecting subspace structures in omics data and understanding the diversity of the biological processes is essential to the full comprehension of biological mechanisms and complicated biological systems. In this thesis, we are developing novel statistical learning approaches to reveal the subspace structures in omics data. Specifically, we focus on three types of subspace structures: low-rank subspace, sparse subspace and covariates explainable subspace. For low-rank subspace, we developed a semi-supervised model SSMD to detect cell type specific low-rank structures and predict their relative proportions across different tissue samples. SSMD is the first computational tool that utilizes semi-supervised identification of cell types and their marker genes specific to each mouse tissue transcriptomics data, for better understanding of the disease microenvironment and downstream disease mechanism. For sparsity-driven sparse subspace, we proposed a novel positive and unlabeled learning model, namely PLUS, that could identify cancer metastasis related genes, predict cancer metastasis status and specifically address the under-diagnosis issue in studying metastasis potential. We found PLUS predicted metastasis potential at diagnosis have significantly strong association with patient's progression-free survival in their follow-up data. Lastly, to discover the covariates explainable subspace, we proposed an analytical pipeline based on covariance regression, namely, scCovReg. We utilized scCovReg to detect the pathway level second-order variations using scRNA-Seq data in a statistically powerful manner, and to associate the second-order variations with important subject-level characteristics, such as disease status. In conclusion, we presented a set of state-of-the-art computational solutions for identifying sparse subspaces in omics data, which promise to provide insights into the mechanism in complex diseases.

Sha Cao, Ph.D., Chair

Chi Zhang, Ph.D.

Jingwen Yan, Ph.D.

Yong Zang, Ph.D.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background

With the advent of high-throughput biotechnology, we could now investigate the genetic material of a biological object with multiple bioassays, including its genomic, transcriptomic, epigenomic, proteomic and metabolic profiles, which are altogether called Omics data [1]. On one hand, the analysis of omics data is often challenged by the high dimensionality of its feature space; on the other hand, the high-dimensional omics data provide us comprehensive assessment of genetic molecules, and great opportunities to derive underlying biological mechanisms for different biomedical problems.

In omics data, the tens of thousands of genetic features are often highly inter-correlated, as the underlying molecules are coordinated into different functional units to perform cellular functions and maintain its viability. In other words, an omics dataset can often be represented by the combination of some subspace structures whose rank and complexity is much less than the original matrix. Those low-rank structures are more interpretable than the original high-dimensional matrix, due to lower dimensionality and/or sparse structures. In machine learning, low-rank representation is usually used for data denoising, missing data imputation, matrix decomposition, submatrix detection, bi-clustering and feature extraction [2]. Meanwhile, the high-dimensional biological data often contains subspace structures that enables intelligent representation and processing. It will be more explainable and useful when a dataset is represented by a set of low-rank subspaces. Therefore, it is reasonable to consider transcriptomic data as a combination of several subspaces, namely, the data is considered as samples approximately drawn from a mixture of several low-rank subspaces. Those underlying low-rank structures in biomedical data may represent many biomedical phenomena and disease mechanisms. In this thesis, I built up a series of methods for low-rank structure detection, explanation and knowledge extraction using transcriptomic data.

1.2 Matrix decomposition and transcriptomic data deconvolution

Matrix decomposition is a technique that breaks down a matrix into the product of two different low-rank matrixes. It is widely used in rank estimation, solving linear systems as well as solving other scientific and engineering applications [3]. The low-rank matrices carry the structure information of the original dataset, and is often more robust in information transfer from one system to another. Hence, finding a subspace structure that can reveal the inherent characteristics in high-dimensional data will enable us to transfer knowledge in a more robust fashion, and maximize the utility of public data.

Tissue transcriptomic data display convoluted signals from different cell types [4]. Deconvoluting cell components and identifying strain-/tissue-/experimental condition-specific cell types and gene expressions are crucial for understanding how experimentally perturbed conditions are associated with cellular level characteristics and cell-cell interactions [5]. Currently, ImmuCC and its varied versions are the only methods specifically focusing on mouse data deconvolution [6]. The core computational algorithm, which was adapted from CIBERSORT designed for human [7], assumes fixed cell types and signature gene expressions (subject to simple transformations) regardless of the experimental conditions of the target data. Meanwhile, multiple deconvolution methods have been developed for investigating the heterogeneous cell types in human cancer or other tissue data [7-16]. TIMER [17] only makes estimations on six immune cell types. xCell [18] make estimations on the higher number of different immune cell types but may fail to detect signals from homogeneous samples. EPIC [9] can directly generate scores interpreted as cell fractions. But they may not be directly applicable to mouse tissue data. First of all, the cell type specific genes for human cells differ from mouse cells; secondly, compared with human, the variations among different mouse tissue samples may be considerably higher, as they are collected from different strains with varied genetic backgrounds and experimental conditions.

In chapter 2, I developed a novel semi-supervised deconvolution method, namely Semi-Supervised Mouse data Deconvolution (SSMD) [19], to infer data/tissue specific cell type marker genes and expression profiles as well as estimate their relative abundances from mouse transcriptomics data.

1.3 Sparse subspace and cancer metastasis prediction

Sparse signal representation has proven to be an extremely powerful tool for representing high-dimensional signals. This success is mainly due to the fact that important patterns have natural sparse representations with respect to fixed bases [20]. Although biomedical datasets are very high-dimensional, most of them lie on low-dimensional sparse subspaces to exhibit similar biological phenomena. This gives us the capability to uncover meaningful information using sparse subspace representations.

Metastatic cancer accounts for over 90% of all cancer deaths and compared with well-confined primary tumors. Metastatic cancer remains incurable because of its systemic nature and the resistance of disseminated tumor cells to existing therapeutic agents [21, 22]. So evaluations of metastasis potential are vital for minimizing metastasis associated mortality and achieving optimal clinical decision-making [23, 24]. Previous work has provided strong evidence indicating that a number of genomic markers in primary tumors form a sparse subspace structure which is associated with the development of metastasis, and those distant metastasis events can be inferred from gene expression profiles within the primary tumor bulk [25, 26]. A recent study used machine-learning techniques to determine metastatic tumor organ of origin using the somatic mutation data [27]. Several studies have defined gene expression signatures that predict overall and metastasis-free survival as well as progression and metastatic growth in breast cancer patients [28-32]. Tang et al. identified 13 genes can be used to predict locoregionally advanced nasopharyngeal carcinoma metastasis in a large cohort study [33]. Deep learning models are applied to image data for metastasis prediction [34]. The under-diagnosis of metastasis events often happens, but most of the existing study doesn't consider this from cancer gene expression aspect. Computationally evaluating a cancer patient's metastasis potential is vital for clinical decision-making and understanding the biological mechanism of metastasis which is the first step toward targeted therapeutics.

In chapter 3, I designed PLUS [35] to detect biologically explainable sparse subspace structure in patients' transcriptomic data and build a positive and unlabeled learning classifier which enables early metastasis event prediction at the pan-cancer level, as well as infer biologically meaningful gene markers for metastasis potential. Upon successfully completing this aim, we delivered a computational tool for predicting cancer metastasis potential with transcriptomic data for clinical and research use.

1.4 Covariance-explainable subspace structure and gene-gene interaction

In human body, the biological functions of different cells are not determined by one single gene but by the modulations of a group of genes. Computational detection of gene expression variations may help us to better understand the gene-gene associations with similar biological functions. Currently, a common approach to investigating grouped gene signals involves pathway enrichment methods and co-expression module detection using either the differentially expressed genes or genes' importance ranking [36, 37]. But the current enrichment or co-expression-based pathway analysis methods suffer from the selection of a proper threshold, and the biggest unmet need in the current pathway-level analysis of scRNA-Seq data is the lack of a rigorous and powerful statistical framework to make inferences on important variables.

Covariance regression has been utilized in studying regression problems when the outcome variable is a covariance matrix [38-44]. In chapter 4, I introduce a statistically powerful framework [45] based on covariance regression, to discover the covariates explainable subspace in covariance matrix and model the pathway level second-order variations using scRNA-Seq data and associate the second-order variations with important subject-level characteristics, such as disease status.

1.5 Contribution of this thesis

- We seek to reveal and interpret biologically meaningful low-rank subspace structures in gene expression data in complicated biological systems.
- We propose the first computational tool which utilizes semi-supervised identification of non-fixed cell types and their marker genes specific to each mouse tissue transcriptomics data.
- We propose a positive and unlabeled learning framework to classify cancer metastasis and specifically address the under-diagnosis issue in studying metastasis potential.

- We firstly integrate the covariance regression technique to model the impact of important variables such as disease status, age, and sex, which is explainable of gene-gene correlation in individual pathways in a statistically powerful manner on scRNA-Seq data.

Chapter 2 Development software for mouse bulk tissue gene expression data deconvolution

2.1 Introduction

The mouse has long served as the premier model organism for studying human biology and disease, due to their striking genetic homologies and physiological similarity to humans, as well as the relatively low cost of maintenance. Currently, thousands of unique inbred strains and genetically engineered mutants have been made available for a wide array of specific disease types [46]. Research on mouse models have provided added impetus and indispensable tool for studying human disease, regarding its initiation, maintenance, progression and response to treatment, as well as evaluating drug safety and efficacy. Amongst all, the ability to examine physiological states and interactions between diseased cells and their microenvironment in vivo represents the most important tool for studying disease dynamics. To this end, numerous omics data have been collected from mouse that vary in terms of genetic perturbations, cell/tissue types, and treatment conditions [47-50]. A strong computational capability is needed to study the interactions of components within the mouse tissue microenvironment subject to different genetic and physiological perturbations, the knowledge gained from which could be projected to human disease scenarios and provide invaluable insight and guidance for effective human therapeutic regimes.

Tissue transcriptomic data display convoluted signals from different cell types [4]. Deconvoluting cell components and identifying mouse strain-/tissue-/experimental condition-specific cell types and gene expressions are crucial for understanding how experimentally perturbed conditions are associated with cellular level characteristics and cell-cell interactions [5]. While multiple deconvolution methods have been developed for investigating the heterogeneous cell types in human cancer or other tissues data [7-16], they may not be directly applicable to mouse tissue data. First of all, the cell type specific genes for human cells differ from mouse cells; secondly, compared with human, the variations among different mouse tissue samples may be considerably higher, as they are collected from different strains with varied genetic background and experimental conditions.

Currently, ImmuCC and its varied versions are the only method specifically focusing on mouse data deconvolution [6]. The core computational algorithm, which was adapted from CIBERSORT designed for human [7], assumes fixed cell type and signatures gene expressions (subject to simple transformations) regardless of experimental conditions of the target data. This assumption becomes problematic as mouse data, which are collected from different strains, have varied genetic background, thus, it is expected the tissue compositions are highly adaptable regarding the existent cell types and their expression profiles [51-53]. Aside from prominent variability in the appearance of cell types and the expression levels of markers genes, mouse data deconvolution also suffers from the following challenges: diverse experimental platforms, prevalently small sample size of mouse experiments, and limited training data sets available for deriving signature genes of cell types.

To address these challenges, we developed a novel semi-supervised deconvolution method, namely Semi-Supervised Mouse data Deconvolution (SSMD), to infer data/tissue specific cell type marker genes and their expression profiles and estimate their relative abundances from transcriptomics data. SSMD is capable to infer the relative proportion of

35 cell types in the blood, inflammatory, cancer, central nervous system and hematopoietic system. To the best of our knowledge, SSMD is the only mouse data deconvolution method considering strain, tissue type and data specificity of cell type specific gene markers. We demonstrated SSMD achieved a high sensitivity in identifying the appearance of immune and stromal cell types in inflammatory tissue and brain cell types in central nervous tissue, and with a high accuracy in estimating their relative proportion on single cell RNA-seq simulated bulk tissue data sets. We also experimentally validated that the cell populations inferred by SSMD accurately recapitulates the true cell proportions measured by fluorescence-activated cell sorting (FACS) on a leukemia bone marrow data. Applications of SSMD on a large collection of public mouse blood, brain, cancer, and other inflammatory tissue data suggested that the method achieved a robust performance throughout diverse types of experimental conditions and platforms including RNA-seq, microarray and immuno-assay. In addition, the software of SSMD grants users to build in their own tissue/data specific knowledge of cell type specific markers to reinforce the method. An R package of SSMD is released through GitHub: https://github.com/xiaoyulu95/SSMD and a R Shiny based web server of SSMD is available at https://ssmd.ccbb.iupui.edu/.

2.2 Materials and Methods

2.2.1 Random walk based identification of cell type specifically expressed genes from tissue data

We applied a non-parametric random walk based approach to screen genes with higher expression in certain cell types comparing to others, using bulk cell training data. On the combined expression matrix containing M genes for N samples of K cell types, we first calculated the expected frequency of each cell type, i.e. dividing the total number of samples for the cell type ($N_k, k = 1, \dots, K$) by the total number of samples N, denoted as $E_k = N_k/N$, $k = 1, \dots, K$. For a given gene g , denote $\boldsymbol{x}$ and $\boldsymbol{x}^k$ as vectors of expression profile for cells of all types and type k. Denote O_{jk} as the percentage of values in $\boldsymbol{x}^k$ that are no less than the jth largest value in vector $\boldsymbol{x}$. A random walk vector $\boldsymbol{d}_{1\times N}$ that describes the non-negative discrepancy between the observed and expected cell type frequency of the gene was defined as $d_j = \sum_{k=1}^{K}\left(O_{jk} - E_k\right)^2, j = 1, \dots, N$, which attains a minimum value of zero at N. A higher peak of the random walk $\boldsymbol{d}_{1\times N}$ suggests gene g is more enriched in certain cell types than the others. Denote m as the index of the maximum of $\boldsymbol{d}_{1\times N}$, i.e. $\mathrm{m} = \mathrm{argmax}\ (\boldsymbol{d}_{1\times N})$, and the cell type frequency at m as $e_k^m = O_{mk} - E_k$. Cell types were further ordered by e_k^m decreasingly, and a labeling matrix L was built such that $L_{g,\mathrm{k}} = 0, if\ e_k^m \le 0$; otherwise, $L_{g,\mathrm{k}} = \frac{1}{p}, if\ \boldsymbol{x}^k$ has the pth largest mean among $\boldsymbol{x}^1, \dots, \boldsymbol{x}^K$.

It is noteworthy the approach can be directly applied to scRNA-seq data for marker training. In this study, due to the relatively limited availability of existing scRNA-seq data, especially the mouse strain and tissue type coverage, we generate core marker list purely by using bulk cell data.

2.2.2 Identification of rank-1 cell type uniquely expressed gene modules

To screen genes that form tight rank-1 modules on various tissue training datasets, SSMD performs a community detection method among the genes specifically expressed in each cell type as stored the labeling matrix. A correlation matrix was first built among cell type specifically expressed genes, and the significance cutoff of correlation was determined by random matrix theory. Random matrix theory (RMT) has been widely used to understand the low-rank structure encoded in biological data. In this study, an RMT based approached developed by Luo et al was used to determine the threshold of significant correlation for each dataset [54]. rm.get.threshold functions in the RMThreshold R package was utilized. Specifically, RMT indicated that the nearest neighbor spacing distribution of eigenvalues will have a characteristic change when the threshold properly separates signal from noise. By removing all the below-threshold correlation elements, the co-expression modules can be more robustly unraveled. Then, hierarchical clustering was performed using the correlation matrix as similarity measure.

Specifically, SSMD gradually increases the height of the hierarchical clustering at which the tree is cut. At each height, the number of genes, the average correlation among the genes, and the rank of the matrix composed of the genes in each of the cluster, is calculated. Here, matrix rank is determined by a modified bi-cross validation (BCV) algorithm. SSMD stops scanning the hierarchical tree if all the clusters contain less than q_0 genes, or the three following criterior is met for all the clusters: (1) with at least q_0 genes, (2) the average correlation among the genes is above the threshold determined by RMT, and (3) the rank of the expression matrix profile of the genes in the cluster is 1. In this study, q_0=7 is used. Such an iterative approach will eventually select the clusters with at least q_0 genes, each of which is considered as possible cell specific marker genes specific to this data set. SSMD merges modules until the canonical correlation between any pair of modules is lower than a cutoff cor_{cut} or the number of current modules is not larger than the total rank of the gene expression profile of the selected data set specific markers genes. In this study, we utilized $cor_{cut} = 0.9$.

2.2.3 A modified Bi-cross validation rank test

Bi-cross validation (BCV) has been developed to estimate the matrix rank for singular value decomposition (SVD) and Non-negative Matrix Factorization (NMF) , which requires a prefixed low dimension K and two low-rank matrices for the approximation $X_{M\times N} = W_{M\times K} \cdot H_{K\times N}$. The error distribution of gene expression data is usually non-identical/independent, mostly because a gene's expression can be affected by its major transcriptional regulators, other biological pathways and experimental bias. Hence undesired biological characteristics and experimental bias may form significant dimensions in a gene expression data [55]. In sight of this, we developed a modified BCV rank test (Algorithm 1) to minimize the effect of the non-i.i.d errors in assessing the matrix rank of a gene expression data.

Algorithm 1: Modified Bi-cross validation matrix rank test

```
Input: Matrix $X_{M \times N}$, parameters $M_0, N_0, R, msp$.
For r=1…R
    $Sample$ row index set $I_r = \left\{\mathrm{i}_1, \mathrm{i}_2, \dots, \mathrm{i}_{M_0} | i_p \in \{1 \dots M\}\right\}, \overline{I_r} = \{1 \dots M\} \backslash I_r$
    $Sample$ column index set $J_r = \left\{\mathrm{j}_1, \mathrm{j}_2, \dots, \mathrm{j}_{N_0} | j_p \in \{1 \dots N\}\right\}, \overline{J_r} = \{1 \dots N\} \backslash J_r$
    $Split\ X$ into four submatrices $\begin{vmatrix} A_r & B_r \\ C_r & D_r \end{vmatrix}$, $where\ A_r = X[I_r, J_r], B_r = X[I_r, \overline{J_r}],$
    $C_r = X[\overline{I_r}, J_r], D_r = X[\overline{I_r}, \overline{J_r}]$
    $For\ k = 1 \dots \min(M_0, N_0)$
        $\mathrm{BCV}(k, r) = \sum_{i=1}^{M_0} \sum_{j=1}^{N_0} \left\| A_r - B_r \widehat{D_r}^{(k)^+} C_r \right\|_F^2$ $(*)$
    End
End
$\mathrm{Rank_x} \leftarrow 0$
$For$ $\mathrm{k} = 1 \dots \min(M_0, N_0)$
    $Do$ t test between $\{\mathrm{BCV}(k, r) | r = 1 \dots R\}$ and $\{BCV(k+1, r) | r = 1 \dots R\}$
    $if$ $(\mathrm{p.value} < 0.01\ \&\ \mathrm{mean}\left(\mathrm{BCV}(k+1, r)\right) - \mathrm{mean}\left(\mathrm{BCV}(k, r)\right) > \mathrm{msp})$
        $\mathrm{Rank_x} \leftarrow k$
End
$Return$ $\mathrm{Rank_x}$
```

$(*)$ Denote the SVD of a matrix D as $D = U\Sigma V'$, and Moore– Penrose inverse of D as $D^+, D^+ = V'\Sigma^+ U$, where Σ^+ is a diganol matrix $\mathrm{diag}\left(\sigma_1^+, \sigma_2^+, \dots \sigma_p^+\right)$ with $\sigma_1^+ \geq \sigma_2^+ \geq \cdots \geq \sigma_p^+ \geq 0$. Define $\widehat{D}^{(k)^+} = \sum_{i=1}^{k} \sigma_i^{\ +} v_i\, u_i$

After running the rank-1 module detection on all the training bulk tissue datasets, those genes commonly identified in the rank-1 modules in more than 40% (70%) data sets were selected as core (stringent) markers. The list of stringent marker sets was derived with more stringent criterion, which is particularly useful for the analysis of small sample sized target data. Core markers of cells in central nervous systems were identified by a similar approach on the brain training tissue datasets. Due to the limitation of hematopoietic system tissue training data, its core markers were selected as the genes specifically over expressed in each hematopoietic cell type, by using the criteria: the gene's expression level is above 10% quantile in one cell type and below 50% in the other cell types. Complete lists of selected core and stringent marker sets were given.

2.2.4 Estimation of cell proportion

Two methods were utilized to estimate cell proportion: (1) SVD based computation. With cell type specific markers derived, the first row base of the gene expression profile of the marker genes is directly utilized as an estimation of the cell proportion, which can be directly computed by SVD. (2) Constraint NMF based computation. With the number of identifiable cell types and cell type specific markers identified, the signature matrix $\tilde{S}_{M_0\times K_0}$ and proportion matrix $\tilde{P}_{K_0\times N}$ can be estimated by minimizing the following objective function:

$$\min_{\tilde{S}_{M_0\times K_0},\tilde{P}_{K_0\times N}} \left(\left\|\tilde{X}_{M_0\times N} - \tilde{S}_{M_0\times K_0}\cdot\tilde{P}_{K_0\times N}\right\|_F^2 + \lambda\cdot \mathrm{trace}\left(\tilde{S}_{M_0\times K_0}{}^{\mathrm{T}}\cdot\left(\mathbf{1}_{\mathrm{M}_0}\mathbf{1}_{\mathrm{K}_0}^{\mathrm{T}} - \mathrm{C}_{M_0\times K_0}\right)\right)\right)$$

, where $\mathrm{C}_{M_0\times K_0}[i,j] = 1$ if gene i is marker of the cell type j, and 0 otherwise. λ is the hyper parameter. In this study, we tuned λ by using single cell data simulated tissue data. λ=10 is empirically utilized in the analysis.

2.2.5 Explanation score and Comparison with state-of-the-arts methods

An explanation score (ES) was utilized to evaluate the goodness that each marker gene's expression is fitted by the predicted cell proportions:

$$EScore(x) = 1 - \sum_{j=1}^{N}(x_j^* - \hat{x}_j)^2 \, / \sum_{j=1}^{N}(x_j^*)^2, \ \widehat{x_j} = \sum_{k=1}^{k_x} \beta_k^x p_j^k \, , \beta_k^x \geq 0$$

where x_j^* is the observed expression of marker gene x in sample j, $\hat{x}_j$ is the explainable expression by cell proportions, obtained by a non-negative regression x on the predicted proportion $p_j^k, k = 1 \ldots k_x$. Here, k_x represents the number of cell types that express x, and β_k^x are the non-negative regression parameters. Intuitively, with correctly selected marker genes, the marker gene's expression can be well explained by the predicted proportions of the cell types that express the gene. Hence, a high ES score is a necessary but not sufficient condition for correctly selected marker genes and predicted cell proportion.

2.2.6 Data used in this study

Bulk cell training data sets: for mouse blood, solid cancer and inflammatory tissue microenvironment, we retrieved 116 datasets of sorted mouse cells of 12 selected cell types, totaling 1106 samples from GEO database. For mouse brain tissue microenvironment, we collected 2130 bulk cell samples of the nine selected cell types in central nerve systems. For mouse hematopoietic microenvironment, two datasets were available that cover 14 hematopoietic cell types. All the bulk cell training data were generated by the Affymetrix GeneChip Mouse Genome 430 2.0 Array platform and normalized with MAS5 method [56]. Samples of the same cell type were further merged together with batch effect removed using Combat [57].

Single Cell RNA-sequencing data: One mouse melanoma scRNAseq data set (6638, 9) was acquired from the Human Cell Atlas database [58]. Three scRNA-seq datasets of lung (4485, 12), pancreas (4405, 8), and small intestine (4764, 10) and two sets of brain tissue (3679, 7 and 1099, 6) were accessed from Mouse Cell Atlas (MCA) data portal [59]. The two numbers in the parenthesis indicate the number of cell samples and cell types of each data set. We specifically selected the cells with UMI more than 500 to exclude low quality cells. Cell labels were either provided in the original data or curated using Seurat v3 with cell type specific genes [60, 61].

Training tissue data from cancer and blood: 33 cancer tissue datasets of 9 cancer types generated by four popular experimental platforms were collected, namely Illumina HiSeq 2000 Mus musculus, Affymetrix Mouse Genome 430 2.0 Array, Illumina HiSeq 2500 Mus musculus and Affymetrix Mouse Genome 430A 2.0 Array from GEO database. Each data set has at least 15 samples. We didn’t consider datasets from immunodeficient mouse, mouse cell lines, and PDX models, as only real cancer or blood micro-environment is considered. A data set of liver tissue collected from 31 mouse strains (GSE55489) were utilized to evaluate the variation of cell type specific markers through different mouse strains [62].

Brain tissue data: 14 datasets of mouse brain tissues generated by two experimental platforms, namely Illumina HiSeq 2500 Mus musculus and Affymetrix Mouse Genome 430 2.0 Array were collected from Gene Expression Omnibus. Datasets were split into sub data sets of different brain regions. Each data set has at least 40 samples. The complete training data information are available.

Hematopoietic System tissue and FACS data: We generated a RNA-seq data set with matched FACS data of bone marrow cells isolated from the hind limbs of C57BL/6, Tet2-/-Flt3ITD , DNMT3A-/-Flt3ITD , and DNMT3A-/-Tet2-/-Flt3ITD mice (n=3 for each group). RNA (600 ng/ sample) was used to prepare single indexed strand specific cDNA library using TruSeq stranded mRNA library prep kit (Illumina). The library prep was assessed for quantity and size distribution using Qubit and Agilent 2100 Bioanalyzer. The pooled libraries were sequenced with 75bp single-end configuration on NextSeq500 (Illumina) using NextSeq 500/550 high output kit. The quality of sequencing was confirmed using a Phred quality score. The sequencing data was next assessed using FastQC (Babraham Bioinfomatics, Cambridge, UK) and then mapped to the mouse genome (UCSC mm10) using STAR RNA-seq aligner [63], and uniquely mapped sequencing reads were assigned by featureCounts. The data were normalized to RPKM. FACS data were collected from same biological prep by IU School of Medicine Flowcytometry Core. Hematopoietic stem cells were identified by lineage negative, C-Kit high and Sca1 high cells, general myeloid progenitor cells were identified by Cd34 and Cd16/32 high cells, mature myeloid cells were identified by Gr1 and Cd11b high cells, and PreB cells were identified by B220 and SSC-A high cells.

2.2.7 Generation of simulated bulk tissue data from scRNA-seq data

Cell types in each scRNA-seq data were labeled by the cell clusters provided in the original works or by using Seurat pipeline with default parameters. Detailed information of the scRNA-seq data and cell type annotation is given. For each data set, we simulate bulk tissue data by: (1) removing insignificantly expressed genes, (2) randomly generate the proportion of each cell type, called true proportion in this paper, that follows a Dirichlet distribution, and (3) draw cells randomly from the cell pool with replacement according to the cell type proportion, and sum up the expression values of all cells to produce a pseudo bulk tissue data. The insignificant expressed genes were identified by left truncated mixture Gaussian model [64, 65]. The Dirichlet distribution matrix was generated with R package "DirichletReg" [66].

2.3 Results

2.3.1 Mathematical consideration and problem formulation

Denote $\tilde{X}_{M\times N}$ as a tissue data of M genes and N samples, a deconvolution analysis assumes $\tilde{X}_{M\times N}$ as the following non-negative product form:

$$\tilde{X}_{M_0\times N} = \tilde{S}_{M_0\times K_0} \cdot \tilde{P}_{K_0\times N} + E, \tilde{S}_{M_0\times K_0} \geq 0, \tilde{P}_{K_0\times N} \geq 0 \quad (1)$$

Here, $\tilde{X}_{M_0\times N}$ represents the observed gene expression matrix of M_0 selected genes (a subset in M) in N tissue samples, and columns in $\tilde{S}_{M_0\times K_0}$, and rows in $\tilde{P}_{K_0\times N}$, denote the expression signatures, and the relative proportions of the K_0 cell types respectively. In the conventional formulation of deconvolution analysis, with fixed M_0 and K_0, $\tilde{S}_{M_0\times K_0}$ and $\tilde{P}_{K_0\times N}$ are solved to minimize the $\mathcal{L}_2$ loss of the above linear equation. Because of the highly varied genetic and phenotypic background of mouse experiment, $\tilde{S}_{M_0\times K_0}$, M_0 and K_0 are usually varied and unknown, i.e. for each $\tilde{X}_{M\times N}$ collected from tissues of certain microenvironment, what cell types are present, what gene markers each cell type expresses and how much they were expressed, could vary drastically due to the genetic and physiological perturbations. Correctly specified cell types K_0, and selected cell type marker genes M_0 can largely increase the prediction accuracy of $\tilde{P}_{K_0\times N}$. Table 2.1 lists the key mathematical definitions utilized in this study.

In this study, we define a cell type k is "transcriptomically identifiable" if its ground-truth proportion $P_{1\times N}^{k}$ and estimated as $\tilde{P}_{1\times N}^{k}$ have high correlation, i.e.. $cor\left(P_{1\times N}^{k}, \tilde{P}_{1\times N}^{k}\right) = 1 - \epsilon$ and ϵ is substantially small, where $\tilde{P}_{1\times N}^{k}$ is the kth row of $\tilde{P}_{K_0\times N}$, and K_0 as the number of "identifiable" cell types. A strong condition for a cell type to be identifiable is that it has uniquely expressed genes [67]. Here we provided a comprehensive mathematical derivation of the relationship between cell type unique expression and identifiability of cell proportion. We derived the identity of cell type uniquely expressed gene markers, denoted as the set G_k, is a necessary but non-sufficient condition for the identifiability of cell type k: – if k is "transcriptomically identifiable", $\tilde{X}_{G_k\times T}$ must be a matrix of rank one, for $\forall\, T \subset \{1, \dots, N\}$. This condition forms the foundation of how SSMD discover cell type marker genes that are not fixed, but instead specific to each dataset. Fortunately, we do not need to scan for all the local rank-1 matrices within $\tilde{X}_{M\times N}$, where M is usually to the tens of thousands. In fact, with an effective knowledge transfer of the gene labels derived from single or bulk cell training data, the genes that are more likely to be cell type specific markers of identifiable cell types can be detected, which forms the core algorithm of SSMD pipeline.

Terminology	Mathematical Definition in this study
Rank-1 matrix	A matrix with rank = 1, i.e. the matrix is generated by the product of two vectors, $\mathrm{X}=\mathrm{A}\cdot\mathrm{B}^{\mathrm{T}}$. In this study, we consider all transcriptomics data are with error. Hence the rank-1 matrix is defined by $\mathrm{X}=\mathrm{A}\cdot\mathrm{B}^{\mathrm{T}}+\mathrm{E}$, where the matrix rank of X is 1 can be computed by the bi-cross validation (BCV) algorithm detailed in Methods.
Local rank-1 matrix	A submatrix with rank = 1, i.e. denoting I and J as the indices of the submatrix, $\mathrm{X}_{\mathrm{I}\times\mathrm{J}}$ is generated by the product of two vectors with error, $\mathrm{X}_{\mathrm{I}\times\mathrm{J}}=\mathrm{A}\cdot\mathrm{B}^{\mathrm{T}}+\mathrm{E}$.
Transcriptomically identifiable cell type	The cell type with a high correlation between the true proportion $\mathrm{P}_{1\times\mathrm{N}}^{\mathrm{k}}$ and estimated $\widetilde{\mathrm{P}}_{1\times\mathrm{N}}^{\mathrm{k}}$
Prediction accuracy	Pearson correlation between true proportion and predicted proportion of each cell type
Detection accuracy	The number of true cell type signature genes were identified as signature genes of an identifiable cell type
Matrix total Rank	The total rank of a data matrix that can be tested by the BCV algorithm

Table 2.1. Definition of mathematical terms in SSMD

2.3.2 SSMD Analysis pipeline

SSMD is a semi-supervised method composed by (1) training a large candidate list of cell type specific marker genes, (2) evaluating the identifiability of each cell type and confirming their marker genes for each to-be-deconvolved data, and (3) estimating the proportion of each cell type.

The training step is to look for genes that are more likely to serve as cell type marker genes through different tissue types and data sets, named as core marker lists. Specifically, we identified the genes that are commonly over expressed in one cell type comparing to the others in bulk cell data and commonly form rank-1 matrices in tissue data, by using a very extensive set of training data sets collected from different mouse strains and tissue types (see details in Methods). Figure 2.1A illustrates the procedure of SSMD to construct cell type core marker lists. On the bulk cell training data, we adopted a random-walk based approach to detect genes that are significantly expressed in higher quantities in one or a few cell types, than others (see details in Methods). As a result, a labeling matrix that annotates cell type specifically expressed genes will be constructed, which forms the first evidence of the potential marker genes for each cell type. Then on each bulk training tissue dataset, we further identified marker genes that form rank-1 submatrices with a community detection approach as detailed in methods. Only those modules, whose genes significantly and consistently over-represent one and only one cell type across multiple training tissue datasets, are selected to form the core marker list. Noted, variations caused by different experiment batches, tissue types and mouse strains were handled by enabling certain errors in the random-walk based cell type specific marker identification, i.e. identifying the genes overly expressed in the cell type comparing to the others in a certain proportion of the collected bulk cell data. In addition, data batch variation was also considered in the bulk data based training step, by identifying the genes commonly serve as cell type specific marker in more than 50% of analyzed bulk tissue training data. The goal of this training procedure is to summarize a relatively large list of commonly observed cell type specific marker genes, which can be used to as semi-supervised information to identify data set specific cell type marker for a further un-supervised deconvolution analysis.

Based on the cell type core markers, the deconvolution of any given bulk tissue dataset is composed by the steps as illustrated in Figure 2.1B. SSMD first identifies all the rank-1 modules on the target dataset by an iterative hierarchical clustering and bi-cross validation approach. Then SSMD selects the rank-1 modules that are likely to be markers of a certain cell type for this data set, if genes in the modules largely overlap with the core marker list of one and only one cell type. Modules that are highly co-linear will be merged. Consequently, genes in each module is called gene markers of one cell type, that satisfy the necessary condition for "transcriptomically identifiable". Notably, two modules may represent the same cell type, and they are treated as marker genes of different subtypes of the cell type. Here, the total number of modules is an estimate of the number of "identifiable" cell types, i.e., K_0. Importantly, SSMD is an "semi-supervised" approach, because the cell marker genes do not solely depend on the training data, but also the co-expression patterns of the marker genes in the target dataset. In other words, SSMD addresses the variability issue of signature genes from one dataset to another, and has the potential to discover cell types not pre-defined. Algorithms of each computational step are detailed in Materials and Methods.

The prediction of the cell type proportions is conducted using a constrained Non-negative Matrix Factorization (NMF) method by solving the following optimization problem:

$$\min_{\tilde{S}_{M_0\times K_0},\tilde{P}_{K_0\times N}}\left(\left\|\tilde{X}_{M_0\times N}-\tilde{S}_{M_0\times K_0}\cdot\tilde{P}_{K_0\times N}\right\|_F^2+\lambda\cdot\mathrm{trace}\left({\tilde{S}_{M_0\times K_0}}^{\mathrm{T}}\cdot(\mathbf{1}_{\mathrm{M}_0}\mathbf{1}_{\mathrm{K}_0}^{\mathrm{T}}-\mathrm{C}_{M_0\times K_0})\right)\right)\quad(2)$$

where $\mathrm{C}_{M_0\times K_0}[i,j]=1$ if gene i is marker of the cell type j, and 0 otherwise. $\mathbf{1}_{\mathrm{d}}$ denotes an all-1 column vector of length d, λ is a hyperparameter selected by cross validation, and other annotations follow equation (1). The constraint matrix $\mathrm{C}_{M_0\times K_0}$ is enforced upon the regular NMF to guarantee similarity of the solved signature matrix $\tilde{S}_{M_0\times K_0}$ and constraint $\mathrm{C}_{M_0\times K_0}$, namely, in the kth column of $\tilde{S}_{M_0\times K_0}$, it should have higher expressions for genes that are markers of cell type k. The solution to (2) is by alternative update where each time one of $\tilde{S}_{M_0\times K_0}$, $\tilde{P}_{K_0\times N}$ is held fixed, and the other is updated. λ can be tuned by using simulated tissue data with known cell proportion. In this study, we tuned λ and empirically select λ as 10 when $\tilde{X}_{M_0\times N}$ is log normalized microarray data or log(X+1) normalized FPKM/CPM/TPM RNA-seq data.

Following these procedures, and on a large collection of mouse bulk cell and tissue training data, we generated core marker gene lists for different tissue microenvironments: (1) for mouse blood, solid cancer and inflammatory tissues, 980 genes of 12 cell types namely T cell, B cell, NK cell, hematopoietic stem cell, monocyte, macrophage, neutrophil, mast cell, adipocytes, fibroblast, dendritic cell, and endothelial cell were discovered (Figure 2.1C); (2) for mouse hematopoietic system, 2877 genes of 14 cell types namely hematopoietic stem cell, common lymphoid progenitor, granulocyte-macrophage progenitors, megakaryocyte lineage-committed progenitor, erythroid cell, megakaryocyte-erythrocyte progenitors, multipotent progenitors, early myeloid progenitor, mature myeloid cell, pre colony forming unit erythroid, pre-megakaryocytic/erythroid progenitor, B cell, CD4+ T and CD8+ T cell were discovered, and (3) for mouse central nervous system tissue, 1570 genes of nine cell types namely ependymal cell, general glial cell, oligodendrocyte, stromal-like cell, Schwann cell, microglial, neuron, and astrocyte were discovered (Figure 2.1D). It is noteworthy that the size of core marker list ranges from 27 to 547 for different cell types. However, our analysis suggested that more than 5 marker genes that form a rank-1 matrix is sufficient for an accurate estimation of cell proportion. Note that, compared with conventional regression-based deconvolution analysis, SSMD only uses labels of the core markers as the semi-supervised information and identifies data set specific cell type markers for a further unsupervised estimation of cell types, which grants a flexibility and robustness to handle the variation of cell type specific marker genes and their expression scale through different mouse strains, tissue types and experimental platforms. In addition, the semi-supervised formulation of SSMD enables the inference of identifiability of each cell type and identification of rare or sub cell types.

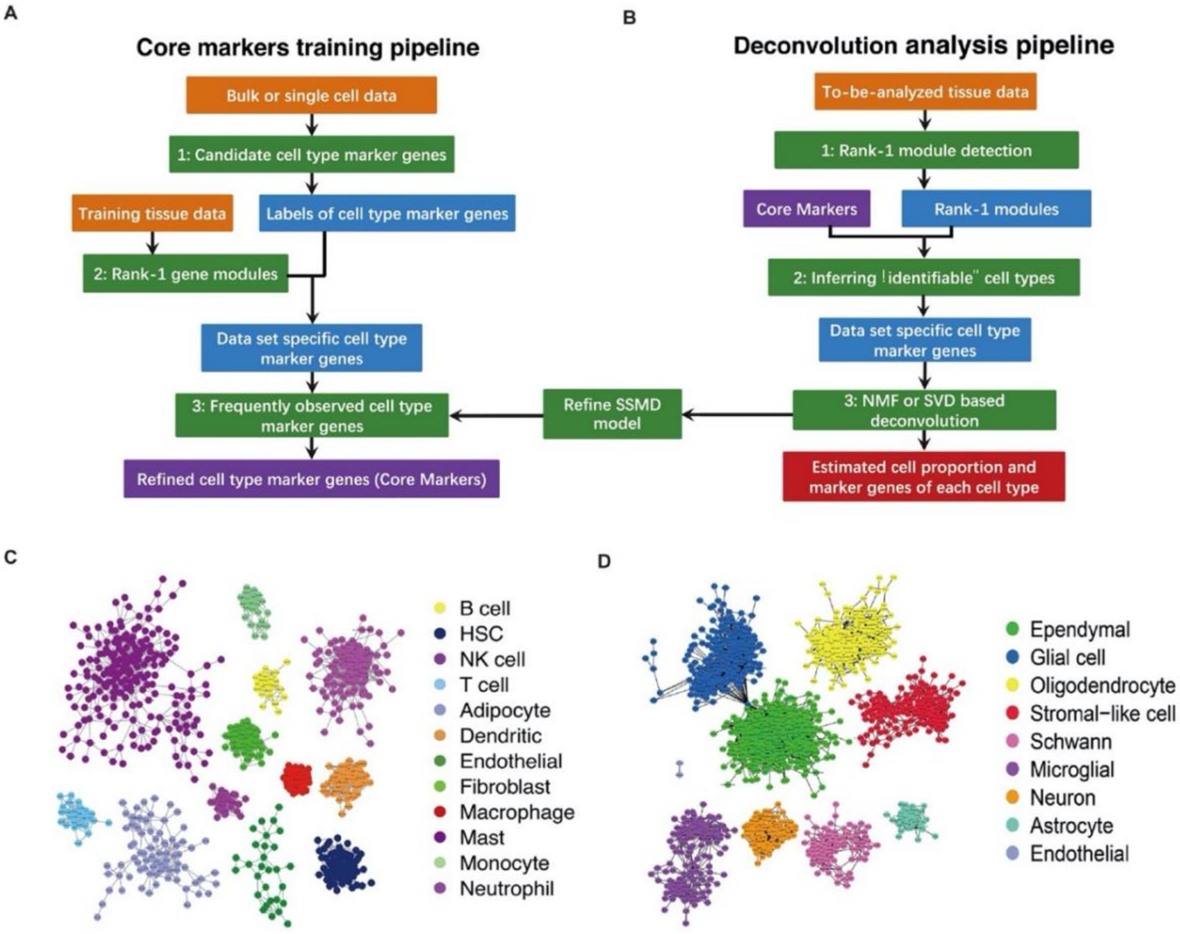

A
Core markers training pipeline
Bulk or single cell data
1: Candidate cell type marker genes
Training tissue data
Labels of cell type marker genes
2: Rank-1 gene modules
Data set specific cell type marker genes
3: Frequently observed cell type marker genes
Refined cell type marker genes (Core Markers)
B
Deconvolution analysis pipeline
To-be-analyzed tissue data
1: Rank-1 module detection
Core Markers
Rank-1 modules
2: Inferring ¦identifiable'' cell types
Data set specific cell type marker genes
3: NMF or SVD based deconvolution
Estimated cell proportion and marker genes of each cell type
Refine SSMD model
C
B cell
HSC
NK cell
T cell
Adipocyte
Dendritic
Endothelial
Fibroblast
Macrophage
Mast
Monocyte
Neutrophil
D
Ependymal
Glial cell
Oligodendrocyte
Stromal−like cell
Schwann
Microglial
Neuron
Astrocyte
Endothelial

Figure 2.1. Analysis pipeline of SSMD and core cell type-specific markers. (A) Analysis pipeline of the core marker training procedure. (B) Analysis pipeline of the deconvolution procedure. In (A) and (B), input data including training and target data, computational procedure and key intermediate outputs were colored by orange, green and blue, respectively. (C) Core markers of 12 cell types in blood, solid cancer and inflammatory tissue. An edge between two genes means the two genes are coidentified as markers of one cell type in more than 50% of the training data sets. (D) Core markers of nine cell types in central nervous system. Notably, core markers for the endothelial cell in the inflammatory tissue and central nervous system were separately trained by comparing with other cell types in the same tissue system.

2.3.3 Benchmarking based on artificial tissue data simulated by using single cell RNAseq data

We first benchmarked SSMD on a set of artificial tissue data simulated from four single cell RNAseq (scRNA-seq) datasets of mouse lung, pancreas, small intestine and melanoma. For each data set, we simulated 100 tissue samples by randomly drawing and mixing cells of different types whose proportions follow random Dirichlet distributions. Prediction accuracy of each cell type was assessed by the Pearson correlation coefficients between its known mixing cell proportions and the predicted relative proportion. We compared SSMD with three state-of-arts deconvolution methods of mouse data, namely ImmuCC (ICC), tissue-ImmuCC (TICC) and EPIC [9]. Our analysis suggested that SSMD achieved 93.2% prediction accuracy on average in the four simulated data sets and 23 out of the 28 cell types (82.1%) are with higher than 0.9 prediction accuracy (Figure 2.2A-D). In contrast, EPIC, ICC and TICC achieved 69.7%, 45.2% and 48.5% averaged prediction accuracy on the cell types covered by these methods, and the proportion of cell types with higher than 0.9 prediction accuracy are 32.2% (9/28), 0% (0/28) and 7.2% (1/14), respectively. We also tested the popular human data deconvolution methods such as CIBERSORT (CIBERSORTx) and TIMER [5, 7], by using the known human and mouse homolog genes. Non-surprisingly, predictions made by CIBERSORT and TIMER on the mouse are less accurate than SSMD. TIMER and CIBERSORT achieved 49.25% and 47.5% averaged prediction accuracy, and the proportion of cell types with higher than 0.9 prediction accuracy are 17.9% (5/28) and 3.6% (1/28).

It is noteworthy that the SSMD enables the detection of sub cell types defined as transcriptomically identifiable. SSMD successfully identified two sub populations of fibroblast cells in the melanoma data and different subtypes of neutrophils in lung and small intestine data. In contrast, ICC, TICC and EPIC are not capable of providing cell subtype predictions due to their fixed cell type assumption.

We also benchmarked SSMD on simulated brain tissue data using two scRNA-seq data of central nervous systems. SSMD achieved more than 0.9 correlation in predicting the cell types microglial, stromal-like, and ependymal subtypes in the simulated tissue data (Figure 2E-F). To the best of our knowledge, SSMD is the first of its kind method to specifically target mouse central nervous system decomposition. To benchmark SSMD, we selected MUSIC as the state-of-the-art method, which requires an additional input of an scRNA-seq data to train context specific gene signatures [68]. Here we first utilized the same scRNA-seq data for tissue data simulation and signature training in MUSIC. Non-surprisingly, MUSIC achieved consistently good predictions (averaged cor=0.99), and the predictions made by SSMD are very close to MUSIC with slightly lower correlations compared with MUSIC under this ideal setup. In sight the possible disparity caused by tissue, strain, and experimental platform variations between the target tissue data and available scRNA-seq data for training cell markers, we also conducted a robustness test of MUSIC and SSMD. Our analysis suggested that MUSIC highly depends on the consistency of cell type specific marker genes and their expression scale between the target tissue and the training scRNA-seq data. In contrast, the de novo data set specific marker identification by SSMD enables a broader application to the tissue data without matched scRNA-seq data. Because EPIC, ImmuCC and tissue-ImmuCC cannot analyze brain tissue data and the melanoma and pancreas tissue were not covered by tissue-ImmuCC, we did not include the comparison with these methods on the brain tissue data.

To further validate the specificity of SSMD, we tested the total rank of the identified marker genes and compared with the number identified cell types (TIMER and CIBERSORT achieved 49.25% and 47.5% averaged prediction accuracy. and the proportion of cell types with higher than 0.9 prediction accuracy are 17.9% (5/28), and 3.6% (1/28).). We also compare the total matrix rank of the marker genes used in other methods and the number of cell types assumed in those methods. Comparing to the fixed number of cell types in other methods, the number of cell types predicted by SSMD better matches the total rank of the expression profile of identified marker genes. Our observation suggested SSMD can correctly estimate the number of cell types and select proper markers for cell type proportion estimation. It is noteworthy the predicted number of cell types may not exactly match the total rank of selected markers because possible co-linearity among the true proportion of the cell types.

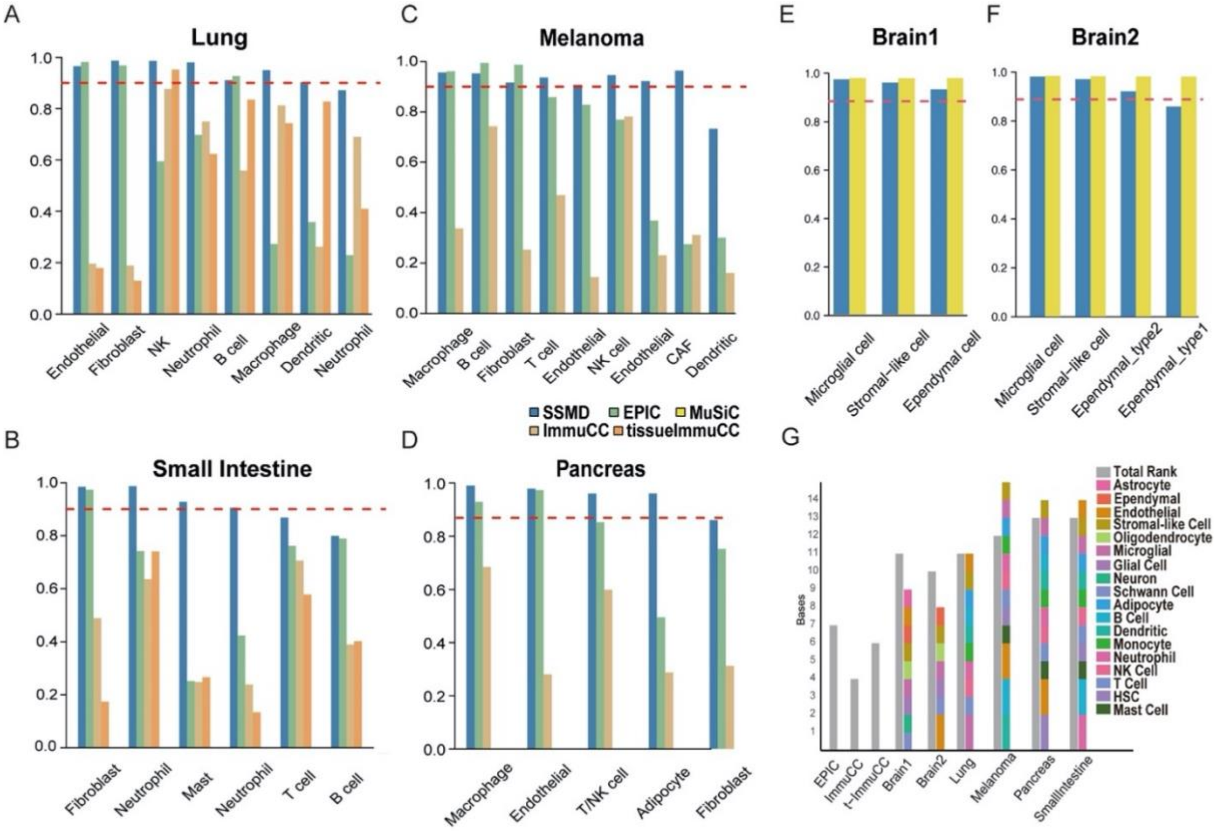


Figure 2.2. Method evaluation on scRNA-seq simulated tissue data. (A–D) Correlation between true and predicted cell proportions in the simulated lung (A), pancreas (B), small intestine (C) and mouse melanoma (D) tissue data. The x-axis represents cell type and y-axis represents prediction accuracy. Predictions made by SSMD, EPIC, ICC and TICC were dark blue, green, yellow and orange colored, respectively. The red dash line represents the 0.9 correlation cutoff. (E, F) Correlation between true and predicted cell proportions in the two simulated brain tissue data. (G) The total rank of the gene expression profile of selected marker genes in the six simulated tissue data (gray), and the total number of cell types identified by SSMD in each data set or assumed in other methods (left three gray bars).

2.3.4 Experimental validation of SSMD by using matched RNA-seq and cell sorting data

We generated a tissue RNA-seq data of 11 mouse bone marrow tissue samples with matched cell counting using Fluorescence activated cell sorting (FACS) (see details in Methods). Application of SSMD on the RNA-seq data identified hematopoietic stem cell (HSC), general myeloid progenitor (GMP), mature myeloid cell and Pre-B cells, and their cell type specific markers. We also observed that the correlation between SSMD predicted and FACS measured amount of HSC, GMP, mature myeloid cell and B cells are 0.92, 0.8, 0.86, and 0.97, respectively, suggesting a high prediction accuracy of SSMD. Figure 2.3A-D shows the correlation between the SSMD predicted cell proportion and the FACS measured cell proportion of the four cell types. Figure 3E-H illustrate the FACS based cell counting of the four cell types. Complete cell type specific markers, cell proportions counted by FACS and predicted by SSMD were given. It is noteworthy that SSMD is not compared with other methods as none of the existing method is capable of predicting proportions of hematopoietic cell types.

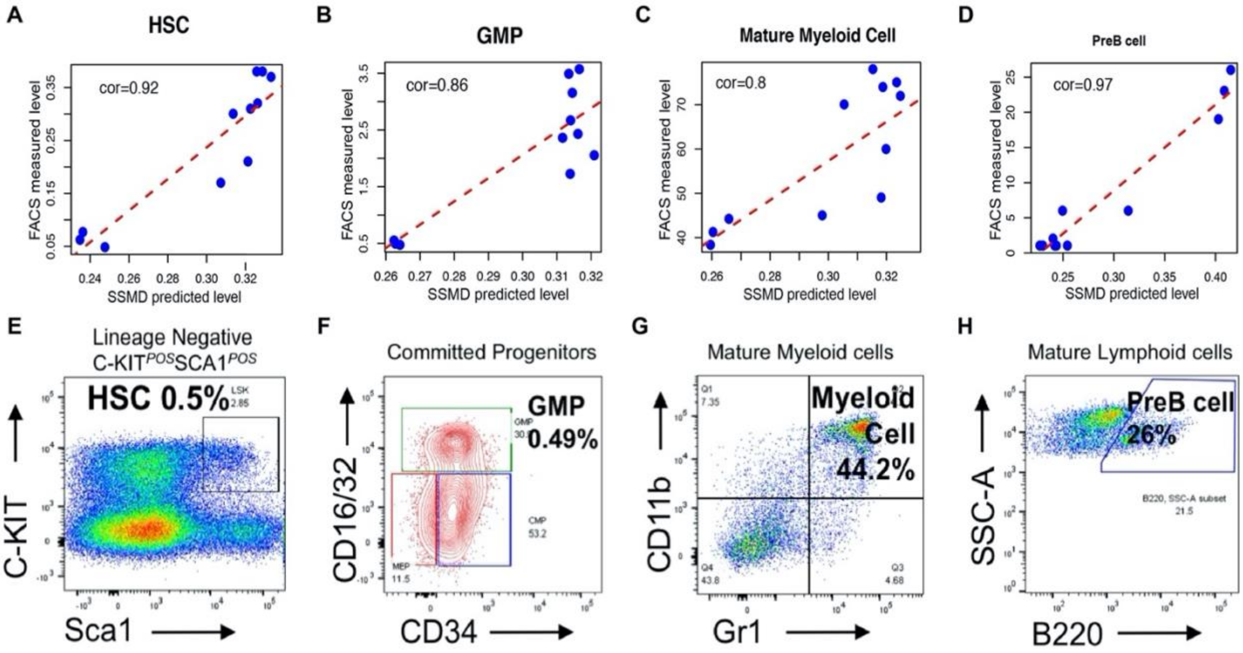


Figure 2.3. Method evaluation on hematopoietic tissue data. (A–D) Correlation between SSMD-predicted (x-axis) and FACS-identified (y-axis) cell proportions of HSC, GMP mature myeloid cell and pre-B cell. (E–H) Marker proteins utilized to identify the four cell types by using FACS. The x- and y-axis of the plots represent the level of cell type markers. The black block in (E), the green block in (F), the upper-right block in (G) and the block in (H) are the sorted HSC, GMP, myeloid and pre-B cell, respectively.

2.3.5 Application of SSMD to real mouse tissue transcriptomics data

We applied SSMD to nine cancer and eight central nervous system tissue data of four different experimental platforms, including one data set measured by immune-assay. On average, SSMD identified more than seven cell types in each of the cancer data, and the number of identified cell types is highly consistent with the total rank of the expression profile of the detected cell type specific marker genes (Figure 2.4A). This indicates that SSMD is capable of capturing the latent structure of the data. We further examined the explanation score (E-score), defined as the averaged absolute residual of the non-negative linear regression of each marker gene's expression on the predicted cell proportion, i.e. the average measure of how the predicted proportions could explain all the marker genes' expression levels. A high E-score is a necessary condition for an accurate cell proportion prediction. On average, the data set specific markers genes of each cell type identified by SSMD achieved 0.73 E-score while the average E-score of the marker genes used by EPIC and ImmuCC is 0.45 and 0.3 (Figure 2.4B). Similarly, application of SSMD on eight central nervous system tissue data identified more than seven cell types on average. The number of identified cell types is highly consistent with the total rank of the gene expression profile of the marker genes (Figure 2.4C). And the marker genes identified by SSMD achieved averaged 0.77 E-score for the cell types in central nervous system (Figure 2.4D). It is noteworthy that multiple marker sets of fibroblasts, myeloid or microglial cells that forming distinct rank-1 bases were identified in numerous data sets, suggesting the possible sub types of these cell types identified by SSMD.

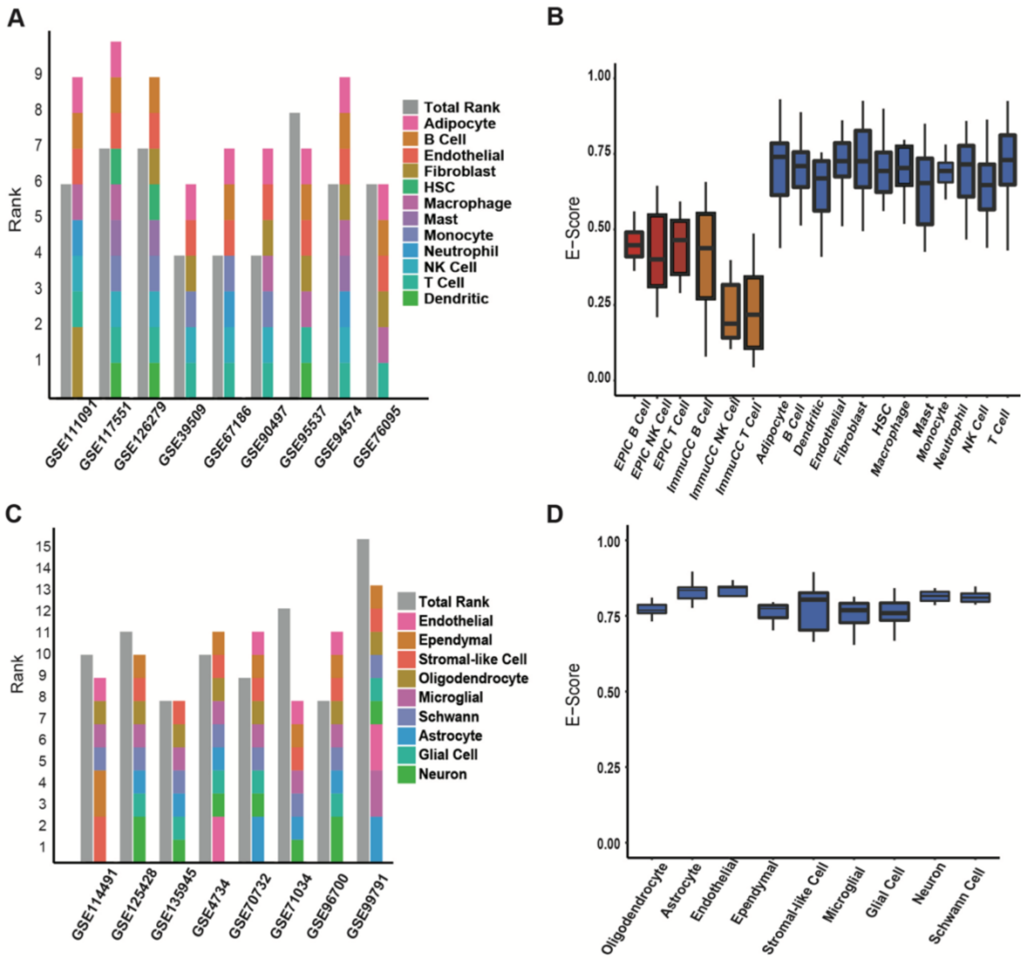


Figure 2.4. Prediction of SSMD on real tissue data. (A, C) The total rank of the gene expression profile of selected marker genes (gray) in different (A) cancer tissue and (C) brain data and the total number of cell types identified by SSMD in each data set (colored). (B, D) E-score for different cell types identified by SSMD (blue) in (B) cancer and (D) brain data set or assumed in other methods (EPIC: red, ICC: yellow).

2.3.6 Robustness analysis

We first evaluated the variation of cell type specific markers through different mouse strains on one transcriptomic dataset of mouse liver tissue samples collected from 31 different mouse strains [62]. To the best of our knowledge, this is the only dataset in the public domain that systematically measured gene expression profiles of the same tissue type for different mouse strains by using the same experimental platform. SSMD was applied to the data of each mouse strain respectively. 9 cell and their sub types were commonly identified in the liver tissue of most strains. The identifiability of the cell types and the detected cell type markers among different strains were compared (Figure 2.5). We analyzed all the identified marker genes that form rank-1 modules, i.e. the necessary condition for gene markers of identifiable cell types, and noticed that only 9.1% of the identified marker genes are shared in more than 50% strains, while 58.4% of the identified marker genes only served as a cell type marker in less than 20% of the analyzed strains, suggesting a high variation of cell type specific markers among different mouse strains, and the necessity to consider strain or data set specificity in deconvolution analysis.

We further examined the robustness of SSMD by evaluating its (1) sensitivity and (2) specificity in identifying cell types specific marker genes and its (3) accuracy in assessing of cell proportions, on the data of different sample sizes. Previous studies revealed that the robustness of the computation of co-expression correlation will decrease when the sample size is below 25. To comprehensively evaluate the method’s robustness, we selected five data sets, namely GSE76095, GSE67186, GSE90885, GSE94574, and GSE126279, with sample size ranging from 15 to 30 and randomly drew samples from each data set to build testing data sets of different sample size. We assumed the cell type markers and cell proportion inferred from whole data as “true” markers and proportions, and evaluated the consistency between the “true” ones and the ones predicted from small sub data sets. Accuracy in cell proportion prediction was assessed by the Pearson correlation between proportions predicted from small data and the “true” proportion on overlapped samples.

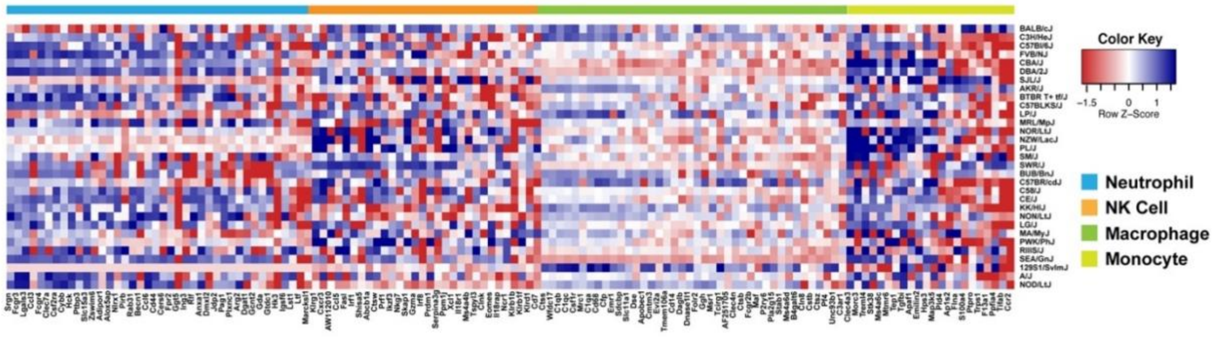


Figure 2.5. Correlation between expression level of strain-specific cell type marker genes and predicted cell proportion. High correlation is a necessary but nonsufficient condition for the genes to serve as marker genes of the cell types in corresponding mouse strain. In the heatmap, x- and y-axis represent genes and mouse strains, respectively. Genes in the core marker list of four selected cell types, namely neutrophil, NK, macrophage and monocyte, were colored on the column side bar.

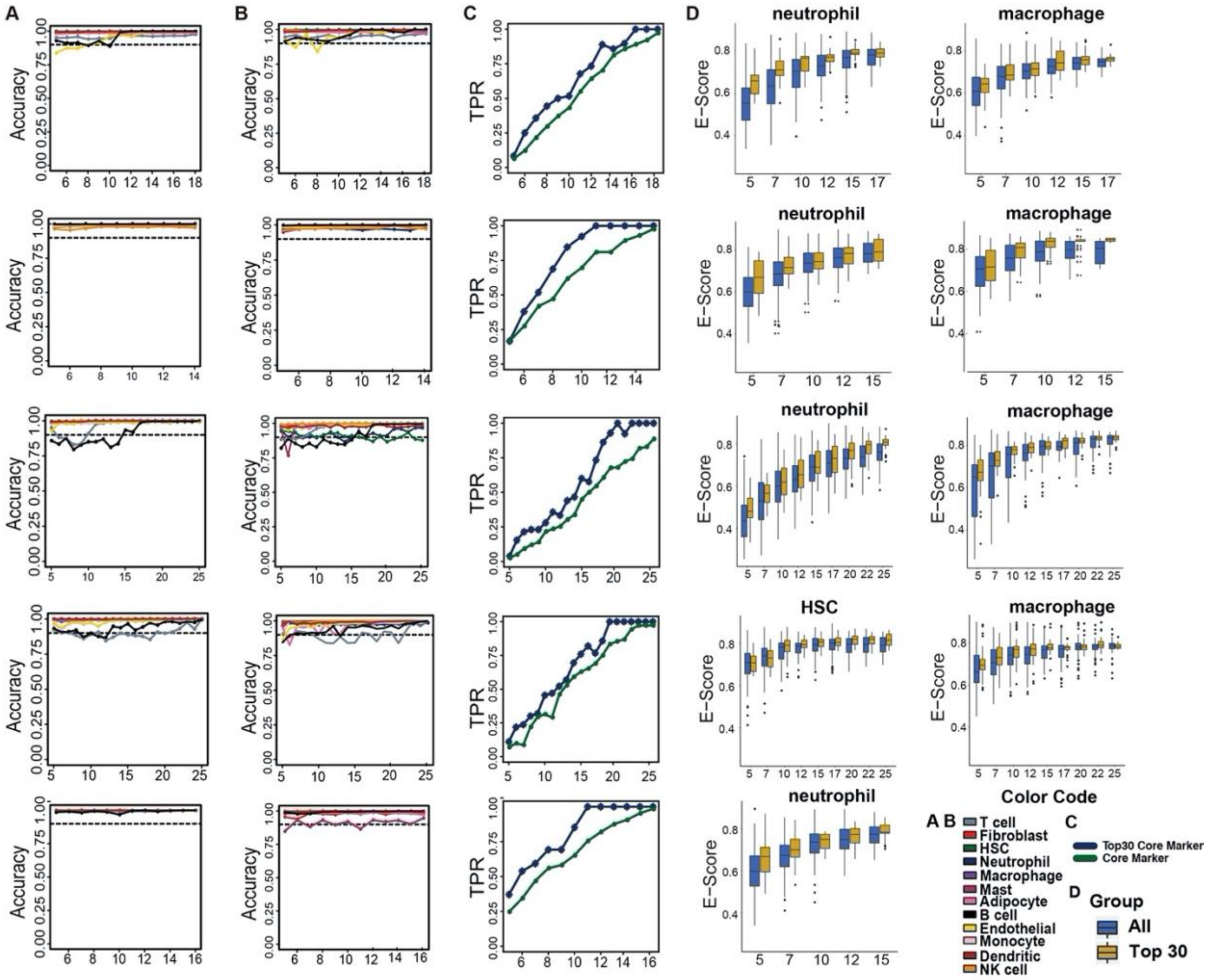


Figure 2.6. Performance evaluation of different sample size. (A) Prediction accuracy (y-axis) in different sample size (x-axis) using all core markers. Accuracy is the Pearson correlation between predicted proportion using only selected small sample and using all samples. (B) Prediction accuracy (y-axis) in different sample size (x-axis) using selected stringent markers. (C) True positive rate (y-axis) of the cell type-specific markers identified by using the stringent markers (blue) and core markers (green) with respect to different sample size (x-axis). (D) E-score for using coexpression modules consisting of all core markers and only selected stringent markers. From top to bottom, the statistics were derived from GSE76095, GSE67186, GSE90885, GSE94574 and GSE126279.

On average, all of the marker genes of the “true” cell types were also identified when sample size is low (Figure 2.6A). In addition, the cell proportion of 92.3%, 94.6% and 98.9% of the correctly identified cell types were with more than 0.9 correlation with their “true” proportions when the sample size is 6, 12 and above 20 (Figure 2.6A). Our analysis suggested a high robustness of the sensitivity and prediction accuracy of SSMD when sample size is as small as 6, i.e. the commonly used sample size in two-condition-comparison experiment (3 samples vs 3 samples). However, as a trade-off, there is a high false discovery rate of cell type specific modules when sample size is small, due to the low specificity of gene co-express analysis. To control the false discoveries on small data sets, we further derived a more “stringent” set of 341 cell type specific marker genes among the core marker set (see details in Methods). Our method validation demonstrated a slight drop of the sensitivity and prediction accuracy when using the stringent marker set on small data set (Figure 2.6B), while the specificity of the identified cell type specific markers increased to from 54.4% to 72.6% when sample size is above 12 (Figure 2.6C). Figure 2.6D illustrates the E-score of the cell type specific marker genes identified by using the core and the more stringent marker set with respect to different sample size. The E-score of the cell types marker genes identified by using the more stringent marker set were significantly higher than the ones identified by using the general core marker sets when sample size is below 10, also demonstrating the stringent core marker sets can effectively increase the analysis specificity when sample size is small.

2.4 Discussion

Over the years, research using well-established mouse models to mimic human conditions have provided extensive insight into the mechanisms underlying many human diseases. We developed SSMD to study mouse tissue microenvironment of complex traits, to mine the interactions of cell components in the microenvironment, which will feed back to studying human microenvironment. In order to have a robust prediction of cell component abundance in mouse tissue, SSMD detects a subset of the genes and identifiable cell types that are the most representative to the tissues to be analyzed, instead of using fixed gene signatures and cell types as in classic deconvolution schemes. The limitation in expression profiling and the intrinsic and mysterious variability in microenvironments excludes the possibility to have a unified set of cell type specific genes that have absolutely constant expression across all conditions. The way SSMD flexibly defines cell type marker genes mitigates the impact of variable marker genes due to experimental platforms and microenvironment alterations. This strategy allows our model to fully recapitulate the disparity of cell types and their marker genes across different microenvironment and data-generating platforms. In addition, the semi-supervised formulation enables the detection of sub cell types, which has been validated on scRNA-seq data simulated tissue data. Hence, a relatively coarse standard for categorizing the cell types was used in training the core marker list, which enabled a high robustness of the core markers. The unsupervised constrained-NMF or SVD-based deconvolution on the selected marker genes further excludes the adversarial batch effects.

It is noteworthy a successful identification of the rank-1 modules depends on a relatively large samples (>25) sharing cell types and marker genes. Currently, SSMD cannot be applied to the data with a single or small sample size. However, we consider such a tradeoff between sample size and prediction robustness is highly worthwhile, especially considering using SSMD as an exploratory tool in large scale publicly available mouse transcriptomics data. After all, the predicted proportions are often to be associated with other biological and clinical features, which will be severely underpowered with a small sample size.

We released a R package of SSMD via https://github.com/xiaoyulu95/SSMD and a web server via https://ssmd.ccbb.iupui.edu/. The input data is a mouse tissue transcriptomics data and user selected tissue specific cell type core marker sets. Currently, SSMD offers general core and stringent marker sets of 6 cell types in blood system, 12 cell types in normal, inflammatory and cancer tissue, 9 cell types in central nerve systems, and 14 cell types in hematopoietic systems. We have a practical guide for using SSMD of different tissues and sample size. The input of SSMD is a mouse tissue expression data set and user selected tissue environment category. The output of SSMD includes the identified data set specific cell type markers and the estimated sample-wise relative proportion of each identifiable cell type. We consider the currently included cell types are comprehensive enough to cover major cell types in mouse. However, the tissue specific cell types (for example, liver cells in liver tissue, colon cells in colon tissue, etc) were not included in our training scope. As forming rank-1 pattern among marker genes is a necessary but non-sufficient condition of identifiable cell types, SSMD R package can also output rank-1 modules that do not enrich the core markers of any cell type, which could possibly be markers of rare cell types. The user could further investigate whether the gene module corresponds to a real cell type or not. Another key feature of the webserver is that users are welcome to contribute their data to reinforce the training of cell type specific marker genes.

Potential future directions of SSMD include (1) enabling identification of cell type specific varied functions, which is not generally available for tissue data analysis in the public domain, (2) identifying data set specific cell type markers forming rank-1 submatrix in a subset of samples, i.e. local rank-1 submatrix, which can benefit from state-of-the-arts subspace clustering methods [69-71] and (3) extending and implementing the semi-supervised framework of SSMD with other state-of-the-arts deconvolution methods by refining data set specific cell marker genes. We anticipate that our computational concept, which is to identify data set specific and computationally “identifiable” cell types and their marker genes, can provide high robustness in deconvolution analysis, by which the predicted cell proportions can be reliably correlated with experimental features to provide biologically meaningful interpretation of the roles of microenvironmental changes in different disease tissues.

Chapter 3 PLUS: predicting cancer metastasis potential based on positive and unlabeled learning

3.1 Introduction

Metastatic cancer is responsible for over 90% of all cancer deaths [23, 24]. Compared with well-confined primary tumors, metastatic cancer remains incurable because of its systemic nature and the resistance of disseminated tumor cells to existing therapeutic agents [21, 22]. Hence, for a substantial number of cancer patients, effective treatment is largely dependent on an understanding of and capacity to interdict metastasis. Cancer metastasis is a multistep process by which cancer cells disperse from a primary site and progressively colonize distant organs. This process is often schematized as a sequence of discrete steps, termed the invasion-metastasis cascade [72-74]. Although advances have accelerated dramatically over the past decade and provided valuable insights regarding the molecular changes in the process of metastasis, metastatic cancer still represents an emerging field replete with major unanswered questions. In this context, evaluating a cancer patient's metastasis potential is vital for clinical decision-making and understanding the biological mechanism of metastasis is the first step towards targeted therapeutics.

Previous work has provided strong evidence indicating that a number of genomic markers in primary tumors are associated with the propensity of a patient to develop metastatic relapse, and that distant metastasis events can be inferred from gene expression profiles within the primary tumor bulk [25, 26]. For example, a recent study used machine-learning techniques to determine metastatic tumor organ of origin using the somatic mutation data [27]. Several studies have defined gene expression signatures that predict overall and metastasis-free survival as well as progression and metastatic growth in breast cancer patients [28-32]. Kikuchi et al. identified 40 metastasis-related genes by comparing lymph node-positive and lymph node-negative lung cancer patients [75]. Schell et al. [76] developed a score that can separate metastatic and non-metastatic tumors. Klein et al. [77] used the Cox multivariable proportional hazard model and survival C-index to evaluate the ability of a genomic classifier to predict metastasis and validated its robust performance. Goossens-Beumer et al. [78] performed differential MicroRNA (miRNA) expression analysis between metastatic and non-metastatic cases to establish miRNA-based metastasis risk predictions. Md Jahid et al. [79] proposed a personalized approach for improving the prediction of breast cancer metastasis.

On one hand, many of these metastasis predictors have been developed for a certain cancer type and thus are less generalizable to other cancer types. In fact, the molecular signatures of cancer metastasis reported in different studies hardly overlap [80]. Recently, the development of high-throughput sequencing technology has produced a large amount of molecular data at the pan-cancer level. Hence, harnessing the power of large-scale projects such as The Cancer Genome Atlas (TCGA) would enable us to systematically study the abnormalities in cancer progression at the molecular level in a statistically more powerful manner. On the other hand, to develop a cancer metastasis predictor, the following challenges remain unsolved by existing methods, which may largely limit the ability to predict early metastasis events and derive biological insights: 1) Metastasis events are not easily detectable, especially when we consider the hibernating disseminating cancer cells; thus, clinical metastasis diagnoses often tend to underestimate metastatic events. For classification-based methods, training a classifier with under-detected metastatic instances may lead to an under-estimated metastasis potential. 2) Many survival-based studies were principally designed to detect molecular markers that can best predict patient overall, progression-free, or metastasis-free survival, instead of directly targeting metastasis events itself. Therefore, the detected markers may not have any functional implication in metastasis. 3) High-dimensional molecular features complicate the classification and feature selection process. Therefore, a substantial refinement of the statistical considerations for model training and marker identification is required in order to increase the power to predict metastatic potential at the pan-cancer level. In summary, the combination of under-detected metastasis events and high-dimensionality in molecular features in transcriptomics data presents both statistical and computational challenges.

We have developed an algorithm called Positive and unlabeled Learning from Unbalanced cases and Sparse structures (PLUS) to address the aforementioned challenges. The ultimate goal of PLUS is to enable early metastasis event prediction at the pan-cancer level, as well as to infer biologically meaningful gene markers for metastasis potential. PLUS belongs to a category of classifiers called positive and unlabeled learning (PU learning). Whereas the input to a binary classifier normally consists of positive and negative incidence sets, in PU learning, a learner has access to only positive and unlabeled incidences, and it is assumed that the unlabeled data can contain both positive and negative incidences [81]. PLUS is particularly well-suited for studying metastasis potential: only patient samples diagnosed as metastatic are available and trustable, called positive samples; the samples that are not diagnosed as metastatic, due to a short follow-up time, are either metastatic or non-metastatic and are thus categorized as the unlabeled samples. In addition, PLUS is built on a penalized likelihood estimation framework for variable selection, and its iterative bootstrapping procedure makes it robust to bias caused by unbalanced allocations of positive and unlabeled samples.

PLUS represents a first-of-its-kind method to specifically address the under-diagnosis issue in studying cancer metastasis potential using the PU learning framework. Its robustness enables the power of big data to be harnessed through integration of large-scale datasets collected from different cancer types. Insights gleaned from this research will prove useful to the early diagnosis and treatment of metastatic disease. We benchmarked PLUS on extensively simulated data sets and demonstrated the superiority of PLUS over all other PU learning methods across all simulation scenarios. Application of PLUS to TCGA pan-cancer gene expression dataset resulted in metastasis potential estimations consistent with the clinical follow-up data. Moreover, PLUS selected a set of genes that are highly predicative of metastasis potential, and the differentiating potency of these genes was validated on independent single-cell RNA-sequencing (scRNA-seq) datasets, as well as existing literature.

3.2 Materials and Methods

3.2.1 Model setup

Let $x \in \mathcal{R}^p, y \in \mathcal{R}$ be a p-dimensional covariate and binary response variable. To model the probability of observing an event y conditioning on covariates x, logistic regression is commonly used to estimate the probability $Pr(y = 1|x)$ when both positive outcomes $y = 1$ and negative outcomes $y = 0$ occur in the observation. In the PU setting, however, only positive labeled instances are observed, while the labels for the remaining instances are unknown, or too noisy, as in the case of cancer metastasis diagnosis. In other words, we denote the observed outcome by z, where $z = 1$ represents positively labeled instances, and $z = 0$ represents unlabeled instances. For a subject i with $z_i = 1$, it clearly follows that $y_i = 1$. However, if $z_i = 0$, then either $y_i = 1$ or $y_i = 0$. The main purpose of PU-learning is to estimate $Pr(y_i = 1|x_i)$ from observed data tuples (x, z). Direct application of logistic regression in which z is treated as the response is severely biased. Because only part of y is observed, the PU problem can be viewed as a missing data problem, and one commonly used method for missing data problems is the EM algorithm [82]. In the following sections, we will first introduce the existing EM algorithm designed for the PU problem and then propose our PLUS algorithm, which is more tailored for the unbalancedness and sparsity issues in cancer metastasis prediction.

3.2.2 Case-control framework

We adopt the case-control framework proposed by Wald et al. [82], i.e., the positive instances are sampled from one distribution, deemed as "cases", while the negative instances are sampled from the other one, deemed as "controls". It is based on two reasonable assumptions:

Assumption 1: Positive instances are completely randomly selected from the positive population. In other words, whether an instance is observed as positive is regardless of its covariates x, that is,

$$Pr(z = 1|x, y = 1, s = 1) = Pr(z = 1|y = 1, s = 1);$$

Assumption 2: The unlabeled instances are a random sampling from the population, that is:

$$Pr(y = 1|x, z = 0, s = 1) = Pr(y = 1|x)$$

Here, $s = 1$ indicates that an instance is in the sample, which is always the case when we are working with the sample. The observed likelihood under this case-control sampling scheme is

$$L^{obs}(\theta|x, z, s = 1) = \prod_i P_\theta(z_i|s_i = 1, x_i)$$

$$= \prod_i P_\theta(z_i = 1|s_i = 1, x_i)^{z_i} \left(1 - P_\theta(z_i = 1|s_i = 1, x_i)\right)^{1-z_i}, \tag{1}$$

and the full likelihood is

$$L^{full}(\theta|x, y, z, s = 1) = \prod_i P_\theta(y_i, z_i|s_i = 1, x_i)$$

$$\propto \prod_i P_\theta(y_i = 1|s_i = 1, x_i)^{y_i} P_\theta(y_i = 0|s_i = 1, x_i)^{1-y_i}. \tag{2}$$

Direct optimization with respect to the observed likelihood function is difficult; thus, an EM procedure is introduced to accomplish the optimization according to the expectation of the observed likelihood.

E-step:

Given the estimated model parameter $\theta^{(k)}$ from the k^{th} iteration, the conditional expectation of full log-likelihood is thus

$$Q\left(\theta\middle|\theta^{(k)}\right) = E\left[\ell^{full}(\theta|x, y, z, s = 1)|x, z, s = 1, \theta^{(k)}\right]$$

$$= \sum_i \{E[y_i|z_i, x_i, s_i$$

$$= 1, \theta^{(k)}] log f_\theta^*(x_i) + (1 - E\left[y_i\middle|z_i, x_i, s_i = 1, \theta^{(k)}\right])\log\,(1 - f_\theta^*(x_i))\},$$

where $f_\theta^*(x_i) = P_\theta(y = 1|x, s = 1)$. We are also aware of

$$E\left[y_i\middle|z_i, x_i, s_i = 1, \theta^{(k)}\right] = P_{\theta^{(k)}}(y_i = 1|z_i, x_i, s_i = 1) = f_{\theta^{(k)}}(x_i)^{(1-z_i)},$$

where $f_{\theta^{(k)}}(x) = P_{\theta^{(k)}}(y_i = 1|x_i)$. This expression holds because when $z = 1$, the outcome y can only equal to 1, and when $z = 0$, we know from assumption 2 that $Pr(y = 1|x, z = 0, s = 1) = Pr(y = 1|x)$.

M-step:

In the M-step, we maximize the expectation of full log-likelihood described in the E-step with a penalty term $\lambda J(\theta)$ to account for the sparsity

$$\theta^{(k+1)} = arg\,max_{\theta}\, Q\left(\theta \middle| \theta^{(k)}\right) + \lambda J(\theta), \tag{3}$$

where λ is a penalty coefficient, and $J(\cdot)$ is a proper regularization function. Here, we adopt the L_1 norm to select informative variables. The penalized likelihood method based on EM was implemented in a similar manner as PUlasso [83].

After the M-step, we obtained $\theta^{(k+1)}$ as well as $f^*_{\theta^{(k+1)}}$, and to obtain $f_{\theta^{(k+1)}}$for the next E-step, we derive the connection between these two terms:

$$f_{\theta}(x) = \frac{(c-1)f^*_{\theta}(x)}{c - f^*_{\theta}(x)}, \tag{4}$$

where

$$c = \frac{Pr(y = 1|s = 1)}{Pr(z = 1|s = 1)}.$$

$Pr(z = 1|s = 1)$ is directly observed in the sample, but $Pr(y = 1|s = 1)$ requires knowledge of the population prevalence $\pi = P(y = 1)$. However, this parameter is unknown in our case, as we do not have prior information on the population prevalence of metastasis, which is indeed what we are seeking. A randomly assigned population prevalence may work as well when the two classes in the population are clearly separated and have a balanced presence; however, as we will see in the simulation data, this approach severely impacts PUlasso performance when the true prevalence is close to 0 or 1, or, in other words, when the population allocation is unbalanced.

3.2.3 PLUS framework

Even with the penalized likelihood estimation to enable feature selection, using such a framework similar to PUlasso for predicting cancer metastasis potential is not applicable for two reasons: 1) the observed unbalancedness, in which there are fewer positive samples (metastatic diagnosis) than unlabeled samples (non-metastatic diagnosis), and 2) unknown population prevalence, meaning that there is no prior knowledge on how many patients are metastatic in the population. Both challenges may significantly impact the performance of the case-control framework.

To address the challenges in existing algorithms, the proposed PLUS algorithm is particularly tailored to deal with potential unbalancedness in both observation and population allocation, along with a sparse data structure. Sparsity is solved by adopting a variable selection procedure, which can be naturally embedded in the EM structure with the LASSO penalty in the M-step. Unbalancedness can occur at 1) population level or 2) observation level as follows. 1) The population prevalence is extreme, that is, π is either close to 0 or close to 1, or 2) the number of observed positives is outnumbered by the unlabeled instances. Both types of unbalancedness, along with the PU setting, makes this problem even more complicated.

We rewrite c in equation (4) as

$$c = 1 + \Pr(y = 1|z = 0, s = 1)\frac{\Pr(z = 0|s = 1)}{\Pr(z = 1|s = 1)}, \tag{5}$$

where both types of unbalancedness are explicitly included. The population unbalancedness is expressed by $Pr(y = 1|z = 0, s = 1)$, since $Pr(y = 1|z = 0, s = 1) =$

$Pr(y=1)$ under assumption 2. Meanwhile, $\frac{Pr(z=0|s=1)}{Pr(z=1|s=1)}$ is a measure for observation unbalancedness. Clearly, when the positive prevalence is high or the unlabeled instances outnumber the positive instances, c will be a relatively large number. From S2 Figure, we find that the larger the c, the little difference between f_θ^* and f_θ. Consequently, when these unbalanced scenarios occur, the traditional EM-based algorithm approach is prone to perform little correction on the f_θ^*. In practice, this behavior causes PUlasso to fail in unbalanced situations.

According to this observation, we propose a new way to transform f_θ^* to f_θ, which does not require a knowledge of the population prevalence $P(y=1)$ and also works for unbalanced scenarios. Unlike the EM algorithm, in which $f_\theta^*(x)$ is always smaller than $f_\theta(x)$, our transformation adopts a bipolar function such that extreme estimated probabilities will become more extreme. Here, we choose a sigmoid function:

$$f_\theta(x) = \frac{1}{1+e^{-\alpha g\left(f_\theta^*(x)-p_0\right)}}, \tag{6}$$

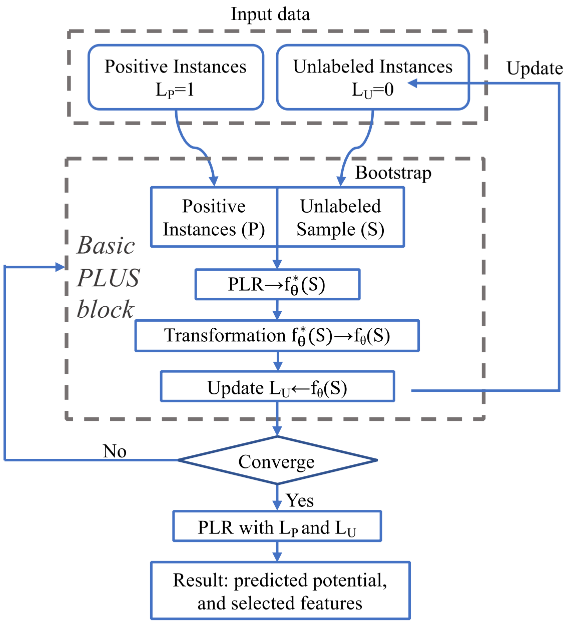
where p_0 is the anchor probability. Based on this sigmoid function, any $f_\theta^*(x)$ larger than p_0 will be projected from 0.5 to 1 or 0 to 0.5 for those smaller than p_0. p_0 is determined by the q_0-th percentile of the estimated probability for the positive cases, or $E(y|x,z=1,s=1)$. Here, we use the predicted probability of the positive samples to help distinguishing the unlabeled instances. PLUS is not sensitive to the choice of q_0 if the rank of probability is applied (see S3 Figure). $g(\cdot)$ is a function that linearly maps $f_\theta^*(x) - p_0$ to an arbitrary symmetric domain of the sigmoid function, for example [−1, 1], calibrated at 0. α is a scale parameter that determines the magnitude of transformation. In practice, this parameter parimarily determines the speed of convergence. We suggest choosing a value between 5 to 10 if the domain is $[-1, 1]$. In S2 Figure, we show a direct comparison of the sigmoid transformation and the EM transformation.

At each iteration, we 1) randomly sample the same number of observed positive instances from the unlabeled set with replacement and 2) conduct a one-step EM calculation, but only use only the new transformation function (6). In this manner, we can handle observation unbalancedness by maintaining a reasonably high ratio at each step. Then we 3) update the estimated probability for each unlabeled instance. Repeat step 1-3 until the estimated probabilities are stabilized. We take advantage of this bootstrap scheme to reduce the noise and increase the robustness. The details of the algorithm, as well as a flowchart, are given in Algorithm 1.

Note that, theoretically, the penalized logistic regression adopted in each iteration requires binary outcomes, while $E[y_i|z_i = 0, x_i, s_i = 1, \theta^{(k)}]$ typically lie between 0 and 1 and are not binary. However, most logistic procedures are currently able to handle non-integer responses, or we can work on augmented dataset as long as weights can be incorporated [24]. Thus, adopting $E[y_i|z_i = 0, x_i, s_i = 1, \theta^{(k)}]$ as responses is computationally feasible.

PLUS Algorithm:

```
Input: $X_{M\times N}$(Covariate matrix), the indices and labels of positive instances $P$, and $L_P$, its size $N_0$, and parameters $q_0, \alpha$.
Output: $f_{\theta final}$
Initialization: Labels for the unlabeled instances $L_U$=0
While stopping criteria are not met, do
    1. Randomly sample $N_0$ unlabeled instances with replacement, denoted as $S$, where $N_0$ is the number of positive instances.
    2. Run a PLR based on all positive instances $P$ and $S$. Record the estimated outcomes for $P$ and $S$, which are $f_\theta^*(P)$ and $f_\theta^*(S)$.
    3. Calculate $p_0$ based on the $q_0$-th percentile of $f_\theta^*(P)$.
    4. Perform a sigmoid transformation on $f_\theta^*(S)$ by $f_\theta(S) = \frac{1}{1+e^{-\alpha g\left(f_\theta^*(S)-p_0\right)}}$.
    5. Update the corresponding labels for S by $f_\theta(S)$:
       $L_U[S] \leftarrow f_\theta(S)$.
End;
Run a PLR based on $L_P$ and the new $L_U$, to yield the final estimation $f_{\theta final}$, which tells the selected features, as well as the class probability, i.e., $P(Y = 1|X)$.
```



To comprehensively assess the proposed method, we designed a series of simulation studies. Four different scenarios for the noise distribution were simulated, corresponding to a balanced population with two well-separated classes (a clear balanced scenario), a balanced population with two classes not well-separated (a noisy balance scenario), an unbalanced population with two well-separated classes (a clear unbalanced scenario), and an unbalanced population with two classes not well-separated (a noisy unbalanced scenario). Alterations to the population unbalancedness and separation are achieved by designing different propensity score functions as follows:

Clear Balanced Scenario: $logit(Pr(Y = 1)) = 2X_1 + 4X_2(X_3 + 5) - 3\,sin(X_4 + X_5) - 0.1X_6^4$

Noisy Balanced Scenario: $logit(Pr(Y = 1)) = 2X_1 + 4X_2X_3 - 3\,sin(X_4 + X_5) - 0.1X_6^4$

Clear Unbalanced Scenario: $logit(Pr(Y = 1)) = 2X_1 + 8X_2X_3 - 3\,sin(X_4 + X_5) - 3(X_6 - 1)^4$

Noisy Unbalanced Scenario: $logit(Pr(Y = 1)) = 2X_1 + 1.5X_2X_3 - 3\,sin(X_4 + X_5) - (X_6 - 1)^4$

The functions look similar, but their non-linear properties are distinguished. Here, we show distributions of $Pr(Y = 1)$ for each of the scenario. In addition, the direction of the population unbalancedness is also altered, causing the positive instances to be either the larger or smaller class.

We simulate the covariate matrix $X_{\mathrm{n\times p}} \sim MVN(0, I_{p\times p})$, where the total number of covariates p can take is 100, 200, and 400 and the sample size n is 2000. For all environments, only six of the variables are relevant to the binary outcome Y. To generate PU instances, we randomly flip the true label 1 to 0, and the probability of a sample being flipped determines the observation balancedness. A higher flipping probability corresponds to a lower observation balancedness, with the sum of these two values equaling 1. The level of observation balancedness ranges from 0.2 to 0.5. Its impact is shown in Figure 3.2 on the x-axis.

3.2.4 Benchmark methods and comparison

PLUS was compared with six state-of-the-arts PU learning or binary classification methods and one oracle reference method: (1) The PUlasso algorithm, which implements the EM algorithm for penalized likelihood estimation [83], and it utilizes a majorization-minimization framework to improve the stability of the EM-algorithm-based solution. The (2) Ada-KNN, (3) Ada-Logit, and (4) Ada-SVM methods all belong to a multi-method wrapper, called AdaSample [84]. AdaSample utilizes an adaptive sampling procedure to estimate the class mislabeling probability and to reduce the risk of selecting mislabeled instances. It is presented as a wrapper that can integrate support vector machine (SVM), k-nearest neighbor (KNN), logistic regression (logit), linear discriminant analysis (LDA), and feature weighted KNN [2]. (5) XGBoost is a state-of-the-art decision-tree-based classification algorithm that uses a gradient boosting framework [85]. (6) The random forest is a classification algorithm consisting of many decisions trees, which uses bagging and feature randomness when building each individual tree in an attempt to create an uncorrelated forest of trees whose combined prediction is more accurate than that of any individual tree [86]. (7) PRL places an L1 penalty on the logistic regression coefficients to enable variable selection. Notably, the PLR method was applied to the synthetic data using the true underlying label for method validation. For real data, PLR was also applied as a baseline method with the observed labels treated as true, which leads to an underestimated metastasis potential. Of all the methods, the Ada- methods, like most of the PU-learning methods, are incapable of variable selection. For each simulation setting, we conducted

100 repetitions. In each repetition, a predictive model is trained by each method, which generates an ROC curve with false and true positive rates at the x- and y- axis and the AUC. The average AUC among the 100 repetitions is used to evaluate the prediction performance for each method and simulation setting.

3.2.5 Data analyzed in this study

TCGA transcriptomics data. We retrieved the RNA-seq data from the PanCanAtlas [87]. The retrieved data includes the expression profiles of 20,531 genes in 10,332 samples from 33 cancer types. We combined COAD and READ into one cancer type called COADREAD. Among all cancer types, we extracted the 20 cancer types with at least one sample that was initially diagnosed as metastatic, with a total sample size of 7,467. Among these samples, 553 were confirmed to be metastatic. The detailed cancer types and corresponding frequencies of initial metastasis diagnosis. We first applied EB++ to remove the batch effect introduced by the different cancer datasets, as suggested by PanCanAtlas. EB++ is a recently developed batched effect removal method for TCGA pan-cancer data analysis that can adjust for sequencer platform differences (https://bioinformatics.mdanderson.org/BatchEffectsViewer/). Specifically, the UNC GA- and BCCA GAII-sequenced samples were separately adjusted to the UNC HiSeq data.

TCGA clinical information. Baseline and follow-up clinical information for all cancer samples was retrieved from the Genomic Data Commons Data Portal. Baseline diagnosis of a distant metastasis event was retrieved from the baseline clinical data, based on whether the patient is diagnosed as stage M1 in the TNM stage information. Specifically, a patient is determined to be diagnosed with metastasis if (1) the patient is in the M1 stage according to at least one of the “ajccmetastasis pathologic pm”, “ajcc metastasis clinical cm”, “clinical M”, and “pathologic M” criteria in the TCGA clinical information or (2) there exists direct evidence of metastasis under one of the following terms: “metastatic dx confirmed by”, “metastatic dx confirmed by other”, “metastatic tumor site”, “metastasis site”, “metastatic site other”, “metastasis site other”, “metastatic tumor indicator”, “metastatic disease confirmed”, “metastatic site”, and “other metastatic site”. Based on these criteria, 553 samples were confirmed to have a metastasis diagnosis while the remainder were treated as unlabeled samples.

To evaluate the accuracy of the predicted metastasis potential, we utilized the available TCGA follow-up data and collected PFS data. We specifically defined events in PFS as a patient having a new tumor event, whether this event was a progression of disease, local recurrence, distant metastasis, new primary tumor at any sites, or deceased with the cancer and no new tumor event, including cases with a new tumor event whose type is N/A, based on which the PFS specific to metastasis events can be computed. The unit of progression-free interval time is days. For the progression events, we collected either new tumor event dx days to or death days to, whichever was applicable, and for the censored cases, we collected either last contact days to or death days to, whichever was applicable. These collection criteria are in accordance with the existing literature [88].

scRNA-seq datasets. Two scRNA-seq data sets for human breast cancer (GSE75688) and head and neck cancer (GSE103322) were retrieved from the Gene Expression Omnibus database [89-91]. The datasets were selected based on the presence of cancer cells of varied metastasis status in the original bulk tissue or bulk cell samples. All data were downloaded as the counts or TPM profiles used in the original work. Cell labels generated in the original work were directly utilized. Specifically, GSE75688 includes data for 281 cancer cells from 10 cancer tissues, of which 3 have a high metastasis potential defined by metastasis to more than 2 lymph nodes and 7 have a low metastasis potential. GSE103322 contains data for 1040 cancer cells from 6 primary head and neck cancer tissues with extracapsular extension, a significant indicator of a metastasis event at a primary site, and 1468 cancer cells from 14 primary cancer tissues without extracapsular extension. Notably, the cancer cells from each scRNA-seq dataset can be classified as having high or low metastasis potential by the provided pathological or phenotypic information.

3.2.6 Analysis of scRNA-seq data

Both GSE75688 and GSE103322 were collected by using C1-Fluidigm or C1-SMART-seq protocol, and the sequencing saturation for both datasets are high. We selected only the malignant cells based on the cell labels provided in the original works. Seurat 3.0 was utilized for basic data processing [92]. All the analyzed cells have at least 1000 UMI measured. Cell clustering analysis was conducted via Seurat 3.0 with default parameters. Specifically, the cell clusters inferred from all genes were based on all genes with significant dispersion detected by Seurat, and the clustering inferred from the metastasis-predictive genes was based on the intersection of the metastasis-predictive genes identified by PLUS from the TCGA dataset and the significantly varied genes in each scRNA-seq dataset. The silhouette width $s(i)$ of each cell and the silhouette coefficient defined below were utilized to determine whether the cells from the tissues with higher metastasis potential were more closely clustered when the metastatic-predictive genes were used. We note that there are only two oracle cell clusters, namely, cells of high and low metastasis potential, in each dataset.

$$a(i|i \in C_m) = \frac{1}{|C_m| - 1} \sum_{j \in C_m, i \neq j} d(i,j)\,, b(i|i \in C_m) = \frac{1}{|C_{nm}|} \sum_{j \in C_{nm}} d(i,j)$$

$$s(i|i \in C_m) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$\text{Silhouette coefficient } (C_m) = mean\big(s(i|i \in C_m)\big),$$

where C_m and C_{nm} represent cells with high and low metastasis potential, $d(i,j)$ is the distance between cells i and j in the dimension-reduced space, and $s(i)$ is the silhouette width of the cell i. Notably, only the cells in C_m are utilized to compute the silhouette width and silhouette coefficient. The rationale here is that the silhouette coefficient can validate the use of the predicted metastasis-predictive genes and can better cluster the cells of high metastasis potential in independent scRNA-seq data. Here, 2D UMAP-based dimension reduction was utilized for the results visualization and computation of the distance $d(i,j)$. The average $s(i)$ values of the cells with high metastasis potential derived from the cell clusters inferred by using all genes or metastasis-predictive genes were compared, where a higher average silhouette width suggests a better clustering of the cells with high metastasis potential.

In addition to the silhouette width, we also assessed whether the PLUS-selected metastatic-predictive genes are more enriched by genes that are significantly associated with the metastasis status in each scRNA-seq dataset. For each gene, the association between the single cells' gene expression and their metastasis potential, which is defined as the metastasis potential of the patients where the cells were derived from, was tested by a Student's t-distribution-based test of their Pearson correlation, with p<0.01 as the significance cutoff. Genes with significant positive association with the metastasis potential are called meatstasis Enrichment of the scRNA-seq data-derived metastasis-associated genes in the TCGA data-derived metastatic-predictive genes was tested by a hypergeometric test.

3.3 Results

3.3.1 Problem formulation and methods overview

Diagnoses of metastatic cancer are often confirmed by detection of tumor masses at a distant site or effusions on clinical examination or by imaging [93]. Unfortunately, there is currently no panel of basic tests that can aid in revealing metastatic tumor events. Hence, many patients that are not diagnosed with metastatic tumors may have developed metastasis, but could not be diagnosed at an early phase due to weak symptoms (Figure 3.1a). Take the cancer patients enrolled in the TCGA project as an example. Among patients initially diagnosed as non-metastatic (M0), a large portion have a good prognosis and do not develop metastasis (M0: NP-Alive in Figure 3.1b). However, a significant portion of these patients do develop metastasis (M0: P-Alive in Figure 3.1b) or die (M0: Deceased in Figure 3.1b) based on their follow-up data. This is especially true for such cancer types as BLCA, ESCA, HNSC, LIHC, LUAD, LUSC, MESO, PAAD, SKCM, and STAD. This trend indicates a possible under-detection of metastasis at initial diagnosis. Given the relatively short follow-up time of these cancer types (Figure 3.1c), we believe that this discrepancy may be even greater if we considered longer follow-up data. In contrast, among patients who are initially diagnosed as metastatic, a majority develop metastasis (M1: P-Alive in Figure 3.1b) or die (M1: Deceased in Figure 3.1b) based on the follow-up diagnosis. This trend indicates that metastasis diagnoses, but not the non-metastasis diagnosis, are often trustable. These observations are the motivation of our proposed PLUS algorithm. PLUS is built upon the framework of PU learning, where patients diagnosed as metastatic and non-metastatic are treated as positive and unlabeled instances, respectively. Abbreviations for cancer types, and their initial diagnosis frequencies are given.

PLUS builds upon the case-control framework proposed in [82], and the modeling formulation of PLUS is detailed in the Materials and Methods section. Different from many other PU-learning algorithms, the solution of PLUS is achieved by incorporating both an EM-type algorithm and a bootstrap technique tailored for three main challenges that are particularly important for predicting cancer metastasis: (1) The genes that are informative in differentiating the true metastatic and non-metastatic classes are a sparse set of the whole transcriptome, representing a sparse structure. (2) The observed positive incidences (M1 diagnoses) are largely outnumbered by the unlabeled samples (M0 diagnoses), indicating observation unbalancedness. (3) One class (true metastatic patients) may be much larger or smaller than the other class (true non-metastatic patients), indicating population unbalancedness. Unfortunately, the true metastasis prevalence is usually unavailable. Specifically, for (1), we introduce a LASSO penalty into the objective function to select informative features [94]. For (2), we recursively bootstrap from the unlabeled instances of equal size to the positive instances to maintain a relatively high information ratio for the subsequent analyses. For (3), a sigmoid transformation is applied to the probability function in the EM procedure to account for the population unbalancedness.

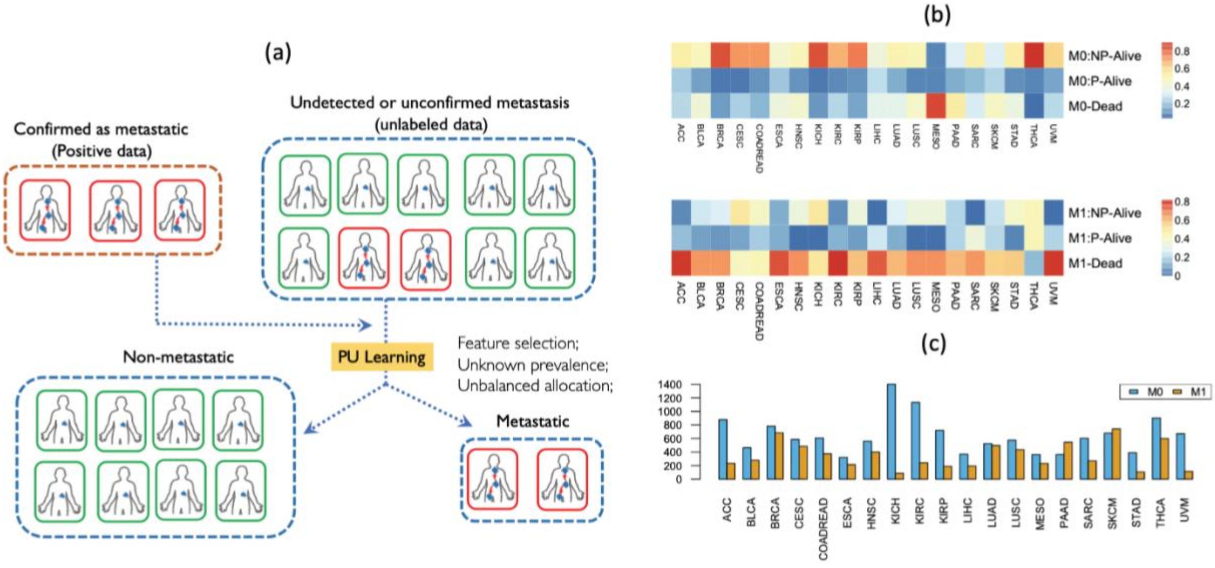


Figure 3.1. The motivation of PLUS. (a) Under-diagnosis occurred among patients who were not diagnosed as metastatic. To predict the metastasis potential for non-metastatic samples, PLUS builds upon the PU learning framework and is specifically designed to recognize the bias in under-diagnosis, and to address the computational challenges in feature selection, unknown prevalence, and unbalanced allocation. (b) For patients who were clinically diagnosed as non-metastatic (M0) at baseline for each cancer type (columns) in TCGA Pan-Cancer study, the top three rows show the proportions of patients with follow-up information who were found alive and with non-progressed disease (NP-Alive), alive and with progressed disease (P-Alive), and deceased (Deceased). The bottom three rows show the same proportions for patients who were diagnosed as metastatic (M1) at baseline. (c) The median follow-up time (y-axis) for patients who were diagnosed as non-metastatic (blue) and metastatic (yellow) at baseline for each cancer type (x-axis).

The general idea of PLUS is that we first bootstrap a subset of unlabeled samples equal in size to the positive set and perform a one-step EM-type procedure to generate the estimated probability for the unlabeled samples. We then repeated this procedure with another sample. The whole process stops when the assigned probability for unlabeled samples is stabilized in the recursive process. Ultimately, the algorithm provides the predicted probability of a sample being positive, denoted as the metastasis potential, as well as genes that may be predictive of metastasis potential.

3.3.2 Method validation on synthetic data

Using extensive simulated datasets, we compared PLUS with four other state-of-the-art PU-learning algorithms, namely, PUlasso [83], Ada-KNN, Ada-Logit, and Ada-SVM [84], as well as three popular binary classification methods, including the penalized logistic regression (PLR), XGBoost, and random forest. The input of each method includes the covariate matrix X (simulated gene expression matrix) and an observed response or label for either positive instances (purely true positive, corresponding to an M1 diagnosis) or unlabeled instances (consisting of true positive and true negative, corresponding to an M0 diagnosis). The output is the predicted probability of a sample being positive (i.e., metastasis potential) for each sample.

Basically, the covariate matrix X and true response y are linked by a logit model. In addition, to mimic the sparseness of gene features, the covariate matrix is simulated to contain both truly predictive covariates, as well as a large number of noisy features that do not contribute to the prediction of the response. To mimic the under-detection of metastasis, a certain portion of positive instances are randomly selected to be flipped to negative, resulting in an observed negative set that consists of both true negatives and true positives.

Both PLUS, PUlasso, and PLR are based on a logit model, and as a wrapper, AdaSample is capable of implementing logistic regression as its core. The XGBoost and random forest approaches are powerful tree-based methods for handling non-linear relationships between the high-dimensional predictors and responses. Hence, simulation data based on the logit model would have a minimal bias towards any methods. The simulation procedure is detailed in the Materials and Methods section. We evaluated the prediction accuracies of the methods in various simulation environments, where four parameters are altered: (i) the ratio of true positive and negative instances in the population, which is usually unobservable in real data scenarios, (ii) the level of separation of the two classes, or the noise level, where a higher noise level means that classes are less separable; (iii) the level of unbalancedness for observed positive and unbalanced instances, and (iv) the number of informative covariates among all features.

We assessed the prediction accuracy of all methods on only the unlabeled instances, which contain both true positive and negative instances. The average Area Under the ROC curve (AUC) calculated using the true labels was obtained from 100 repetitions of each simulation setting as the metric for methods evaluation. A higher AUC indicates a better prediction. Here, PLR was applied to provide an “oracle” estimation, as we intentionally provided the PLR with the true positive and negative label information, while the remaining methods were all provided with positive and unlabeled information. Hence, the performance of the PLR served as an oracle prediction that can be made if the given labels are all correct.

Figure 3.2 compares the performance of the PLUS, PUlasso, Ada-KNN, Ada-Logit and Ada-SVM, XGBoost, and random forest, as well as the reference method, PLR, which was provided with the true labels, under different settings: (1) Different levels of observation unbalancedness. On the x-axis of each figure, a higher ratio indicates that the number of observed positive samples is closer to that of the unlabeled samples, with 0.5 indicating that the numbers are equal. (2) Different levels of the population unbalancedness. From the left-most to middle to right-most images, we show results for simulation settings with more true positive than true negative samples, more true negative than true positive samples, and equal true positive and true negative samples. (3) Different noise levels of the logit model. From top to bottom, we show results for simulation settings in which the true positive and true negative sets are well-separated (top) and not well-separated (bottom).

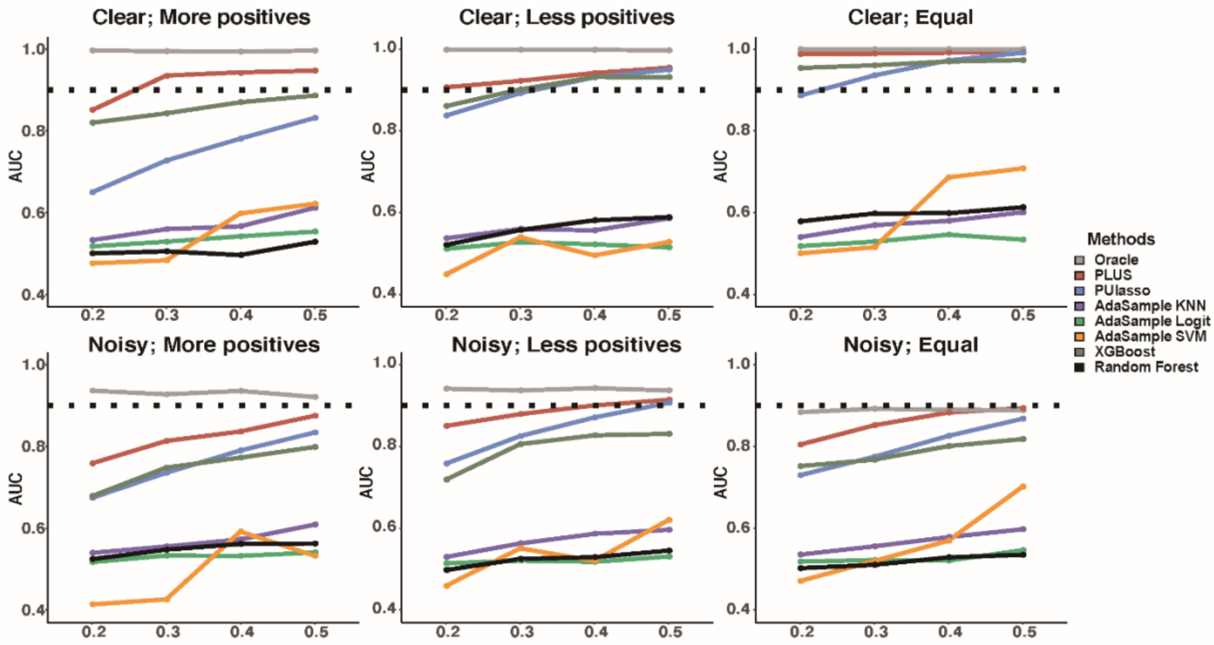


Figure 3.2. Performance comparisons of the competing methods on simulated data. The top panel shows results for simulation settings in which the positive and negative sets are clearly separable, while the bottom panel shows results for simulation settings in which the two sets are less separable. The left-most column shows results for simulation settings in which there are more true positive samples than true negative samples, the middle column shows results for less true positive samples than true negative samples, and the right-most column shows results for an equal number of true positive and negative samples. The dotted line corresponds to an AUC of 0.9. Here, the y-axis represents the AUC. And x-axis shows levels of observation unbalancedness, where a larger number indicates that the number of observed positive samples is closer to that of the unlabeled samples, with 0.5 indicating that the numbers are equal.

In summary, the three Ada- methods and the random forest all performed substantially worse than the other methods, with an average AUC of less than 0.6 for all simulation parameter settings. This result is probably due to the inability to handle high-dimensional features for Ada- methods or to handle unlabeled cases for the random forest, worsened by the population and observation unbalancedness issues. In general, all methods tend to have a higher AUC as the ratio of the observed positive samples increases (change on the x-axis in Figure 3.2). Among the methods, PLUS is the most robust to the ratio of the true positive samples, as it is specifically designed to handle the population unbalancedness issue with a bootstrapping procedure. In contrast, the performances of PUlasso and XGBoost fall sharply as the ratio of observed positive samples decreases. When the true positive and true negative sets become less separatable, all the methods tend to perform worse; here, PLUS still maintains an AUC above 0.8 for almost all scenarios, while PUlasso and XGBoost drop well below 0.8 for most cases. Interestingly, all tested methods tend to achieve the best performance when the ratio of true positive samples is lower (middle column), while they have the worst performance for a higher ratio of true positive samples (left-most column). This trend is reasonable because a higher rate of true positive samples corresponds to a higher contamination of true positive samples in the unlabeled cases, making it more difficult to achieve an accurate estimate on the distribution of true negative samples among unlabeled cases. After all, the best scenario for a prediction arises when all unlabeled cases are truly negative, with the least contamination of true positive samples. When we varied the ratio of informative features, our analysis suggested that the performances did not vary much for PLUS or PUlasso; however, this parameter moderately affects the performance of XGBoost and the random forest, and severely impacts the performance of the AdaSample methods. This result arises from the built-in model selection capability for PLUS and PUlasso whereas AdaSample methods cannot handle high-dimensional features, and the random forest and XGBoost are known to suffer from scalability/memory issues with high-dimensional features.

Overall, our analysis clearly suggests that PLUS achieves the best performance under all parameter settings over the other tested methods. The AUC of PLUS, averaging over 0.8 for all settings, is closest to the optimal AUC obtained by PLR. Notably, in our real data analysis on TCGA dataset shown in the next section, we have much fewer observed positive samples and a high-dimensional set of features, even though we do not know the ratio of true positive samples. Hence, we expect that PLUS will perform better than the other methods. The complete statistics of the methods evaluation based on simulations are provided at the GitHub link: https://github.com/xiaoyulu95/PLUS.

3.3.3 TCGA pan-cancer data analysis

Next, we applied PLUS to the transcriptomic profiles of all 7,467 cancer samples from 20 cancer types in the TCGA cohort. Among these, only 12 cancer types have at least 10 samples confirmed as metastatic at initial diagnosis, totaling 553 samples across the 12 cancer types. These 553 samples are treated as our observed positive samples, while the remainder are treated as unlabeled samples. Details on data pre-processing and sample metastasis diagnosis are provided in the Materials and Methods section. We applied PLUS, PLR, PUlasso, and XGBoost to this pan-cancer dataset. All four methods obtained the estimated metastasis potential as the probability of being metastatic for all the samples. Note that because we cannot observe whether metastasis developed in each patient, validating the classification accuracy using the ROC curve is impossible. Instead, we evaluate the performance of the methods on this real dataset by examining the association between the predicted metastasis potential with the progression-free survival (PFS) extracted from the TCGA clinical follow-up data, using only those patients that were initially diagnosed as non-metastatic at the time of tumor tissue collection. The event of disease progression is defined in the Materials and Methods section. The predicted metastasis potentials obtained by the four methods for all samples, as well as the PFS data for each sample.

	M0 sample	**PLUS**	**PLR**	**PUlasso**	**XGBoost**
ACC	62	**0.019**	1.000	1.000	1.000
BLCA	346	0.123	1.000	1.000	1.000
BRCA	1054	0.062	0.590	1.000	1.000
BRCA_TNBC	111	**<0.001**	1.000	0.887	1.000
CESC	294	**<0.001**	**<0.001**	**<0.001**	0.394
COADREAD	528	0.214	0.073	1.000	1.000
ESCA	167	**0.007**	0.349	1.000	1.000
HNSC	514	**<0.001**	**<0.001**	**0.003**	1.000
KICH	64	**<0.001**	**<0.001**	0.261	1.000
KIRC	452	**<0.001**	**<0.001**	**<0.001**	1.000
KIRP	278	**<0.001**	**<0.001**	**<0.001**	1.000
LIHC	367	**0.009**	**0.046**	**0.005**	1.000
LUAD	490	**<0.001**	**<0.001**	**0.003**	1.000
LUSC	495	0.062	0.091	**0.042**	0.387
MESO	84	**0.004**	**0.001**	**0.010**	<0.001
PAAD	174	**0.009**	**0.044**	**0.016**	1.000
SARC	203	**0.020**	0.145	0.139	1.000
SKCM	380	0.219	1.000	1.000	0.272
STAD	396	**0.019**	0.091	0.821	1.000
THCA	491	**0.020**	0.243	**0.042**	1.000
UVM	75	**<0.001**	**0.022**	**0.024**	1.000

Table 3.1. Significance of the association between PFS and metastasis potential predicted by PLUS (column 3), PLR (column 4), PUlasso (column 5) and XGBoost (column 6) using only the patient samples not diagnosed as metastatic (M0 samples). The p-values are adjusted for multiple comparisons. The second column shows the number of M0 samples for each cancer type.

Recognizing that different cancer types have different baseline metastasis potentials, we conducted an association analysis of between patients' PFS data and their predicted metastasis potential given by each method. The association analysis was conducted using the Cox proportional-hazards model considering continuous predictors [95]. The significance of associations between PFS and predicted metastasis potential, given by PLUS, PUlasso, PLR and XGBoost, is presented in Table 3.1, and the p-values were adjusted for multiple comparisons via the Holm method [96]. Clearly, 16 cancer types showed a significant association between the PLUS-predicted metastasis potential with PFS (p-value < 0.05), with higher predicted metastasis potentials related to worse survival outcomes. For BRCA and LUSC, we observed marginally significant associations (p-value < 0.08). Notably, for BRCA, we observed a significant association for its most aggressive subtype, namely, the triple negative breast cancer (TNBC) (p-value < 0.001). We did not observe significant associations between PLUS prediction and PFS for BLCA, COADREAD, or SKCM. We argue that the follow-up time for BLCA and COADREAD is too short (see Figure 3.1c), with the follow-up occurring before a metastasis event could be confirmed. Overall, we have demonstrated that for patients initially diagnosed as non-metastatic, PLUS is able to predict the metastasis potential for these patients and detect possible under-diagnosis incidences; moreover, its predicted metastasis events are in strong accordance with true metastasis events based on the follow-up data. For PLR, PUlasso, and XGBoost, strong associations were identified between the predicted metastasis potential and PFS in 9,10, and 1 cancer types, respectively. To visually compare the performances of the four methods, we also compared PFS for high and low metastasis potential groups within each cancer type. Specifically, for each cancer type, we divided the patients (initially diagnosed as non-metastatic) into two equal-sized groups, one with high metastasis potential and another with lower potential, according to each method. Figure 3.3 shows a survival comparison between the two groups for all 20 cancer types and 1 subtype based on PLUS and the three other methods. Our results and comparisons clearly demonstrate the advantage of applying PLUS for predicting metastasis potential, owing to its robustness in the PU learning setting, with possible unbalancedness in sample collection and population incidence.

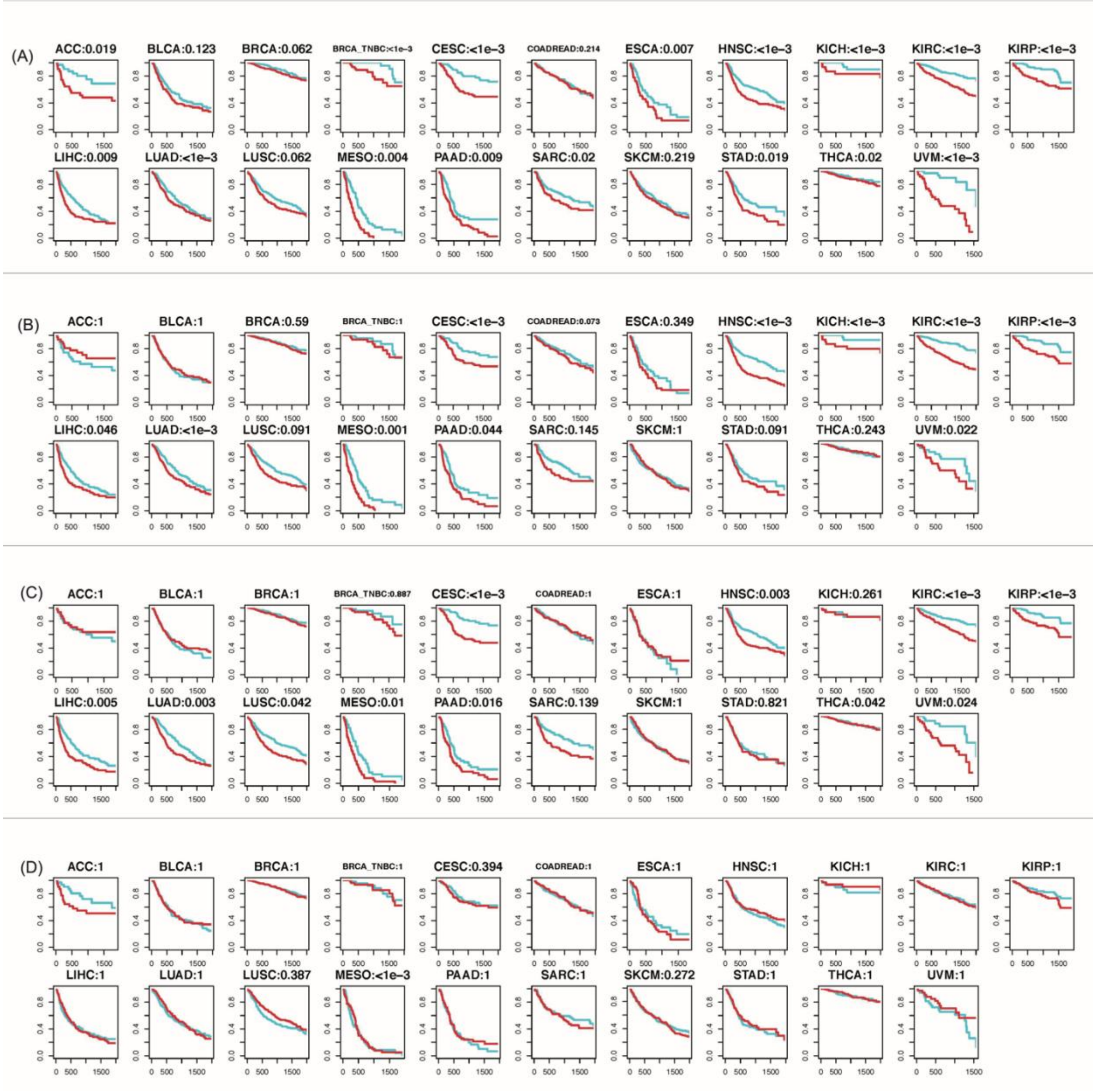


Figure 3.3. Progression-free survival curves of patients with different metastasis potentials. Higher metastasis potential (red) and lower metastasis potential (blue) for 21 cancer (sub)types are predicted by (A) PLUS; (B) PLR; (C) PUlasso; and (D) XGBoost. The x-axis represents time in days and y-axis represents the percentage of patients without metastasis event at a particular time point.

Together, Table 2.1 and Figure 3.3 strongly suggest that the metastasis potential predicted by PLUS is highly consistent with the actual follow-up data for many different cancer types. This finding has three primary implications: (1) Gene expression at early stages can predict the propensity of patients to subsequently develop metastasis. (2) PLUS is the first prediction tool for cancer metastasis that works for a general set of cancer types, by harnessing the power of large-scale data integration. (3) The success of PLUS in predicting cancer metastasis potential further confirms that cancer metastasis is often under-detected, posing a threat to timely disease management.

3.3.4 Functional mechanism of the metastasis predictive genes

A total of 191 metastasis-predictive genes were identified by PLUS that optimally predict metastasis potential in the TCGA pan-cancer data. A complete list of the 191 genes is provided. We first evaluated functional clusters of these genes by a pathway enrichment test against Msigdb canonical pathways and Gene Ontology [97]. The top 50 enriched pathways are mainly related to (1) responses to the oxidative stress such as hydroperoxide; (2) the regulation of calcium ion transport, and (3) responses to cytokines. More details are given. These pathways are known to be closely associated with cancer metastasis. In particular, hydrogen peroxide has been viewed as a "fertilizer“ of inflammation, cancer metabolism and metastasis [98], and metastasis is the route for cancer cells to escape from the oxidative stress [99]. Moreover, the calcium ion is a ubiquitous second messenger that acts as crucial regulator of cell migration [100], and cytokines are central mediators in remodeling the local microenvironment to support the growth, survival, and invasion of primary tumors and enhance metastatic colonization [101].

We further investigated the correlations between all the individual genes and the PLUS predicted metastasis potential and selected the genes with significant positive correlations. Similar enrichment analysis demonstrated that the pathways positively associated with metastasis potential are non-surprisingly well-known metastasis-related pathways, and the most highly enriched genes are related to the immune system and inflammatory responses, extracellular matrix organization, and angiogenesis. This functional enrichment analysis presents partial evidence for the concordance of the PLUS-selected genes with the current body of literature. Below, we will examine whether these genes are truly potent in differentiating the metastasis potential of cancer cells using single-cell data.

3.3.5 Validation of metastasis predictive genes in independent scRNA-seq datasets

To validate the metastasis-predictive genes selected by PLUS from TCGA pan-cancer data, we collected two scRNA-seq datasets of human breast and head and neck cancer. Both data sets contain cancer cells from cancer bulk tissue samples with different metastasis statuses. We first conducted cell clustering analysis on each dataset by using (1) general genes with high expression dispersion and (2) the 191 metastasis-predictive genes identified by PLUS. As reported in the original works [89-91], cancer cells from different patients possess strong inter-tumoral heterogeneity and tend to cluster together. Hence, the cancer cells in both breast and head and neck cancer data sets can be separated into two groups of primary cancer cells with high and low metastasis potential. A silhouette coefficient [102] was applied to determine whether cells of different metastasis potential are closer together in certain cell clustering results. Specifically, a larger silhouette coefficient value indicates that cells tend to be clustered together if they have similar metastasis potentials.

In the scRNA-seq data for breast cancer (GSE75688), the cancer cells were collected from three cancer tissues with high metastasis potential and seven tissues with low metastasis potential, determined by the number of lymph node metastases. Our analysis gave silhouette coefficients of 0.07 and 0.36 for the cells with high metastasis potential in the formed clusters when using all genes (see Figure 3.4a) and the PLUS-selected metastasis predicative genes (see Figure 3.4b), respectively. In the scRNA-seq dataset for head and neck cancer (GSE103322), cancer cells were collected from 6 patients with pathologically detected extracapsular extension, a significant indicator of a metastasis event at the primary site [103], and 14 patients without extracapsular extension. Cell clusters inferred by using all genes form distinct patient-specific groups (see Figure 3.4c), which does not show a strong dependency on the extracapsular extension event. In contrast, the cell clusters inferred from the PLUS metastasis-predictive genes clearly form two groups of cells, one from extracapsular extension cancer and one group of cells from the 14 cancer samples with lower metastasis potential (see Figure 3.4d). The average silhouette coefficients for the cells from extracapsular extension cancer are 0.1 and 0.3 in the cell clusters obtained by using all genes and the PLUS metastasis predictive genes, respectively. In addition to the performance of cell clustering analysis, we observed that TCGA-derived metastasis-predictive genes are significantly enriched by metastasis-potential-associated genes (24/147, p = 0.02 and 28/162, p = 0.0059), compared with background result (3037/31656 and 1977/21030) for the breast and head and neck cancer data, respectively.

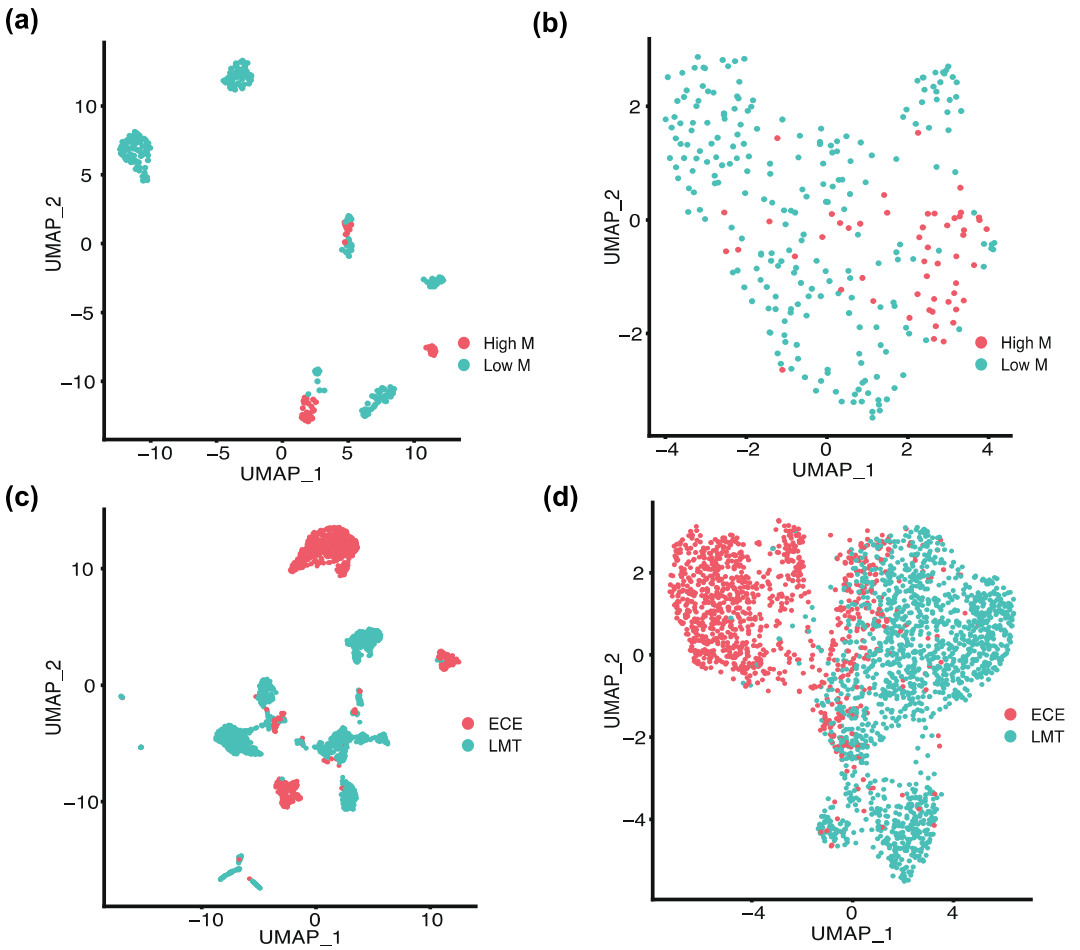


Figure 3.4. Functional validation of the metastasis-predictive genes inferred by PLUS from TCGA data. (a-b) Cell clusters of the cancer cells with high (High_M, red) and low (Low_M, blue) metastatic potential obtained by using all genes (a, silhouette coefficient = 0.07) and the PLUS selected metastasis predictive genes (b, silhouette coefficient = 0.36) in the scRNA-seq dataset for breast cancer (GSE75688). (c-d) Cell clusters of cancer cells from cancer tissues with (ECE, red) and without (LMT, blue) extracapsular extension obtained by using all genes (c, silhouette coefficient = 0.1) and the PLUS-selected metastasis-predictive genes (d, silhouette coefficient = 0.3) in the scRNA-seq dataset for head and neck cancer (GSE103322).

Our analysis of independent scRNA-seq data sets clearly demonstrates that the 191 metastasis-predictive genes derived by PLUS from TCGA pan-cancer data are relevant to metastasis. Details regarding the analysis approaches applied to the scRNA-seq data are provided in the Materials and Methods section. Complete statistics and codes of the scRNA-seq data analysis are available at https://github.com/xiaoyulu95/PLUS.

3.3.6 Robustness analysis

To evaluate the robustness of PLUS on the TCGA pan-cancer data, we intentionally removed the data from one cancer type at each time, and then ran PLUS using only the remaining samples. We examined the robustness of PLUS based on the overlap of the selected genes, as well as the correlations of the predicted metastasis potential on the M0 samples. We showed the number of overlapping genes and correlations of predicted metastasis potential, for any two PLUS predictions made with two different cancer type data removed. We also included the prediction with no samples removed (labeled as "ALL"). On the left panel, we observed that the correlation between any two predicted metastasis potential is consistently high, with a minimum correlation of 0.43, and a median of 0.90. For the overlapping genes, the minimum number of overlapping genes is 34, with a median of 116, for any pair of predictions. Note that PLUS implements a sparsity assumption on gene selection using L_1 penalty, which may make the gene selection less stable. It is well known that sparsity and stability are at the odds of each other, especially when there is strong feature collinearity in the data [104], which may be the case for our gene expression data. Interestingly, six genes, including ALS2, DAPL1, HS6ST1, IGFBP2, MGC12982, PPIAL4C, are selected in all predictions, i.e., when no sample is removed, or one cancer type data is removed. The six genes are highly potential to be indicators of metastasis potential given their robustness, though they certainly warrant further experimental validation.

3.4 Discussion

Metastasis is the major cause of cancer-related deaths, and evaluations of metastasis risk are essential for tailored treatment of cancer patients. Existing computational tools for predicting the cancer metastasis potential fall under two categories: 1) methods that build a classifier using the clinical metastasis diagnoses as responses and 2) methods that evaluate the behavior of gene features found to be significantly associated with metastasis-related survival outcomes. Such predictors exist in many even for the same cancer type; however, selected gene features rarely overlap, not to mention the little consistency of metastasis predictor genes among different cancer types. Thus, there is an urgent need for a powerful tool to characterize the cancer metastasis potential and to delineate the important gene features of cancer metastasis that is applicable across a wide span of cancer types.

Traditional classification methods for predicting cancer metastasis overlook an important fact in cancer metastasis diagnosis: while it is easy to confirm metastasis events with detected metastatic cancer cells in lymph nodes or distant locations, it is much more challenging to confirm non-metastasis events. Disseminating cancer cells may undergo hibernation, temporarily causing few or no complications in the patients, and clinical procedures are often not sufficiently accurate to capture ongoing events. In both cases, we see that cancer metastasis events tend to be under-diagnosed. Comparing the initial metastasis diagnosis and follow-up metastasis occurrence in TCGA pan-cancer clinical data confirmed this unfortunate finding: despite an initial non-metastatic diagnosis, many patients of various cancer types develop metastasis in the following years (see Figure 1b). A good classifier for metastasis should be designed to account for this under-diagnosis issue. However, finding prognostic markers from survival data by treating metastasis as a censored event may not reveal the genes with true biological and functional relevance to metastasis.

Our proposed PLUS algorithm builds on the framework of PU learning by considering patients with metastasis diagnosis as positive instances and the remainder as unlabeled instances, meaning they are either metastatic or non-metastatic. Under this framework, the selected genes become truly relevant to the biology of metastasis. Indeed, the classifier given by PLUS rendered concordance between the predicted cancer metastasis and observed metastasis survival outcomes in the follow-up data for almost all cancer types considered. The selected genes were found to perform functions consistent with experimental research findings and are capable of clustering the single cells based on their levels of metastasis potential. PLUS fully exploits the power of big data by training on ~7,000 patients samples, where only a very small portion are diagnosed as metastatic samples. The superiority of PLUS over other methods lies in its tailored designed that overcomes the high-dimensionality of gene features, the unbalancedness issue in instance allocation (more non-metastatic than metastatic diagnoses), and the possible unbalancedness in the underlying population distribution (unknown population prevalence of metastasis), which fully recapitulates the case of cancer metastasis. The computational tool designed and insights gained from this research will prove useful to the diagnosis and treatment of clinical metastatic disease.

Notably, while different cancer types and subtypes may have different metastasis mechanisms, the successful application of PLUS to pan-cancer data demonstrates its power to identify common hallmarks for early metastasis prediction across cancer types, confirming the accuracy, reliability, and robustness of this model. In addition, the gene markers identified by PLUS are related to early metastasis events, including a series of actions for invading cancer cells to overcome stromal barriers, survive in the circulation system, and settle and colonize at a distant metastasis site, which have been revealed as common metastasis hallmarks for diverse cancer types. In fact, researchers have been harnessing the power of big data by integrating the omics data of multiple cancer types to find biomarkers that underlie a common pathway of oncogenesis and particularly the EMT process [105, 106]. As a result, a pan-cancer EMT signature gene has been discovered that is independent of cancer types [107]. These findings suggest the rationality of applying PLUS to pan-cancer data.

Chapter 4 Detecting cell-type specific variations of within pathway interaction in AD using covariance regression

4.1 Introduction

In 2021, an estimated 6.2 million Americans aged 65 and older were living with Alzheimer's Disease (AD), and this number is projected to reach 13.8 million by 2060 [108]. As the population ages, AD and related dementias will become a major public health concern. Currently, no existing treatments can reverse, stop, or slow down the associated progressive neuronal and neurophysiological changes that occur in patients. Therefore, a deeper understanding of the disease pathology on a molecular and cellular level is critical to better understand the mechanisms behind AD and devise treatments to slow down the disease progression. Currently, large-scale multi-omics data from brain tissues, have been collected, and analysis of multi-omics data on the postmortem human brain bulk tissues has led to numerous findings including those on the epigenomic and transcriptomic signatures of AD [109-121]. However, the bulk tissue profiles fail to capture the cell type-specific abnormality in the disease progression. For example, the following pathways are known to be varied in AD for only specific cell types: synapse pathway in neuron cells [122, 123], inflammatory pathways in microglia [124, 125], protective myelin sheaths retraction in oligodendrocyte [126-128], etc. Recent advancement in single-cell technology provides new avenues for molecular profiling at the single-cell resolution [129], which improves the -omics studies by making increasingly greater precision and granularity possible [130-132], and this is particularly true for AD research [133-139]. Both gene-level and pathway-level analysis have been carried out using single cell data. It has been known that compared to the pathway-level analysis, single-gene analysis may miss important effects on pathways, realizing that cellular processes often affect sets of genes acting in concert [140, 141]. For example, genes in the same pathway may present weak but consistent patterns, but they are very likely to be disregarded due to low statistical power. Moreover, knowledge on single genes is less robust and transferable across studies because when different groups study the same biological system, the list of statistically significant genes from the studies may show distressingly little overlap [36].

Currently, a common approach to aggregate the individual gene signals involves pathway enrichment methods and co-expression module detection using either the differentially expressed genes or genes' importance ranking [36, 37]. On one hand, the current enrichment or co-expression based pathway analysis methods suffer from the selection of a proper threshold, and the biggest unmet need in the current pathway-level analysis of scRNA-Seq data is the lack of a rigorous and powerful statistical framework to make inferences on important variables such as disease status, sex and age [142, 143] with limited sample size; on the other hand, existing research on AD disease gene and pathway detection are mainly focused on first-order analysis, while few methods were designed to model the changes in the interactions of the genes in a statistically solid manner, namely, the gene-gene covariance structure in the same pathway associated with the disease progression. Biological processes are not chiefly controlled by individual proteins, but rather by a complex system-level network of molecular pathways [140, 141]. Understanding the changes of molecular interaction in biological pathways may lead to discoveries of pathogenically dysregulated pathways that are not detectable by merely analyzing absolute abundance level changes. And such discoveries are equally crucial to understanding complex phenotypes in AD. In summary, despite the power of scRNA-Seq technology in dissecting the cell type heterogeneity and delivering high-resolution molecular mechanisms for AD, current existing analytical approaches lack sufficient statistical power to detect robust or higher order changes on a pathway level, that could largely completement traditional single-gene or enrichment analysis methods [144].

To bridge this gap, we introduce a statistically powerful framework based on covariance regression [145], to model the pathway level second-order variations using scRNA-Seq data, namely, single cell Covariance Regression (scCovReg), and to associate the second-order variations with important subject-level characteristics, such as disease status. Covariance regression has been utilized in studying regression problems when the outcome variable is a covariance matrix [38-44]. In our case, when studying the impact of subject-level characteristics on within-pathway gene-gene correlations, the pathway covariance structures of single cells are regressed over the covariates. We call this the covariates-explainable Gene-Gene Correlation (eGGC). Importantly, this covariance regression model will enable us to draw inference on the statistical significance of the considered factors, as well as their interactions, on how well they could explain the gene correlation changes among the single cells. In addition, for each pathway, our covariance regression-based model enables finer analysis on its individual gene members. For example, one may find the correlation patterns among a few genes in a pathway are the best differentiable among healthy control and AD patients, while the correlation changes among other genes are not explainable by the disease status.

We applied the scCovReg pipeline on the Religious Orders Study and Memory and Aging Project (ROS/MAP) single nucleic RNA-Sequencing dataset, for 10,402 pathways collected from the Gene Ontology (GO) database [146]. It remains our key novelty in using the covariance regression technique for modeling scRNA-Seq data to detect the cell type specific second order changes of many pathways that are possibly attributable to important subject-level variables. We have discovered that: 1) most of the pathways tend to have lower eGGC strengths in AD cells than in healthy cells for cell types including astrocyte, neuron, and microglial in females, and neuron cells in males. 2) Compared to males, females demonstrate a much larger number of pathways with significantly different eGGC levels between AD and healthy conditions, in neuron and astrocyte cells. 3) By categorizing the pathways into 17 categories, we consistently observed that among females, the pathways are more likely to have differential eGGC levels between AD and healthy astrocyte and neuron cells. 4) Compared to healthy subjects, different subtypes of neurons and astrocytes tend to be more homogeneous in terms of eGGC in AD patients among females, indicating a loss of functional specialization in diseased cells for female. This is not observed in male subjects. 5) Compared with the traditional first-order based pathway enrichment method, the scCovReg pipeline gives rise a much larger number of pathways that are significantly different between AD and healthy conditions for most of the cell types considered, particularly in female subjects. This indicates the necessity of delineating the molecular level changes for AD pathological pathways in terms of their second order changes. Ultimately, our approach may aid the transition from a limited single-alteration perspective in disease to a comprehensive network-based mindset, which will potentially result in precision medicine paradigms for disease diagnosis and treatment.

4.2 Materials and Methods

4.2.1 Data access and processing

Religious Orders Study and Memory and Aging Project (ROS/MAP): The ROS/MAP dataset is a longitudinal cohort of aging and dementia in elderly nuns, priests, and brothers. The cohort includes rich clinical data collected annually, detailed post-mortem pathological evaluations, and extensive genetic, epigenomic, transcriptomic, proteomic, and metabolomic bulk tissue profiling. The ROS/MAP single-cell data were collected from 24 AD patients and 24 healthy controls with matched age and sex. In total, ~80,000 single cells spanning six cell types were profiled. Among them, five cell types including astrocyte, microglia, neuron, oligodendrocyte, and oligodendrocyte progenitor cell (OPC) have sufficient number of cells present in each individual, and were kept for further analysis.

GSE157827: This dataset contains 21 prefrontal cortex tissue samples from patients with AD (8 male and 4 female) and NC subjects (6 male and 3 female). Single-nucleus RNA sequencing was conducted, and 179392 single cells were collected, where four cell types are present: astrocytes, neuron, microglia and oligodendrocyte. This dataset serves as a validation dataset.

4.2.2 Covariance regression on scRNA-Seq data

A covariance regression approach [145] was employed to capture the population/individual variations in gene-gene interaction. This approach was originally designed to study the variations in brain functional connectivity. Here, we applied it to the single-cell sequencing data to reveal the association of gene-gene interaction with AD, sex, and their interaction and examine the discrepancy between subgroups with three comparisons, including (1) AD vs. control in female, (2) AD vs. control in male, and (3) male vs. female in AD. The covariance regression approach assumes that there exists a linear projection such that in the projection space (called a component), data variation satisfies a log-linear model of the covariates of interest. Thus, the approach uncovers the variations in gene-gene interaction at a network level. In addition, compared to a pair-wise modeling approach, statistical power is significantly improved by using a covariance regression approach. Before running the covariance regression, two steps of data processing were performed: a screening step to remove genes with 0 expression in over 80% of cells and a data-transformation step to make the distribution close to normal. The covariance regression was applied to each cell type and each pathway separately. For each model, multiple components might be identified, where the number of components was determined using the average deviation from diagonality metric with a threshold of two as suggested. For each component, subgroup comparisons were performed, and the p-value and 95% confidence interval were obtained from 500 bootstrap samples. Considering multiple components, pathways, and cell types, a FDR less than 0.05 is considered to be significant.

4.2.3 Pathway collection and organization

The human pathways are well organized into different hierarchies, and eventually, findings from all pathways, which form the basic circuit of cellular functions, could be further organized and visualized to get a more comprehensive understanding of the general functional groups that show abnormalities in gene-gene interactions. We here collected in total 10,402 pathways from GO, with 7658 biological processes; 1006 cellular compartments; 1738 molecular functions. 17 pathway categories are curated by hand, and the total number of pathways in each category is shown.

4.2.4 Single cell clustering and subtype annotation

Processed data was downloaded from AD Knowledge Portal [147]. Under R v3.5.2, function readMM in package Matrix and CreateSeuratObject in Seurat package was used to generate Seurat object. Then, we selected 70633 single cells with more than 200 unique molecular identifiers (UMIs) and mitochondrial content less than 10 percent for further analysis. The original paper provided cell type and sub cell type annotation for all the cells. We directly applied their annotation into our analysis.

4.2.5 Measuring within cell type heterogeneity

Intraclass correlation coefficient (ICC) was first introduced [148] as a modification of Pearson correlation coefficient and widely used to evaluate reflects the variation between 2 or more raters who measure the same group of subjects [149]. Here ICC is utilized to evaluate the eGGC variations among different sub-cell types in each patient group. We calculated eGGC in sub-cell type level for each patient. After this, we get a matrix which columns represent different sub-cell types of one cell type and rows represent individuals. Then ICC was calculated for each patient group which shows if eGGC highly variant among sub-types in the same condition. Higher ICC means the eGGC is more consistent between subtypes.

4.3 Results

4.3.1 Analytical pipeline

The scCovReg pipeline takes as input the single cell gene expression profiles collected from subjects across different pathological or demographic conditions, as well as a set of curated biological pathways whose correlation pattern changes across conditions are to be studied. The core algorithm of scCovReg is covariance regression [145], which was originally designed to study the variations in brain functional connectivity using functional imaging data. Application of covariance regression method on single cell expression data for detecting gene-gene correlation changes is innovative. For a pathway or a set of selected genes, scCovReg makes statistical inferences on how well the genes' correlation pattern for cells of the same type could be explained by subject-level covariates of interest. Considering the existence of sex discrepancy in AD pathology, throughout the analysis, we included sex as a predictor when comparing samples from AD-pathology versus non-pathology groups. Overall, our covariates of interest are disease status, sex, and their interaction.

As shown in Figure 4.1, for each cell type and each curated pathway, correlation matrices among genes are first constructed for each subject from normalized single cell expression data (I, II). The covariates explainable correlation network is then solved by covariance regression, which maximizes a likelihood function connecting the variance of the weighted expression of genes in the pathway, and the covariates, under certain distribution and regularity assumptions (III). Importantly, the weights of the genes indicate the contribution of each gene: a higher weight means that the corresponding gene tends to dominate the changes of the correlation network, i.e., its correlation with other genes are more important in determining the variations of eGGC with respect to the independent variables. Such a process is applied to all the curated pathways and all the cell types, using single cell data collected from subjects of AD and healthy control conditions for both female and male subjects (IV). In total, we collected 10,402 pathways from GO and analyzed 5 cell types including neuron, astrocytes, oligodendrocyte, oligodendrocyte progenitor cell (OPC), and microglia collected by the ROS/MAP cohort. We ran the scCovReg pipeline for each cell type separately, considering the intrinsically different capture or dropout rate of each cell type [64, 150, 151].

Output of the analytical pipeline is the statistical inferences for sex and disease status in explaining the genes' correlation network variations in all collected pathways across all different cell types. Specifically, for each pathway and cell type, we obtain how significant the independent predictors including sex, disease status and their interaction, could explain the variations of the gene correlation matrices calculated on the single cells. For each varied pathway, the important genes will be identified as those with larger weights, and the direction of association between predictors and pathway second-order changes could be revealed as the signs associated with each predictor. The significance of explainability for the predictors is evaluated through a nonparametric bootstrap procedure. A pathway that could be significantly explained is called a differentially correlated pathway (DCP). Similar to multiple linear regression, contrast analysis for each pair of disease-sex condition could be also obtained, namely diseased vs healthy condition in both female and male subjects respectively.

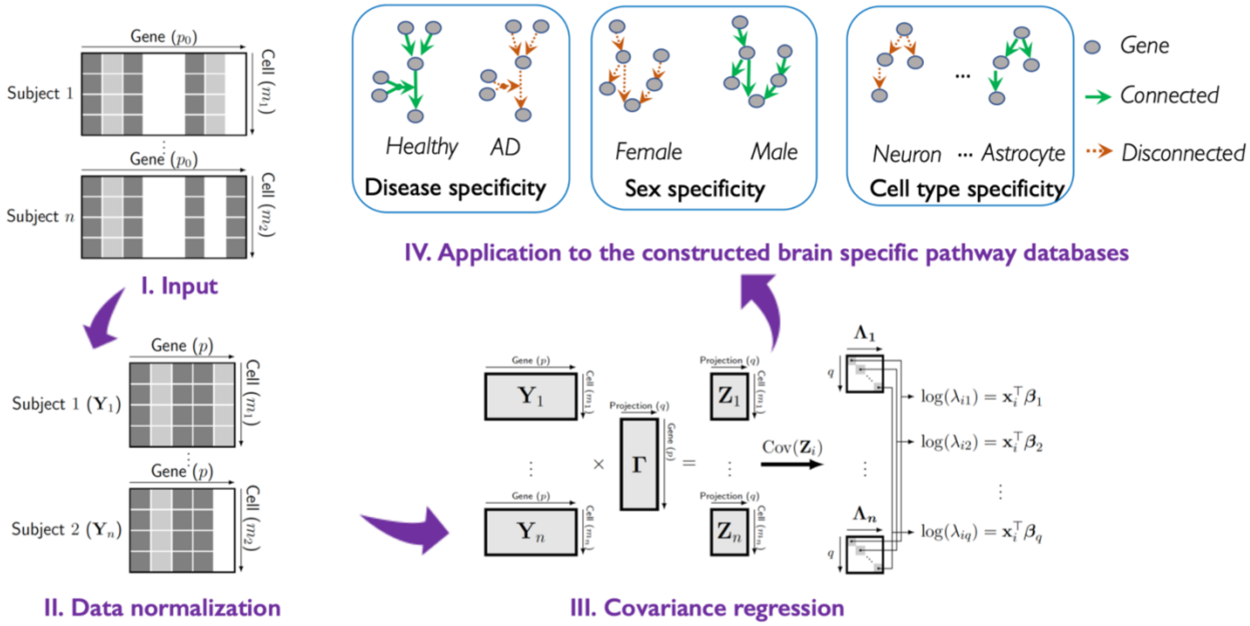


Figure 4.1. scCovReg analytical pipeline to study disease- and sex-specific abnormalities in gene-gene correlation network for curated pathways in cell-type-specific manner.

In total, we identified 3,513 unique GO pathways whose eGGC are significantly different between AD and healthy subjects for both genders respectively, in at least one of the five major cell types (Fig. 4.2 A-B).

4.3.2 Overall trend of pathway connectivity and difference with enrichment-based analysis

To better look for the sex specificity in AD, we did three contrast analyses: AD versus healthy in female or male, and female versus male in AD patients, based on the covariance regression analysis results. In Figure 4.2A, we present the overall trend of the pathway eGGC in each contrast, and compared how in general AD and sex modify the pathway eGGC in the five cell types. Here each dot on the x-axis represents one pathway, and the y-axis shows the covariance regression coefficients for AD vs. control in the females (top row), and males (middle row), and the coefficients for male vs. female in AD patients (bottom row). Note that covariance regression strives to identify all components of the feature correlation network that are explainable by the predictors, where each component may have a different set of weight parameters for the features, and regression coefficients for the predictors. In other words, for each pathway, we may obtain more than one sub-networks that are explainable by the predictors, and the level of explain-ability and direction of association may differ for the different sub-networks. Again, the different sub-networks could be characterized by the different weight vectors for the genes in the pathway. The exact numbers of DCPs, and DCPs with positive and negative coefficients under different contrasts are shown in Table 4.1. Considering the large number of pathways being studied, we control the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses [152].

As shown in Figure 4.2A, for female (top row), diseased astrocyte, microglia, neuron, and OPC all showed a strong trend of pathway “decoupling” with weaker eGGC strength in AD vs healthy patients, and such decoupling is particularly true for astrocyte, microglia, neuron; while on the contrary oligodendrocyte tend to have pathways better coupled in AD condition vs healthy condition. There is no obvious trend for OPC. If we look at the DCPs only, we see the same trend. For neuron, 1,810 of the 1,939 DCP were less correlated in AD vs healthy, while only 322 of the 1,939 DCP are more correlated in AD. Note that when a pathway has two components with opposite signs of associations with the predictor, we may count the pathway twice, one towards the positive count, and another towards the negative count. Similarly, for astrocyte, majority (1,739) of the 1,848 DCPs are less connected in AD with weaker eGGC strength, and majority (127) of the 132 DCPs in OPC are less coupled in AD. On the contrary, 258 of the 315 DCPs in oligodendrocyte are positively associated with AD, meaning better coupling of the genes in these pathways under diseased condition. For microglia, we see only a very small fraction of the pathways showing differential correlation patterns between AD and disease. We suspect that the difficulty in detecting difference in microglia is due to large amount of zero expression genes.

For male (middle row), the general trend of pathway correlation in AD vs healthy conditions seem to hold the same as for female, except for microglia where better coupled pathways seem to dominate in AD. When looking at the DCPs only, the total number of DCPs in each cell type is much less in male than in female, and the numbers are negligible for astrocyte and microglia. In neuron, pathways tend to be less connected with 64 DCPs less connected, given the overall negative coefficients for neuron cells in this contrast group. This is in contrast to other cell types, where oligodendrocyte has 302 pathways much stronger connected. When we compare female to male AD patients (bottom row), clearly, there is a strong sex discrepancy for different cell types, particularly for astrocyte, neuron cells and oligodendrocyte, as we see the coefficients tend to be away from zero. Notably, for OPCs cell types, eGGC varied between male and female AD patients. This is consistent with the current findings on the role of sex in AD pathology [142, 143]. Overall, the numbers of DCPs for non-neuronal populations were substantially smaller, probably owing to reduced power in lower-abundance cell types. These contrasting observations on the number and dominant directionality of DCPs reveal a heterogeneous response to AD pathology between cell types and sex groups—a recurrent theme that we observed throughout the study. These indicate that all major cell types are affected at the transcriptional level by AD pathology and sex, and that single-cell-level resolution is critical because changes in gene expression—including directionality—can be conditional on cell type and sex. We also observed that in healthy condition, female tend to have weaker connection compared with males in neuron, oligodendrocyte and OPC cells. Notably, compared with traditional pathway-enrichment based analysis that aim to detect to pathways enriched by genes with significant first-order changes, our scCovReg pipeline detected a lot more pathways that show second-order changes, as shown in Figure 4.2B. This indicates that covariance regression is more statistically powerful in detecting abnormal pathways with variations on the level of gene-gene correlation; however, the detected DCPs may not necessarily have changes on the level mean expressions.

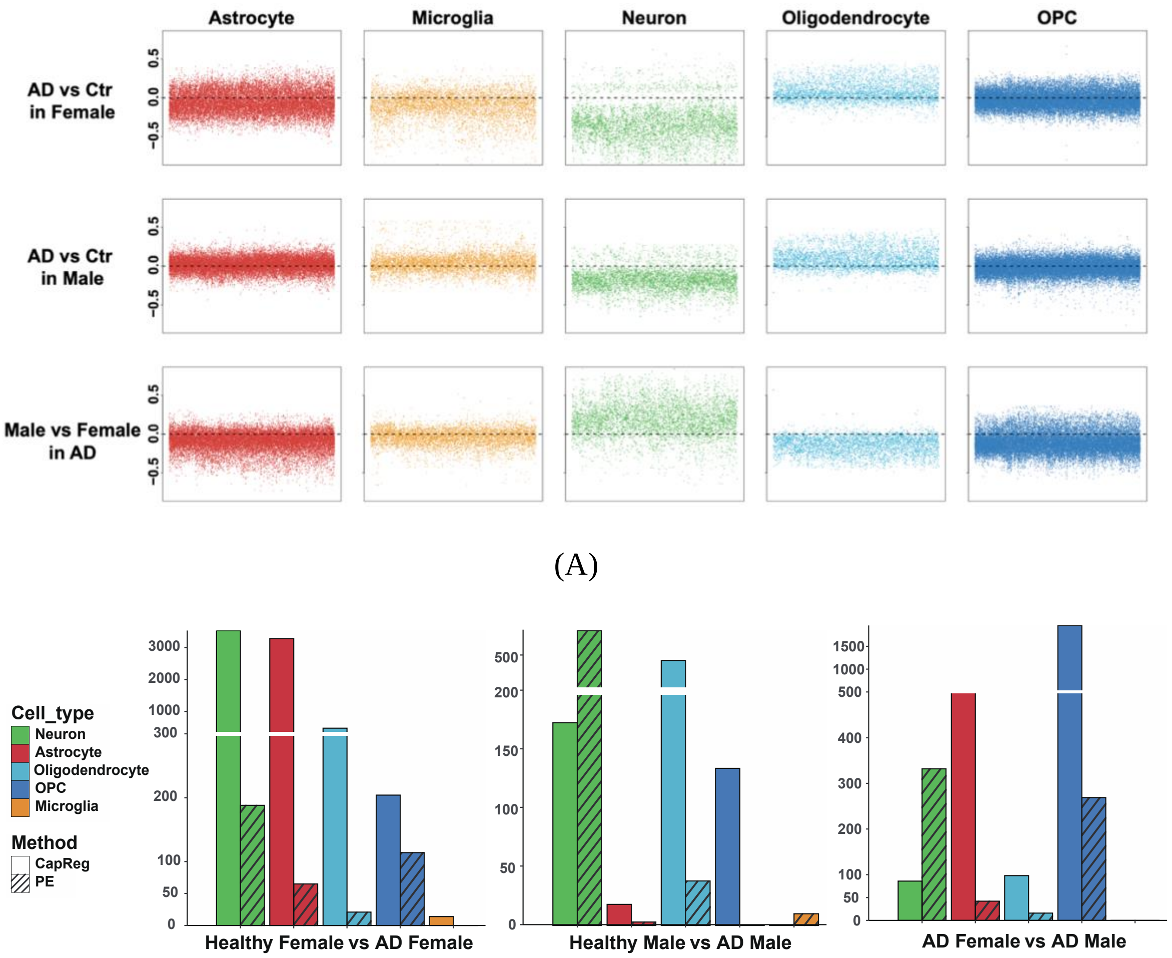
	Pathway with Result	AD-Healthy Female	AD-Healthy Male	Male-Female Healthy	Male-Female AD
Astrocyte	5218	440	5	59	6
Microglia	2772	0	0	0	0
Neuron	4258	480	71	22	38
Oligodendrocyte	2651	392	459	6	1
OPC	6223	11	14	80	50

(A)

	Pathway with Result	AD-Healthy Female	AD-Healthy Male	Male-Female Healthy	Male-Female AD
Astrocyte	2888	1739	10	2090	279
Microglia	1510	11	0	3	0
Neuron	2309	1810	64	30	32
Oligodendrocyte	1604	82	2	587	77
OPC	3361	127	51	2226	1135

(B)

Table 4.1. Number of (A) positive and (B) negative DCPs in each cell type under each contrast.



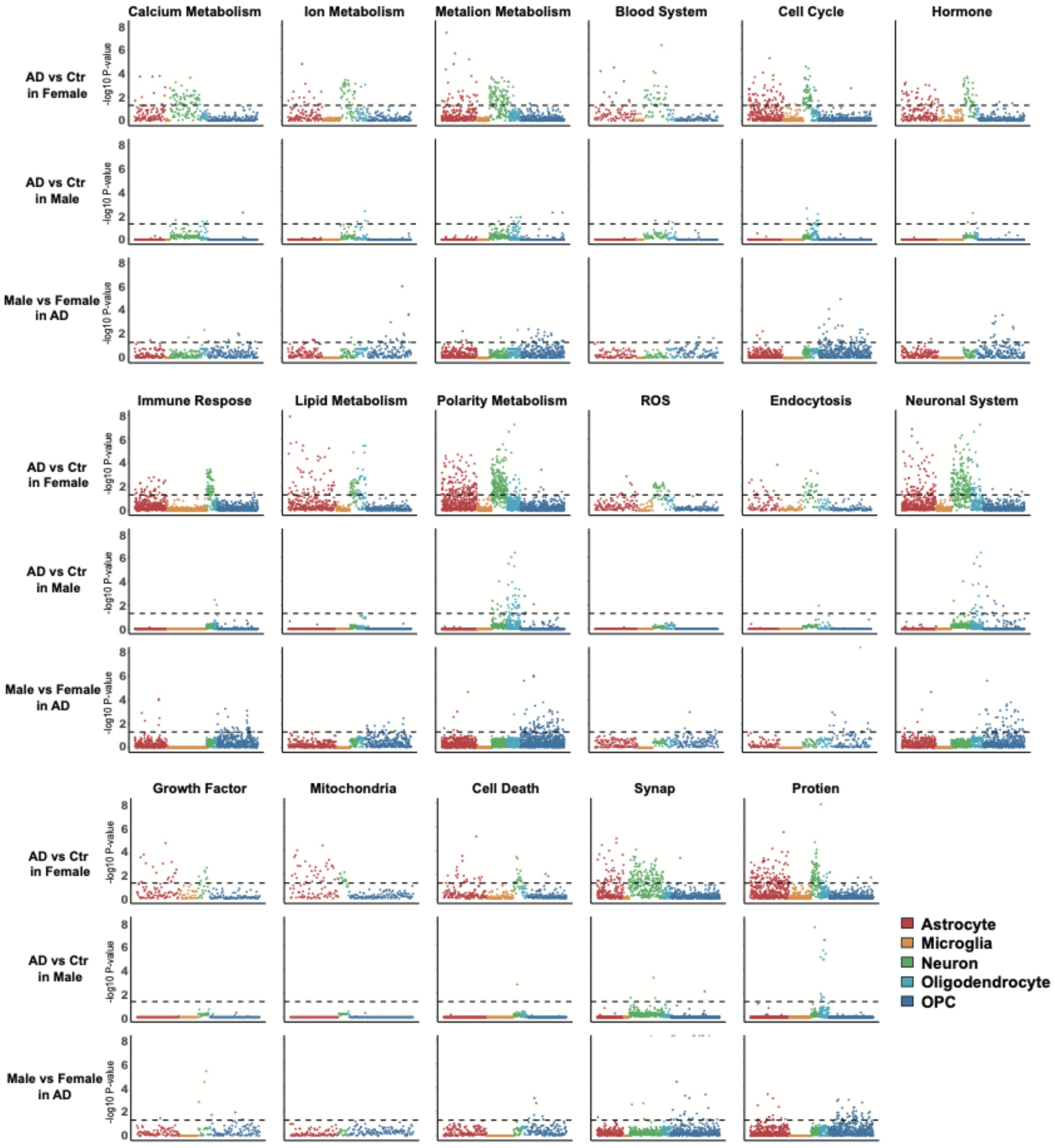
(A)

(B)

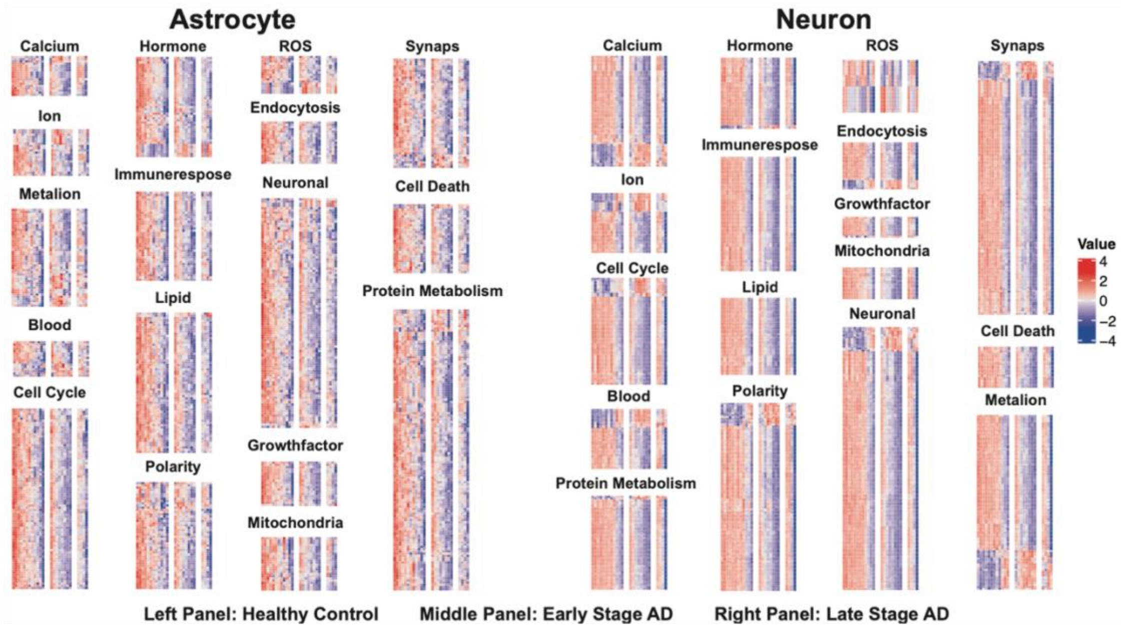
Figure 4.2. (A) Overall trend. (B) Number of significant pathways in CapReg and PE

4.3.3 The patterns of pathway connectivity for known pathway categories

While Figure 4.2 presented an overall trend of the pathway connectivity, in order to get a more mechanistic understanding of the AD based on the analysis results, we further look into the detailed pathway level alterations. We organized pathways in large and distinct categories, and examined the impact of AD on the pathway connectivity in a sex- and cell type-specific manner. The pathways are organized based on the original structures of the pathway database. For example, a pathway category "lipid metabolism" in GO contains 314 pathways that are related to cell lipid metabolism functionals. In total, 17 pathway categories have been summarized, which includes neuronal system, the immune system, general metabolism and lipid metabolism, oxidative stress [153], cell polarity and stress response [154], mitochondrial activity [153, 155], blood–brain barrier dysfunction [156], endocytosis [157], energy metabolism [158], lipid metabolism [159], ion transport and metal ion metabolism [160, 161], calcium regulation [162], hormomne regulation [163], protein homeostasis [164], cell cycle [165, 166], cell death [167], growth factor [166], glia, and neural plasticity and synaptic functions [168]. These pathway categories are selected based on the criterion that 1) they were reported to be related to AD in a broad sense; 2) the number of pathways in the category exceeds a certain number. For each pathway category, we then investigated how the cell type, sex, and disease status might affect the member pathways' connectivity.


Calcium Metabolism
Ion Metabolism
Metalion Metabolism
Blood System
Cell Cycle
Hormone
Immune Respose
Lipid Metabolism
Polarity Metabolism
ROS
Endocytosis
Neuronal System
Growth Factor
Mitochondria
Cell Death
Synap
Protien
AD vs Ctr in Female
AD vs Ctr in Male
Male vs Female in AD
-log10 P-value
0
2
4
6
8
Astrocyte
Microglia
Neuron
Oligodendrocyte
OPC

(A)



(B)

Figure 4.3. (A) AD specific pathways; (B) heatmap of pathway connectivity score compared across conditions.

The pathway connectivity pattern is summarized in Figure 4.3. As shown in Figure 4.3A, each column panel represents one pathway category, and each dot represents one pathway in the category, with different colors for different cell types; and the y-axis shows the negative (FDR adjusted) log p-value of the covariance regression for comparing: AD vs control in females (top row); AD vs control in males (middle row); female vs male in AD patients (bottom row). Larger values on the y-axis mean more significant pathway-level difference in the contrast group, and the dot horizontal line corresponds to adjusted p-value of 10E-2. Among the three contrast groups, the female AD and female control groups are shown to have the largest difference in different pathway categories, as the number of significant DCPs are the largest in this contrast group. In comparing female AD and female controls, astrocyte and neuron cells have the largest number of DCPs, and this is especially true for categories metal ion transport (column #3), lipid metabolism (column #8), cell polarity (column #9), neuronal activity (column #12), and protein metabolism (column #16). Interestingly, the neuron cells seem to have a relatively large number of DCPs in the calcium/ion/metal ion homeostasis, blood circulation, cell cycle response, hormone regulation, immune response category in female AD vs control contrast, and astrocyte showed relatively large number of DCPs in categories including metal ion homeostasis, cell cycle, hormone regulation, growth factor and mitochondria dysfunction. In comparing the male AD and male controls, the significant DCPs is mainly reflected by oligodendrocyte cells, and then neuron cells in cell polarity (column #9), neuronal activity (column #12). In comparing female AD and male AD, the significant DCPs is mainly reflected by OPC cells in cell polarity (column #9), neuronal activity (column #12). For microglia cells, probably due to the low number of cell count, very few significant DCPs were detected throughout all contrasts.

While it is believed that programmed cell death occurs to neuron cells because of the abnormalities in amyloid-β (Aβ) production and clearance [169], we indeed see that a small number of DCPs appear in both astrocyte and neuron cells.

In Figure 4.3B, we further had a more detailed and visual examination of the detailed pathways in each individual patient sample, and the samples (rows) are ordered by their disease stage. It seems that genes are less connected at late stage of AD in most pathways.

These analyses clearly demonstrated the capability of using covariance regression to tease out the pathway-level abnormality in a sex, disease, and cell type specific manner. In summary, AD and sex specific variations in pathway connectivity were observed in certain cell types and biological systems. We also presented the total number of significant pathways in each cell type for each pathway category in Table 4.2. Considering the false discovery rate with multiple tests, we used a p-value cutoff of 10E-5. We also summarized the number of significant pathways in Table 4.2.

	Total Pathways	Astrocyte	Microglia	Neuron	Oligodendrocyte	OPC
Calcium	70	9/20 (45.0%)	0/3 (0.0%)	26/30 (86.7%)	0/13 (0.0%)	0/32 (0.0%)
Ion	76	13/21 (61.9%)	0/11 (0.0%)	18/22 (81.8%)	2/18 (11.1%)	0/26 (0.0%)
Metalion	150	26/48 (54.2%)	0/16 (0.0%)	44/51 (86.3%)	3/32 (9.4%)	2/57 (3.5%)
Blood	42	8/16 (50.0%)	0/3 (0.0%)	14/14 (100.0%)	0/5 (0.0%)	0/16 (0.0%)
Cellcycle	180	33/50 (66.0%)	2/28 (7.1%)	34/39 (87.2%)	6/31 (19.4%)	1/74 (1.4%)
Hormone	165	22/31 (71.0%)	0/21 (0.0%)	24/28 (85.7%)	2/11 (18.2%)	3/39 (7.7%)
Immunerespose	451	26/69 (37.7%)	0/84 (0.0%)	49/51 (96.1%)	0/23 (0.0%)	2/86 (2.3%)
Lipid	314	38/56 (67.9%)	0/16 (0.0%)	27/32 (84.4%)	18/30 (60.0%)	0/53 (0.0%)
Polarity	290	86/109 (78.9%)	0/45 (0.0%)	88/97 (90.7%)	17/76 (22.4%)	6/136 (4.4%)
Ros	75	8/23 (34.8%)	0/8 (0.0%)	19/23 (82.6%)	2/11 (18.2%)	1/23 (4.3%)
Endocytosis	50	11/14 (78.6%)	0/10 (0.0%)	12/14 (85.7%)	0/12 (0.0%)	0/18 (0.0%)
Neuronal	216	53/78 (67.9%)	0/35 (0.0%)	69/75 (92.0%)	16/50 (32.0%)	3/94 (3.2%)
Growthfactor	41	12/17 (70.6%)	0/7 (0.0%)	6/12 (50.0%)	0/2 (0.0%)	0/19 (0.0%)
Mitochondria	85	16/17 (94.1%)	0/0 (NA	11/14 (78.6%)	0/2 (0.0%)	0/22 (0.0%)
Synap	151	28/38 (73.7%)	0/8 (0.0%)	63/82 (76.8%)	1/25 (4.0%)	1/70 (1.4%)
Death	144	15/29 (51.7%)	0/18 (0.0%)	14/17 (82.4%)	0/11 (0.0%)	2/26 (7.7%)
Protein	249	64/79 (81.0%)	0/44 (0.0%)	37/45 (82.2%)	6/45 (13.3%)	3/89 (3.4%)

Table 4.2. Significant/number of pathways with results in AD-Healthy Female contrast

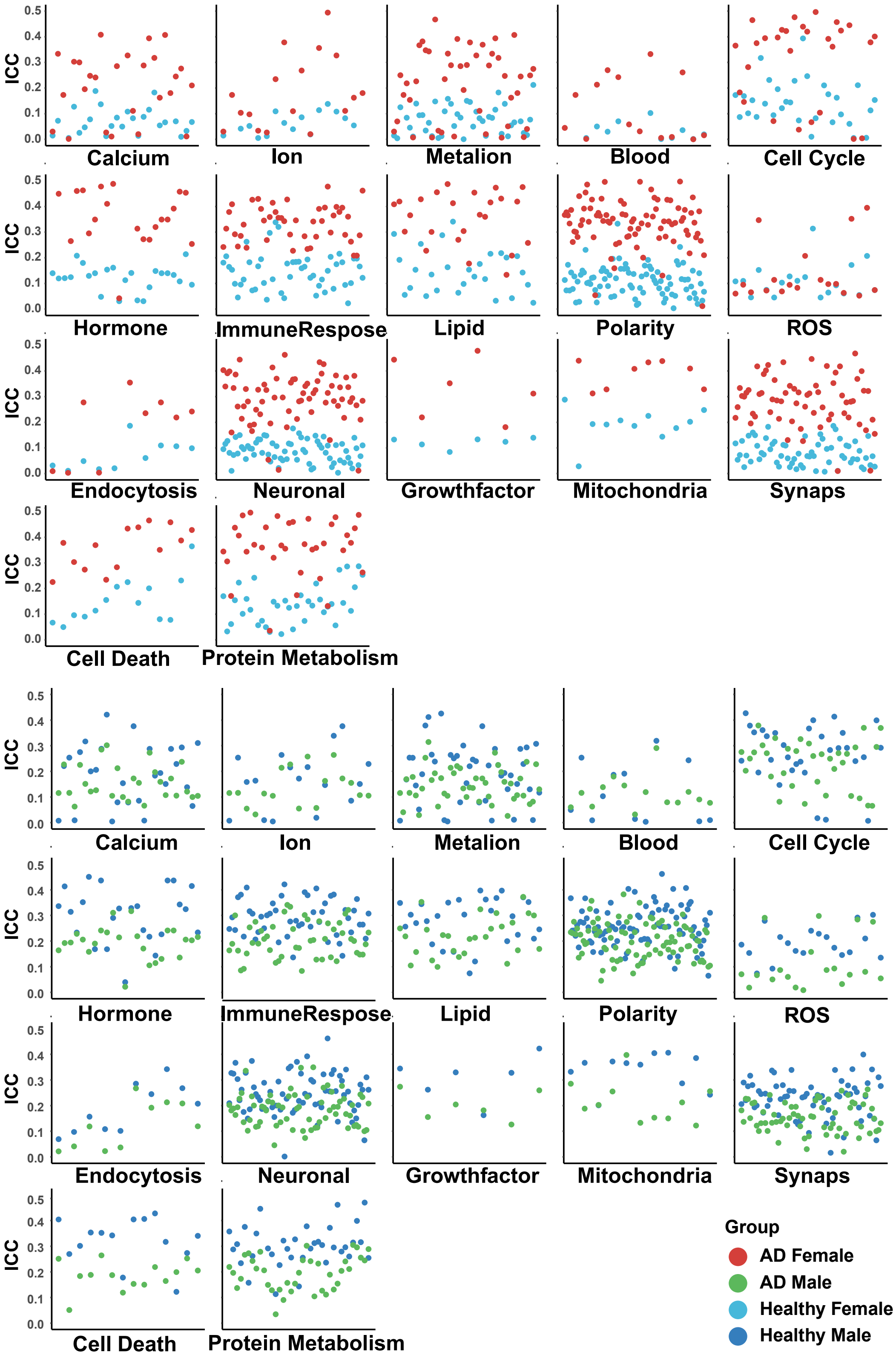


Figure 4.4. Cell subtype heterogeneity analysis

4.3.4 Within cell type heterogeneity of pathway connectivity

To dissect cell-type heterogeneity, we next investigated the variations of pathways with the same cell types, and on a sub-cluster level. The sub-clusters are directly retrieved from the ROSMAP paper [147], which includes 13 excitatory-neuron (Ex), 12 inhibitory-neuron (In), 4 astrocytes (Ast), 5 oligodendrocytes (Oli), 3 oligodendrocyte progenitor cell (Opc), and 4 microglia (Mic) sub-clusters. According to [147],the identified subpopulations were not exclusively enriched with cells from any single individual. We define a “purity” score that is calculated as the of the intra-cluster correlation (ICC) of the averaged connectivity score of different cell subtypes (methods). For one cell type, the “purity” score measures the variations of the connectivity scores for each pathway among different sub-clusters of a cell type.

From Figure 4.4A, we could see that in general, the different subtypes of neuron cells are more homogeneous in female AD than in female control cells, as neuron subtypes in female AD patients (red) usually have consistently higher purity scores than neuron subtypes in female healthy patients (sky blue). However, this is not the case for male subjects. As shown in Figure 4B, the purity scores in male AD and male healthy subjects are not always separable in different pathway categories.

4.3.5 Compare findings with ROSMAP paper

Original paper finds that in comparison of AD late-pathology and no-pathology patients, upregulated genes are involved in protein folding, including molecular chaperones, and are also associated with autophagy, apoptosis, and the generalized stress response [147]. To test the consistency of scCovReg results compared to the original paper, we tested applied our scCovReg method to pathways associated with upregulated genes.

Autophagy and Apoptotic signaling related pathways tend to be less connected in astrocyte and neuron cells.

4.3.6 Validation on independent datasets

In order to validate our covariance regression based analytical pipeline, we applied this pipeline to a similar but smaller independent dataset (GSE157827, see Data access and processing in Methods). After similar preprocessing step as for ROSMAP dataset, we applied scCovReg on three cell types, including astrocyte, neuron and oligodendrocyte. The microglia cell was not included in the analysis because the less sufficient of cell numbers. As shown in Figure 4.5, the x-axis shows pathway level connectivity (covariance regression coefficients) estimated in ROSMAP data, and the y-axix shows coefficients in GSE157827. The solid blue line shows the regression line of pathway level connectivity from two datasets, and the dotted red line shows the y=x line. For all three cell types, the estimations given by ROSMAP and GSE157827 are highly consistent, as their linear regression coefficients are all highly significant (<10E-5). In particular, the R squared for neuron cell is as high as 0.521. For astrocyte and oligodendrocyte, the R square values are not as high. We suspect this is because some patients in GSE157827 don't have sufficient cell numbers for these two cell types. These demonstrate the high robustness of scCovReg.

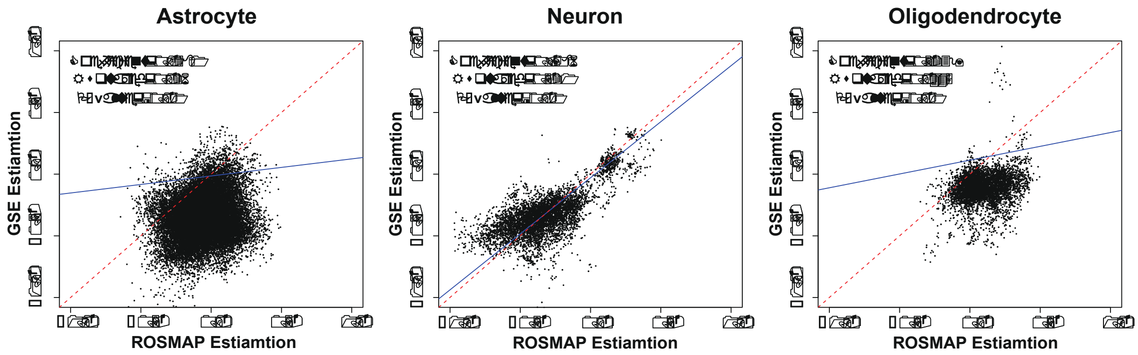


Figure 4.5. Validation of scCovReg on independent dataset. x-axis shows pathway level connectivity (covariance regression coefficients) estimated in ROSMAP data, and y-axis shows coefficients in GSE157827. A linear regression was conducted by regressing the GSE157827 estimate on the ROSMAP estimate, and the regression coefficient, R square and p-value are shown in the plot for each cell type, and the regression line (blue solid line), as well as the y=x line (red dotted line) are also shown.

4.4 Discussion

Many currently untreatable diseases, including AD, arise due to variations in, and through a combination of, multiple modulators of genetic, epigenetic, and environmental nature. Unfortunately, how such modulators lead to a specific disease phenotype or inflict a vulnerability to some cells and tissues but not others remain largely unknown and unsatisfactorily addressed. The analysis of cell-specific gene-gene interaction networks may shed light on organization of biological systems and subsequently to disease vulnerabilities. The emergence of single cell technology is promising to detect cell-type specific changes, however, deriving the variabilities of gene interaction networks across different cell, phenotypes and disease contexts remains a challenge. Currently, a two-step approach is usually adopted to identify disease pathways using single-cell RNA-Seq (scRNA-Seq) data in a cell type-specific manner: (1) unsupervised clustering of single cells to delineate the cell type identities or states [170-173] followed by differential gene expression analysis [174-178] to identify cell type and disease-specific genes, and (2) pathway enrichment and co-expression module detection methods using either the differentially expressed genes or genes' importance ranking [36, 37]. Apparently, it fails to detect the second-order changes of the pathways.

The challenges of systematically investigating the variations in gene “interactome” lies in the following aspects: 1) the scale of interactions of genes. 2) The variations of biological processes are associated with AD pathology in a cell type and sex specific manner. De novo construction of biological networks will need to take the combination of the conditions into consideration, namely, disease status, sex, and cell type. This fine segmentation of the samples may lead to low statistical power. 3) There may exist many biological processes or pathways that underlie the AD progression, hence studying the whole transcriptome-level interactome variations may not pinpoint the specific biological processes or pathways. 4) There is a large number of curated pathways in databases, such as Gene Ontology, that serve as prior information on the structure of the genes' interactome. To this end, we study variations of interactome at the level of existing pathway/biological process, and our scCovReg piepline links gene-gene interaction network variabilities to AD formation in a sex and cell type specific manner, revealing a viable and reproducible experimental solution to obtaining rigorous context-dependent gene-gene interactions.

Our analytical pipeline, scCovReg tool represents the first-of-its-kind to capture the variations of the interactions among genes in a pathway, and the detected pathway-level disease abnormalities are more robust for knowledge transfer from one study or platform to another, as it allows for different activation forms of gene combinations. scCovReg harnesses the power of scRNA-Seq data to discover AD-associated molecules and pathways in a cell type, disease and sex specific manner, and to articulate the disease mechanisms on a cellular, molecular and physiological level, and laying the foundation to develop prevention/intervention strategies. Current single gene or enrichment-based pathway analysis tends to overlook pathways with abnormal gene-gene interactions, due to low statistical power and lack of a sufficient model to detect such interactions. In AD research, this is further challenged that the abnormalities in AD often occur in a cell type and sex specific manner. While single-cell technology dissects the cell type level variations, the sample size is still too low due to high sequencing costs. We addressed the challenges by directly modeling the pathway-level variations and make inference on key variables, such as disease status and sex, using covariance regression model. Several challenges exist that warrants further research into this direction as discussed below.

Some limitations of the work exist. In the study, a large number of pathways and several cell types are considered, and the large number of performed tests make the statistical power a big issue. In the future, we believe constructing a brain microenvironment specific pathway database will be helpful in boosting up the statistical power, and it will be highly beneficial to the scientific community. The original covariance regression approach was proposed based on the assumption that the data are normally distributed. Applying to scRNA-Seq data, with appropriate data transformation, the normality assumption is assumed to hold. Another issue in scRNA-Seq is the existence of missing data. The current strategy is to include a screening step to remove genes with missing data in over 25% of the subjects. After this step, a large number of genes are removed from the analysis. An alternative is to propose a distribution-free approach and at the same time to consider the fact of zero inflation. The covariance regression model was applied on single cells of each cell type separately. However, even for the same cell type, there might exist several sub-cell types. A solution is to examine the heterogeneity of the pathway connectivity score, and determine whether there is a need to break the cell type into various subtypes.

Chapter 5 Conclusions

In this research, we aim to reveal the biological meaningful subspace structures in omics data. Specifically, we focus on three types of subspace structures, bi-cluster low-rank subspace, sparse subspace and covariates explainable subspace.

For bi-cluster low-rank subspace, we try to explain cell type specific expression, which bulk RNA-seq data is mixture signal of different cell types. Due to experimental limitation, we applied a semi-supervised model to detect cell type specific low-rank structures and predict their relative proportions across different samples. Our prediction can help biologist to better understand the cell component and further investigate the disease mechanism.

Next, we shift our focus to identify disease-driven sparse subspace. We proposed a novel statistical model PLUS that could identify cancer metastasis related genes and under detected cases. In TCGA data, PLUS predicted metastasis cases have worse progression free survival. Besides, our method is with better performance compared with other methods in simulated data. Moreover, the identified metastasis related genes are robust in independence scRNA-seq datasets.

Lastly, to discover the covariates explainable subspace in covariance matrix, we referenced a covariance regression approach, namely, scCovReg. We utilized scCovReg to model the pathway level second-order variations using scRNA-Seq data and to associate the second-order variations with important subject-level characteristics, such as disease status. Our finding provides a unique angle to study gene connection abnormalities in Alzheimer disease.

References

1. Buermans, H. and J. Den Dunnen, Next generation sequencing technology: advances and applications. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 2014. 1842(10): p. 1932-1941.
2. Udell, M. and A. Townsend, Why are big data matrices approximately low rank? SIAM Journal on Mathematics of Data Science, 2019. 1(1): p. 144-160.
3. Agarwal, M. and R. Mehra, Review of matrix decomposition techniques for signal processing applications. Int. Journal of Engineering Research and Applications, 2014. 4(1): p. 90-93.
4. Hackl, H., et al., Computational genomics tools for dissecting tumour–immune cell interactions. Nature Reviews Genetics, 2016. 17(8): p. 441.
5. Li, B., et al., Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome biology, 2016. 17(1): p. 174.
6. Chen, Z., et al., Inference of immune cell composition on the expression profiles of mouse tissue. Scientific reports, 2017. 7: p. 40508.
7. Newman, A.M., et al., Robust enumeration of cell subsets from tissue expression profiles. Nature methods, 2015. 12(5): p. 453.
8. Wang, X., et al., Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. 2019. 10(1): p. 380.
9. Racle, J., et al., Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife, 2017. 6.
10. Newman, A.M., et al., Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol, 2019. 37(7): p. 773-782.
11. Li, B., et al., Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol, 2016. 17(1): p. 174.
12. Gaujoux, R. and C.J.B. Seoighe, CellMix: a comprehensive toolbox for gene expression deconvolution. 2013. 29(17): p. 2211-2212.
13. Frishberg, A., et al., Cell composition analysis of bulk genomics using single-cell data. Nat Methods, 2019. 16(4): p. 327-332.
14. Finotello, F. and Z.J.C.I. Trajanoski, Immunotherapy, Quantifying tumor-infiltrating immune cells from transcriptomics data. 2018. 67(7): p. 1031-1040.
15. Abbas, A.R., et al., Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. 2009. 4(7): p. e6098.
16. Abbas, A., et al., Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. 2005. 6(4): p. 319.
17. Li, T., et al., TIMER2. 0 for analysis of tumor-infiltrating immune cells. Nucleic acids research, 2020. 48(W1): p. W509-W514.
18. Aran, D., Z. Hu, and A.J. Butte, xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome biology, 2017. 18(1): p. 1-14.
19. Lu, X., et al., SSMD: a semi-supervised approach for a robust cell type identification and deconvolution of mouse transcriptomics data. Briefings in bioinformatics, 2021. 22(4): p. bbaa307.
20. Wright, J., et al., Sparse representation for computer vision and pattern recognition. Proceedings of the IEEE, 2010. 98(6): p. 1031-1044.

21. Gupta, G.P. and J. Massagué, Cancer metastasis: building a framework. Cell, 2006. 127(4): p. 679-695.
22. Steeg, P.S., Tumor metastasis: mechanistic insights and clinical challenges. Nature medicine, 2006. 12(8): p. 895-904.
23. Steeg, P.S., Targeting metastasis. Nature reviews cancer, 2016. 16(4): p. 201-218.
24. Mehlen, P. and A. Puisieux, Metastasis: a question of life or death. Nature reviews cancer, 2006. 6(6): p. 449-458.
25. Bernards, R. and R.A. Weinberg, Metastasis genes: a progression puzzle. Nature, 2002. 418(6900): p. 823-823.
26. Riggi, N., M. Aguet, and I. Stamenkovic, Cancer metastasis: a reappraisal of its underlying mechanisms and their relevance to treatment. Annual Review of Pathology: Mechanisms of Disease, 2018. 13: p. 117-140.
27. Jiao, W., et al., A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nature Communications, 2020. 11(1): p. 728.
28. van't Veer, L.J., et al., Expression profiling predicts outcome in breast cancer. Breast Cancer Research, 2002. 5(1): p. 57.
29. Van't Veer, L.J., et al., Gene expression profiling predicts clinical outcome of breast cancer. nature, 2002. 415(6871): p. 530-536.
30. Chang, H.Y., et al., Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS biology, 2004. 2(2).
31. Huang, E., et al., Gene expression predictors of breast cancer outcomes. The Lancet, 2003. 361(9369): p. 1590-1596.
32. Wang, Y., et al., Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet, 2005. 365(9460): p. 671-679.
33. Tang, X.-R., et al., Development and validation of a gene expression-based signature to predict distant metastasis in locoregionally advanced nasopharyngeal carcinoma: a retrospective, multicentre, cohort study. The Lancet Oncology, 2018. 19(3): p. 382-393.
34. Zhou, L.-Q., et al., Lymph node metastasis prediction from primary breast cancer US images using deep learning. Radiology, 2020. 294(1): p. 19-28.
35. Zhou, J., et al., PLUS: Predicting cancer metastasis potential based on positive and unlabeled learning. PLoS computational biology, 2022. 18(3): p. e1009956.
36. Subramanian, A., et al., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 2005. 102(43): p. 15545-15550.
37. Wang, X.-L. and L. Li, Cell type-specific potential pathogenic genes and functional pathways in Alzheimer’s Disease. BMC Neurology, 2021. 21(1): p. 381.
38. Hoff, P.D., A hierarchical eigenmodel for pooled covariance estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2009. 71(5): p. 971-992.
39. Fox, E.B. and D.B. Dunson, Bayesian nonparametric covariance regression. The Journal of Machine Learning Research, 2015. 16(1): p. 2501-2542.

40. Zou, T., et al., Covariance regression analysis. Journal of the American Statistical Association, 2017. 112(517): p. 266-281.

41. Zhao, Y., et al., Covariate Assisted Principal regression for covariance matrix outcomes. Biostatistics, 2021. 22(3): p. 629-645.

42. Seiler, C. and S. Holmes, Multivariate heteroscedasticity models for functional brain connectivity. Frontiers in neuroscience, 2017. 11: p. 696.

43. Zhao, Y., et al., A whole-brain modeling approach to identify individual and group variations in functional connectivity. Brain and behavior, 2021. 11(1): p. e01942.

44. Zhao, Y., B. Caffo, and X. Luo, Principal regression for high dimensional covariance matrices. Electronic Journal of Statistics, 2021. 15(2): p. 4192-4235.

45. Lu, X., et al., Cell-type specific variations of within pathway interaction in AD using covariance regression. Alzheimer's & Dementia, 2021. 17: p. e057873.

46. Beck, J.A., et al., Genealogies of mouse inbred strains. Nature genetics, 2000. 24(1): p. 23-25.

47. Mund, J.A., et al., Genetic disruption of the small GTPase RAC1 prevents plexiform neurofibroma formation in mice with neurofibromatosis type 1. Journal of Biological Chemistry, 2020: p. jbc. RA119. 010981.

48. Huang, M., et al., Sestrin 3 Protects Against Diet-Induced Nonalcoholic Steatohepatitis in Mice Through Suppression of Transforming Growth Factor β Signal Transduction. Hepatology, 2020. 71(1): p. 76-92.

49. Pandey, R., et al., SHP2 inhibition reduces leukemogenesis in models of combined genetic and epigenetic mutations. Journal of Clinical Investigation, 2019. 129(12): p. 5468-5473.

50. Zhang, C., S. Cao, and Y. Xu, Population dynamics inside cancer biomass driven by repeated hypoxia-reoxygenation cycles. Quantitative Biology, 2014. 2(3): p. 85-99.

51. Marques, S., et al., Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. Science, 2016. 352(6291): p. 1326-1329.

52. La Manno, G., et al., Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. Cell, 2016. 167(2): p. 566-580 e19.

53. Codeluppi, S., et al., Spatial organization of the somatosensory cortex revealed by osmFISH. Nat Methods, 2018. 15(11): p. 932-935.

54. Luo, F., et al., Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC bioinformatics, 2007. 8(1): p. 299.

55. Lopez, R., et al., Deep generative modeling for single-cell transcriptomics. 2018. 15(12): p. 1053.

56. Pepper, S.D., et al., The utility of MAS5 expression summary and detection call algorithms. BMC bioinformatics, 2007. 8(1): p. 273.

57. Johnson, W.E., C. Li, and A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 2007. 8(1): p. 118-127.

58. Regev, A., et al., Science forum: the human cell atlas. Elife, 2017. 6: p. e27041.

59. Han, X., et al., Mapping the mouse cell atlas by microwell-seq. Cell, 2018. 172(5): p. 1091-1107. e17.

60. Stuart, T., et al., Comprehensive Integration of Single-Cell Data. Cell, 2019.
61. Butler, A., et al., Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature biotechnology, 2018. 36(5): p. 411.
62. Church, R.J., et al., A systems biology approach utilizing a mouse diversity panel identifies genetic differences influencing isoniazid-induced microvesicular steatosis. Toxicological Sciences, 2014. 140(2): p. 481-492.
63. Dobin, A., et al., STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013. 29(1): p. 15-21.
64. Wan, C., et al., LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data. Nucleic acids research, 2019. 47(18): p. e111-e111.
65. Zhang, Y., et al., M3S: A comprehensive model selection for multi-modal single-cell RNA sequencing data. BMC bioinformatics, 2019. 20(24): p. 1-5.
66. Maier, M.J., DirichletReg: Dirichlet regression for compositional data in R. 2014.
67. Chang, W., et al., ICTD: A semi-supervised cell type identification and deconvolution method for multi-omics data. bioRxiv, 2019: p. 426593.
68. Wang, X., et al., Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature communications, 2019. 10(1): p. 1-9.
69. Wan, C., et al., Denoising individual bias for a fairer binary submatrix detection. arXiv preprint arXiv:2007.15816, 2020.
70. Wan, C., et al., Fast And Efficient Boolean Matrix Factorization By Geometric Segmentation. arXiv, 2019: p. arXiv: 1909.03991.
71. Chang, W., et al., Supervised clustering of high dimensional data using regularized mixture modeling. arXiv preprint arXiv:2007.09720, 2020.
72. Talmadge, J.E. and I.J. Fidler, AACR centennial series: the biology of cancer metastasis: historical perspective. Cancer research, 2010. 70(14): p. 5649-5669.
73. Valastyan, S. and R.A. Weinberg, Tumor metastasis: molecular insights and evolving paradigms. Cell, 2011. 147(2): p. 275-292.
74. Robinson, D.R., et al., Integrative clinical genomics of metastatic cancer. Nature, 2017. 548(7667): p. 297-303.
75. Kikuchi, T., et al., Expression profiles of non-small cell lung cancers on cDNA microarrays: identification of genes for prediction of lymph-node metastasis and sensitivity to anti-cancer drugs. Oncogene, 2003. 22(14): p. 2192-2205.
76. Schell, M.J., et al., A composite gene expression signature optimizes prediction of colorectal cancer metastasis and outcome. Clinical Cancer Research, 2016. 22(3): p. 734-745.
77. Klein, E.A., et al., Decipher genomic classifier measured on prostate biopsy predicts metastasis risk. Urology, 2016. 90: p. 148-152.
78. Goossens-Beumer, I.J., et al., MicroRNA classifier and nomogram for metastasis prediction in colon cancer. Cancer Epidemiology and Prevention Biomarkers, 2015. 24(1): p. 187-197.
79. Jahid, M.J., T.H. Huang, and J. Ruan, A personalized committee classification approach to improving prediction of breast cancer metastasis. Bioinformatics, 2014. 30(13): p. 1858-1866.
80. Fan, C., et al., Concordance among gene-expression–based predictors for breast cancer. New England Journal of Medicine, 2006. 355(6): p. 560-569.

81. Elkan, C. and K. Noto. Learning classifiers from only positive and unlabeled data. in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008.

82. Ward, G., et al., Presence-only data and the EM algorithm. Biometrics, 2009. 65(2): p. 554-563.

83. Song, H. and G. Raskutti, PULasso: High-dimensional variable selection with presence-only data. Journal of the American Statistical Association, 2019: p. 1-30.

84. Yang, P., et al., AdaSampling for positive-unlabeled and label noise learning with bioinformatics applications. IEEE transactions on cybernetics, 2018. 49(5): p. 1932-1943.

85. Chen, T. and C. Guestrin. Xgboost: A scalable tree boosting system. in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

86. Ho, T.K. Random decision forests. in Proceedings of 3rd international conference on document analysis and recognition. 1995. IEEE.

87. Hoadley, K.A., et al., Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell, 2018. 173(2): p. 291-304. e6.

88. Liu, J., et al., An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell, 2018. 173(2): p. 400-416. e11.

89. Puram, S.V., et al., Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell, 2017. 171(7): p. 1611-1624. e24.

90. Chung, W., et al., Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nature communications, 2017. 8(1): p. 1-12.

91. Kim, N., et al., Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nature communications, 2020. 11(1): p. 1-15.

92. Stuart, T., et al., Comprehensive integration of single-cell data. Cell, 2019. 177(7): p. 1888-1902. e21.

93. National Collaborating Centre for, C., National Institute for Health and Clinical Excellence: Guidance, in Diagnosis and Management of Metastatic Malignant Disease of Unknown Primary Origin. 2010, National Collaborating Centre for Cancer (UK)

94. Tibshirani, R., Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 1996. 58(1): p. 267-288.

95. Andersen, P.K. and R.D. Gill, Cox's regression model for counting processes: a large sample study. The annals of statistics, 1982: p. 1100-1120.

96. Holm, S., A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, 1979: p. 65-70.

97. Consortium, G.O., Gene ontology consortium: going forward. Nucleic acids research, 2015. 43(D1): p. D1049-D1056.

98. Lisanti, M.P., et al., Hydrogen peroxide fuels aging, inflammation, cancer metabolism and metastasis: the seed and soil also needs "fertilizer". Cell cycle (Georgetown, Tex.), 2011. 10(15): p. 2440-2449.

99. Pani, G., T. Galeotti, and P. Chiarugi, Metastasis: cancer cell's escape from oxidative stress. Cancer Metastasis Rev, 2010. 29(2): p. 351-78.

100. Prevarskaya, N., R. Skryma, and Y. Shuba, Calcium in tumour metastasis: new roles for known actors. Nature Reviews Cancer, 2011. 11(8): p. 609-618.

101. Yao, M., et al., Chapter Eight - Cytokine Regulation of Metastasis and Tumorigenicity, in Advances in Cancer Research, D.R. Welch and P.B. Fisher, Editors. 2016, Academic Press. p. 265-367.

102. Rousseeuw, P.J., Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987. 20: p. 53-65.

103. Stitzenberg, K.B., et al., Extracapsular extension of the sentinel lymph node metastasis: a predictor of nonsentinel node tumor burden. Annals of surgery, 2003. 237(5): p. 607.

104. Xu, H., C. Caramanis, and S. Mannor, Sparse algorithms are not stable: A no-free-lunch theorem. IEEE transactions on pattern analysis and machine intelligence, 2011. 34(1): p. 187-193.

105. Priestley, P., et al., Pan-cancer whole-genome analyses of metastatic solid tumours. Nature, 2019. 575(7781): p. 210-216.

106. Koplev, S., et al., Integration of pan-cancer transcriptomics with RPPA proteomics reveals mechanisms of epithelial-mesenchymal transition. PLoS computational biology, 2018. 14(1): p. e1005911.

107. Mak, M.P., et al., A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. Clinical Cancer Research, 2016. 22(3): p. 609-620.

108. 2021 Alzheimer's disease facts and figures. Alzheimers Dement, 2021. 17(3): p. 327-406.

109. Twine, N.A., et al., Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. PloS one, 2011. 6(1): p. e16266.

110. van den Hove, D.L.A., et al., Epigenome-wide association studies in Alzheimer’s disease; achievements and challenges. Brain Pathology, 2020. 30(5): p. 978-983.

111. Nativio, R., et al., Dysregulation of the epigenetic landscape of normal aging in Alzheimer’s disease. Nature Neuroscience, 2018. 21(4): p. 497-505.

112. Verheijen, J. and K. Sleegers, Understanding Alzheimer disease at the interface between genetics and transcriptomics. Trends in Genetics, 2018. 34(6): p. 434-447.

113. Gjoneska, E., et al., Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer’s disease. Nature, 2015. 518(7539): p. 365-369.

114. Roubroeks, J.A., et al., Epigenetics and DNA methylomic profiling in Alzheimer's disease and other neurodegenerative diseases. Journal of neurochemistry, 2017. 143(2): p. 158-170.

115. Ernst, J., et al., Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat Biotechnol, 2016. 34(11): p. 1180-1190.

116. Ham, S. and S.-J.V. Lee, Advances in transcriptome analysis of human brain aging. Experimental & Molecular Medicine, 2020. 52(11): p. 1787-1797.

117. Keil, J.M., A. Qalieh, and K.Y. Kwan, Brain Transcriptome Databases: A User&#039;s Guide. The Journal of Neuroscience, 2018. 38(10): p. 2399.

118. Wang, Y., et al., N(6)-methyladenosine RNA modification regulates embryonic neural stem cell self-renewal through histone modifications. Nat Neurosci, 2018. 21(2): p. 195-206.

119. Swartzlander, D.B., et al., Concurrent cell type–specific isolation and profiling of mouse brains in inflammation and Alzheimer's disease. JCI insight, 2018. 3(13).

120. Grubman, A., et al., A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nature neuroscience, 2019. 22(12): p. 2087-2097.

121. Corces, M.R., et al., Single-cell epigenomic identification of inherited risk loci in Alzheimer's and Parkinson's disease. bioRxiv, 2020: p. 2020.01.06.896159.

122. Ingelsson, M., et al., Early Aβ accumulation and progressive synaptic loss, gliosis, and tangle formation in AD brain. Neurology, 2004. 62(6): p. 925-931.

123. Coleman, P.D. and P.J. Yao, Synaptic slaughter in Alzheimer's disease. Neurobiology of aging, 2003. 24(8): p. 1023-1027.

124. Giulian, D., Microglia and the immune pathology of Alzheimer disease. The American Journal of Human Genetics, 1999. 65(1): p. 13-18.

125. Sarlus, H. and M.T. Heneka, Microglia in Alzheimer's disease. The Journal of clinical investigation, 2017. 127(9): p. 3240-3249.

126. Desai, M.K., et al., Early oligodendrocyte/myelin pathology in Alzheimer's disease mice constitutes a novel therapeutic target. The American journal of pathology, 2010. 177(3): p. 1422-1435.

127. Wu, Y., et al., Alterations of myelin morphology and oligodendrocyte development in early stage of Alzheimer's disease mouse model. Neuroscience letters, 2017. 642: p. 102-106.

128. Saab, A.S. and K.-A. Nave, Myelin dynamics: protecting and shaping neuronal functions. Current opinion in neurobiology, 2017. 47: p. 104-112.

129. van der Wijst, M.G.P., et al., Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet, 2018. 50(4): p. 493-497.

130. Hwang, B., J.H. Lee, and D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med, 2018. 50(8): p. 96.

131. Buenrostro, J.D., et al., Single-cell chromatin accessibility reveals principles of regulatory variation. Nature, 2015. 523(7561): p. 486-90.

132. Cao, J., et al., Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science, 2018. 361(6409): p. 1380-1385.

133. Alsema, A.M., et al., Profiling Microglia From Alzheimer's Disease Donors and Non-demented Elderly in Acute Human Postmortem Cortical Tissue. Front Mol Neurosci, 2020. 13: p. 134.

134. Leng, K., et al., Molecular characterization of selectively vulnerable neurons in Alzheimer's disease. Nature Neuroscience, 2021. 24(2): p. 276-287.

135. Otero-Garcia, M., et al., Single-soma transcriptomics of tangle-bearing neurons in Alzheimer's disease reveals the signatures of tau-associated synaptic dysfunction. bioRxiv, 2020: p. 2020.05.11.088591.

136. Grubman, A., et al., A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat Neurosci, 2019. 22(12): p. 2087-2097.
137. Mathys, H., et al., Single-cell transcriptomic analysis of Alzheimer's disease. Nature, 2019. 570(7761): p. 332-337.
138. Zhou, Y., et al., Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer’s disease. Nature Medicine, 2020. 26(1): p. 131-142.
139. Jiang, J., et al., scREAD: A Single-Cell RNA-Seq Database for Alzheimer's Disease. iScience, 2020. 23(11): p. 101769.
140. Gardy, J.L., et al., Enabling a systems biology approach to immunology: focus on innate immunity. Trends in immunology, 2009. 30(6): p. 249-262.
141. Barabási, A.-L., N. Gulbahce, and J. Loscalzo, Network medicine: a network-based approach to human disease. Nature reviews genetics, 2011. 12(1): p. 56-68.
142. Barnes, L.L., et al., Sex differences in the clinical manifestations of Alzheimer disease pathology. Archives of general psychiatry, 2005. 62(6): p. 685-691.
143. Ferretti, M.T., et al., Sex differences in Alzheimer disease—the gateway to precision medicine. Nature Reviews Neurology, 2018. 14(8): p. 457-469.
144. Cline, M.S., et al., Integration of biological networks and gene expression data using Cytoscape. Nature protocols, 2007. 2(10): p. 2366-2382.
145. Zhao, Y., et al., Covariate assisted principal regression for covariance matrix outcomes. Biostatistics, 2019.
146. Ashburner, M., et al., Gene ontology: tool for the unification of biology. Nature genetics, 2000. 25(1): p. 25-29.
147. Mathys, H., et al., Single-cell transcriptomic analysis of Alzheimer’s disease. Nature, 2019. 570(7761): p. 332-337.
148. Fisher, R.A., Statistical methods for research workers, in Breakthroughs in statistics. 1992, Springer. p. 66-70.
149. Koo, T.K. and M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of chiropractic medicine, 2016. 15(2): p. 155-163.
150. Kharchenko, P.V., L. Silberstein, and D.T. Scadden, Bayesian approach to single-cell differential expression analysis. Nature methods, 2014. 11(7): p. 740-742.
151. Qiu, P., Embracing the dropouts in single-cell RNA-seq analysis. Nature communications, 2020. 11(1): p. 1-9.
152. Benjamini, Y. and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 1995. 57(1): p. 289-300.
153. Manoharan, S., et al., The Role of Reactive Oxygen Species in the Pathogenesis of Alzheimer's Disease, Parkinson's Disease, and Huntington's Disease: A Mini Review. Oxidative medicine and cellular longevity, 2016. 2016: p. 8590578-8590578.
154. Feng, B., et al., Planar cell polarity signaling components are a direct target of β-amyloid–associated degeneration of glutamatergic synapses. Science Advances. 7(34): p. eabh2307.

155. Cenini, G. and W. Voos, Mitochondria as potential targets in Alzheimer disease therapy: an update. Frontiers in pharmacology, 2019. 10: p. 902.

156. Sweeney, M.D., A.P. Sagare, and B.V. Zlokovic, Blood–brain barrier breakdown in Alzheimer disease and other neurodegenerative disorders. Nature Reviews Neurology, 2018. 14(3): p. 133-150.

157. Burrinha, T., et al., Upregulation of APP endocytosis by neuronal aging drives amyloid-dependent synapse loss. J Cell Sci, 2021. 134(9).

158. Butterfield, D.A. and B. Halliwell, Oxidative stress, dysfunctional glucose metabolism and Alzheimer disease. Nature Reviews Neuroscience, 2019. 20(3): p. 148-160.

159. Barbash, S., et al., Alzheimer's brains show inter-related changes in RNA and lipid metabolism. Neurobiology of disease, 2017. 106: p. 1-13.

160. Atwood, C.S., et al., Role of free radicals and metal ions in the pathogenesis of Alzheimer’s disease. Metal ions in biological systems, 2018: p. 309-364.

161. Wang, L., et al., Current understanding of metal ions in the pathogenesis of Alzheimer’s disease. Translational Neurodegeneration, 2020. 9(1): p. 10.

162. Pchitskaya, E., E. Popugaeva, and I. Bezprozvanny, Calcium signaling and molecular mechanisms underlying neurodegenerative diseases. Cell calcium, 2018. 70: p. 87-94.

163. Janicki, S.C. and N. Schupf, Hormonal influences on cognition and risk for Alzheimer's disease. Current neurology and neuroscience reports, 2010. 10(5): p. 359-366.

164. Cheng, J., et al., The emerging roles of protein homeostasis-governing pathways in Alzheimer's disease. Aging Cell, 2018. 17(5): p. e12801.

165. Moh, C., et al., Cell cycle deregulation in the neurons of Alzheimer's disease. Results and problems in cell differentiation, 2011. 53: p. 565-576.

166. Raina, A.K., et al., The role of cell cycle-mediated events in Alzheimer's disease. International journal of experimental pathology, 1999. 80(2): p. 71-76.

167. Terry, R.D., Cell death or synaptic loss in Alzheimer disease. Journal of Neuropathology & Experimental Neurology, 2000. 59(12): p. 1118-1119.

168. Coleman, P., H. Federoff, and R. Kurlan, A focus on the synapse for neuroprotection in Alzheimer disease and other dementias. Neurology, 2004. 63(7): p. 1155-1162.

169. Wang, J., et al., A systemic view of Alzheimer disease—insights from amyloid-β metabolism beyond the brain. Nature reviews neurology, 2017. 13(10): p. 612-623.

170. Kiselev, V.Y., et al., SC3: consensus clustering of single-cell RNA-seq data. Nature methods, 2017.

171. Bacher, R. and C. Kendziorski, Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol, 2016. 17: p. 63.

172. Trapnell, C., et al., The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol, 2014. 32(4): p. 381-386.

173. Wang, B., et al., Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods, 2017. 14(4): p. 414-416.

174. Kharchenko, P.V., L. Silberstein, and D.T. Scadden, Bayesian approach to single-cell differential expression analysis. Nature methods, 2014. 11(7): p. 740.

175. Finak, G., et al., MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome biology, 2015. 16(1): p. 278.

176. Vu, T.N., et al., Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics, 2016. 32(14): p. 2128-2135.

177. Wu, Z., et al., Two-phase differential expression analysis for single cell RNA-seq. Bioinformatics, 2018. 1: p. 9.

178. Wan, C., et al., LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data. Nucleic Acids Research, 2019. 47(18): p. e111-e111.

Curriculum Vitae

Xiaoyu Lu

Education

Indiana University–Purdue University Indianapolis, Indiana, USA *2017-2022*

Ph.D. in Informatics with Bioinformatics Specification
Minor: Biostatistics

Shandong University, Shandong, China *2013-2017*
B.S. in Statistics
Minor in Finance

Experience

Research Assistant - Indiana University School of Medicine *2017-2022*
Bioinformatics Internship - Bristol Myers Squibb *Jun-Aug 2021*
Single Cell Sequencing and Spatial Transcriptomics Co-Op - Merck & Co. *Sep 2021* - *Feb 2022*

Conferences Attended

AACR Annual Meeting 2019. Atlanta, Georgia, USA. Mar. 2019
2nd CCBB retreat, IU School of Medicine. Indianapolis, Indiana, USA. Oct. 2021
AACR Annual Meeting 2022. New Orleans, Louisiana, USA. Apr. 2022

Publications

1. **Xiaoyu Lu** and et al. (2020) SSMD: A semi supervised approach for a robust cell type identification and deconvolution of mouse transcriptomics data. Briefings in Bioinformatics (IF=11.622). GitHub
2. **Xiaoyu Lu**, Junyi Zhou and et al.. (2022) PLUS: Predicting cancer metastasis potential based on positive and unlabeled learning. PLoS Computational Biology. GitHub

3. Zhigang Cai, **Xiaoyu Lu** and et al. (2020) Hyperglycemia cooperates with Tet2 heterozygosity to induce leukemia driven by pro-inflammatory cytokine induced lncRNA Morrbid. Journal of Clinical Investigation (IF=11.864).
4. **Xiaoyu Lu** and et al. Cell-type specific variations of within pathway interaction in AD using covariance regression. 2021 Alzheimer's Association International Conference
5. **Xiaoyu Lu** and et al. An integrated web server for multi-omics data deconvolution. AACR 2020
6. **Xiaoyu Lu** and et al. A new deconvolution algorithm for accurate assessing immune and stromal cell populations in mouse transcriptomic data. AACR 2019
7. Zhou, Yi, et al. Acid-Base Homeostasis and Implications to the Phenotypic Behaviors of Cancer. bioRxiv.
8. Norah Alghamdi and et al. (2021) A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data. Genome Research (IF=9.043).
9. Silpa Gampala; Fenil Shah; **Xiaoyu Lu** and et al. (2021) Ref-1 Redox Activity Alters Cancer Cell Metabolism in Pancreatic Cancer: Exploiting This Novel Finding as a Potential Target. Journal of Experimental & Clinical Cancer Research. (IF=11.16).
10. Xu, Chengxian, et al. (2021) BATF Regulates T Regulatory Cell Functional Specification and Fitness of Triglyceride Metabolism in Restraining Allergic Responses. The Journal of Immunology
11. Rupert, Joseph E., et al. (2021) Tumor-derived IL-6 and trans-signaling among tumor, fat, and muscle mediate pancreatic cancer cachexia. Journal of Experimental Medicine
12. Jiannan Liu and et al. (2020) Transcription factor expression as a predictor of colon cancer prognosis: a machine learning practice. BMC Medical Genetics.
13. Menghao Huang and et al. (2019) Sestrin 3 Protects Against Diet- Induced Nonalcoholic Steatohepatitis in Mice Through Suppression of Transforming Growth Factor beta Signal Transduction. Hepatology (IF=14.679).
14. Changlin Wan and et al. (2019) LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data. Nucleic Acids Research (IF=11.797).
15. Samuel A Miller and et al. (2019) Lysine-specific demethylase 1 mediates AKT activity and promotes epithelial-mesenchymal transition in PIK3CA mutant colorectal cancer. Molecular Cancer Research (IF=4.630).
16. Chang, Wennan, et al. (2019) ICTD: A semi-supervised cell type identification and deconvolution method for multi-omics data. bioRxiv
17. Ballinger T, Marino N, German R, et al. Prospective, placebo-controlled, randomized study of metformin for breast cancer prevention in overweight/obese women. AACR 2020
18. Changlin Wan and et al. A statistical model to reveal transcriptional regulatory state in tumor single-cell RNA-seq data. AACR 2019

Services

Reviewer, SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2020, 2021
Reviewer, IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020, 2021
Reviewer, International Conference On Research In Computational Molecular Biology (RECOMB), 2021
Reviewer, Bioinformatics, 2019
Reviewer, International Conference on Intelligent Biology and Medicine (ICIBM), 2019