Machine Learning Approaches for the Integrative Analysis of Multi-omics Data

BY

SHANG GAO

B.S., Nankai University, 2011
M.S., Nankai University, 2014

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Bioinformatics
in the Graduate College of the
University of Illinois at Chicago, 2022

Chicago, Illinois

Defense Committee:

Jalees Rehman, Chair and Advisor
Yang Dai, Advisor
Salman Khetani
Beatriz Penalver Bernabe
Xiaowei Wang, Pharmacology and Regenerative Medicine

This thesis is dedicated to my mother, Rong Xiang, my father, Shixiao Gao and my wife, Shuang Song, without whom it would never have been accomplished.

ACKNOWLEDGMENTS

Firstly, I would like to express my sincerest gratitude to my advisors Dr. Jalees Rehman and Dr. Yang Dai, for their guidance, patience, and support during my graduate research. Dr. Rehman taught me how to think like a scientist by encouraging me to explore novel research questions, and by providing continuous guidance on how to assess the rigor of the research. Dr. Dai helped me develop my bioinformatics and computational biology skills. her guidance and advice on mathematical modeling was invaluable for the successful completion of all my projects.

I would also like to acknowledge the members of my thesis committee, Drs.Salman Khetani, Beatriz Penalver Bernabe, and Xiaowei Wang for their time and expertise in shaping my thesis project.

In addition, I would like to thank every member in the Rehman lab and the Dai lab for providing constructive computational and biological feedback on my research.

Most importantly, I would like to thank my family for their understanding, patience, and support throughout my entire Ph.D. journey.

CONTRIBUTION OF AUTHORS

In this dissertation, chapter 2 includes a previous publication:

Lianghui Zhang*, **Shang Gao*** , et al. "***Single-cell Transcriptomic Profiling Identifies Dynamic Inflammatory and Regenerative Endothelial Cell Subpopulations Following Vascular Injury***". JCI Insight. 2022: 2379-3708.

* indicates equal co-first authorship

Lianghui Zhang designed and conducted the mouse breeding, single cell isolation and experimental studies with histological sections and RNA FISH and prepared the manuscript sections involving biological experiments. Shang Gao analyzed the single cell RNA-seq data, visualized the RNA-seq analysis results, prepared all manuscript sections involving computational analysis. Zachary White conducted immunofluorescence experiments. Yang Dai provided the advice on data analysis and revised the manuscript. Asrar B. Malik provided guidance on the experimental design and revised the manuscript. Jalees Rehman conceived and designed the experiments, interpreted data, helped prepare the manuscript, and revised the manuscript.

In chapter 3, it includes a previous publication:

Shang Gao, Yang Dai, Jalees Rehman. "***A Bayesian inference transcription factor activity model for the analysis of single-cell transcriptomes***." Genome Res (2021).

Shang Gao designed and built the Bayesian hierarchical model, curated and processed the data, performed computational analysis, validated the results, prepared the initial manuscript draft, and revised the manuscript. Yang Dai provided advice on

modeling and data analysis and revised the manuscript. Jalees Rehman conceived and designed the computational approach, interpreted data, prepared the initial manuscript draft, and revised the manuscript.

In chapter 4, it includes a previous publication:

Shang Gao, Jalees Rehman, Yang Dai. "***Assessing comparative importance of DNA sequence and epigenetic modifications on gene expression using a deep convolutional neural network***". Comput Struct Biotechnol J 20, 3814-3823 (2022)

Shang Gao designed and built the deep learning framework, curated and processed the data, performed computational analysis, validated the results, prepared the initial manuscript draft, and revised the manuscript. Jalees Rehman provided advice on biological interpretation and revised the manuscript. Yang Dai conceived and designed the model, interpreted the data, prepared the initial manuscript draft, and revised the manuscript.

Table of Contents

LIST OF FIGURES

LIST OF ABBREVIATIONS

EC	Endothelial Cell
LPS	Lipopolysaccharide
DE	Differential Expression
TF	Transcription Factor
GO	Gene Ontology
ARI	Adjust Rand Index
RI	Rand Index
NMI	Normalized Mutual Information
UMAP	Uniform Manifold Approximation
CNN	Convolutional Neural Network
IG	Integrated Gradient
CMP	Common Myeloid Progenitor
MEP	Megakaryocyte–Erythroid Progenitor
GMP	Granulocyte-Macrophage Progenitor
DMR	Differentially Methylated Regions
DAR	Differential Attribution Regions

Summary

Multi-omics data refers to a collective term of different levels of “-omic” sequencing datasets. It can provide a comprehensive understanding of the regulatory mechanisms that link genotype to phenotype. A large number of “omics” datasets have been generated from studies in various biological scenarios but there is a need for advanced analytical tools that derive biological insights from these datasets.

In this thesis, I will develop three projects focusing on integrating multiple “omics” datasets to answer different biological questions. In the first project, I used dimension reduction and clustering methods on the single cell RNA-seq data to identify the cellular heterogeneity in the lung endothelial cells during lung injury and regeneration. We identified three major subpopulations in lung endothelial cells at baseline and time points post injury. One subpopulation is enriched for the expression of immune-related genes while another subpopulation is enriched for the expression of developmental genes. In the second project, I built a Bayesian Inference model, BITFAM, to infer the transcription factors activities in single cells by integrating single cell RNA-seq data and bulk ChIP-seq data. I was able to validate that BITFAM could indeed infer the transcription factors activities using known biological functions as well as a publicly available dataset in which transcription factor deletion was achieved by CRISPR/Cas9 deletion. In the third project, I assessed the attribution of DNA sequences and epigenetics modifications on gene expression by a deep learning framework, iSEGnet. I investigated the optimal regions that achieve the best prediction of gene expression. I also explored the most important regions and epigenetics modifications that impact gene expression as well as the regulatory mechanisms in these regions.

Chapter 1 Introduction

1.1 Multi-omics data

The comprehensive understanding of biological processes can leverage knowledge derived from distinct and complementary approaches to elucidate the molecular and cellular landscape of cells, such as the analyses of the genome, epigenome, transcriptome, proteome, and metabolome[1]. With the important development of sequencing technologies, biological research can increasingly leverage such comprehensive data generated at these levels, which together are referred to as “multi-omics” data[1-3].

1.1.1 Genomics

Genomics refers to the study of whole genetic information in an organism. It includes complete information that control every biological process in the living organisms[4]. Large DNA molecules (chromosomes) are used to keep the information. As DNA has four types of nucleotides, A, T, C and G, this genetic information is stored as distinct combinations of nucleotides in particular regions. Along the DNA, different regions will provide distinct sequences with specific functions. For example, gene coding regions are transcribed into mRNA and translated into protein sequences. The gene regulatory regions such as promoters and enhancers have the abilities of combining with other factors and regulating the biological procedures. The mutations of nucleotides in DNA may lead to significant changes of the organisms in their phenotypes because DNA variants or mutations impact downstream transcription, translation and protein function. Whole-genome sequencing (WGS) was developed based on next generation sequencing (NGS) to identify the information in genome[5]. It can reveal the whole DNA sequence in a

given genome. It has been widely used in the characterizing the mutations that drive the disease progression and predicting the risk of cancer development[6]. From whole-genome sequencing, we can also have the copy number variants (CNVs) which provides information on the number of copy variants in a given tissue[7].

1.1.2 Epigenomics

Epigenomics is the study of how gene expression by the modification of DNA without directly changing the DNA sequence. The regulations on epigenome level include changing the DNA structure to control the binding between DNA and other regulatory factors or modifying DNA base pairs and DNA-associated histones to alter their activities[8, 9]. The epigenome is dynamically changed by environmental factors and can vary across individual cells or cell types during homeostasis as well as in response to external cues such as pathogens [10]. There are multiple types of epigenomic modifications that have been identified and are essential to the regulation of gene expression[9].

1.1.2.1 Chromatin accessibility

The nuclear DNA of eukaryotes binds to proteins called histones. DNA wraps around histone proteins, thereby forming a beaded structure[11]. Such structures can also be further folded, condensed, and assisted by other architectural proteins to form chromosomes. DNA replication and gene transcription require unraveling the higher-order structure of DNA[11, 12]. The unfolding here can occur in certain segments, thus regulating what regions would be available for gen transcription. The regions could be the gene coding sequences or the regulatory regions to which regulatory proteins (such as transcription factors and cofactors) bind. The shifts of chromatin accessibility at key regions in the genome significantly impact the gene expression[13].

MNase-seq and DNase I hypersensitivity assay (DNase-seq) have been developed to investigate chromatin accessibility[14, 15]. The main idea of these two techniques is similar: Only accessible (unfolded) DNA can be cleaved by DNase (MNase or DNase I); therefore sequencing of cleaved DNA provides a snapshot of which DNA regions are accessible and which are not [14, 15].

Another method, ATAC-seq (Assay for Transposase Accessible Chromatin with high-throughput sequencing) also investigates chromatin accessibility[13, 16]. DNA transposition is a procedure in which the DNA sequence is transferred from one region of the chromosome to another, which is realized by DNA transposase. Such transposition insertion into DNA requires that the chromatin at the insertion site be open. Otherwise, it will be attached to higher-order structures. Therefore, it is necessary to artificially add a transposition complex carrying a known DNA sequence tag (that is, a transposase with a sequencing tag) into the nucleus and then use the tag of the known sequence to build a library and sequence it to know which regions are open chromatin[16]. Single cell ATAC-seq has been developed to identify the chromatin accessibility at a single cell level. It provides epigenetic information at a higher resolution in different cell types and states and reveals deeper insights into gene regulatory mechanisms[17].

1.1.2.2 Histone modification

As introduced in the previous section, DNA wraps around histones. The modifications of histones affect the affinity of histones and DNAs, thereby changing chromatin's loose or condensed state or affecting the affinity of transcription factors and gene regulatory regions[18]. There are several types of modifications on different positions of histones, for example, trimethylation of lysine 4 on histone H3 (H3K4me3), acetylation

of lysine 9 of histone H3 (H3K9Ac), and so on. Different histone modifications are associated with different functions in the regulation of gene expression[18, 19]. These regulations via histone modifications are the most important and diverse regulation of epigenetics, which acts on almost all biological processes, such as transcriptional activation/inactivation, chromosome packaging, and DNA damage/repair[20].

ChIP-Seq is widely used to quantify the histone modifications. ChIP-Seq is a technique for the genome-wide study of DNA-proteins interactions, histone modifications, and nucleosomes[21]. It contributes to understanding the gene expression regulation and the functional structure of chromosomes[22]. In ChIP-Seq experiments, first, the DNA in the cell is crosslinked with the protein. Then the cell is lysed, and the chromatin is randomly cut by ultrasonic or enzyme treatment. Using the specific reaction of the antigen and antibody, the DNA fragments bound to the target protein are precipitated. Next these DNA fragments are released by reverse crosslinking. Finally, the sequences of the DNA fragments are obtained by sequencing[22, 23]. ChIP-Seq obtains the different histone modification profiles. The histone modification ChIP-Seq has been widely used to understand cell identity and disease state. There are publicly available databases (ENCODE[24], Roadmaps[25]) with comprehensive histone modification ChIP-Seq data in different cell types and conditions.

1.1.2.3 Transcription factor binding

Transcription factors (TFs), known as trans-acting factors, refer to a type of proteins that can specifically interact with cis-regulatory elements of genes and activate or inhibit gene transcription. TF has essential functions in the regulation development and the immune response. The functional study of TF and its interacting factors is crucial to

understanding their roles in signaling cascades, providing theories for basic research and production applications.

ChIP-Seq is also widely used to investigate the binding of DNA and transcription factors[26]. By using the specific antibody of the transcription factor, the sequences of the DNA fragments that bind with transcription factor are obtained by sequencing. There are publicly available databases (GTRD[27]) with transcription factors ChIP-Seq data in different cell types and conditions.

1.1.2.4 DNA methylation

The DNA methylation has essential functions in multiple biological process[28], such as gene expression regulation, chromosome stability, X chromosome inactivation[29]. In mammals, DNA methylation occurs primarily at the fifth carbon atom of the cytosine base, forming 5-methylcytosine or 5-methylcytosine (5-mC)[30]. DNA methylation is a critical epigenetic modification that regulate gene expressions. The methylation on the CpG sites at gene promoters are proved that can repress gene expression[30, 31].

Whole Genome Bisulfite Sequencing (WGBS) is developed to analyze DNA methylation at single-base resolution[32, 33]. WGBS combines sodium bisulfite conversion of sequences with high-throughput DNA sequencing. The sodium bisulfite reaction protects methylcytosine from conversion, while unmethylated cytosine is converted to uracil. After PCR, the unmethylated cytosines are converted to thymines, while the methylated cytosines will appear as cytosines. In the sequencing data, we can determine the percentage of methylated cytosines by checking the ratio of thymines and cytosines[34].

1.1.3 Transcriptomics

Transcriptome is the sum of all RNAs that are transcribed by a living cell. It is an important method to study cell phenotype because a cell's phenotype is in large part due to the presence of proteins that are translated from RNA. Transcriptomics is a study about gene transcription and transcriptional regulation in cells at the overall level. RNA-sequencing (RNA-seq) has become an indispensable tool for analyzing gene expression at the whole transcriptome level and studying differential splicing of mRNA in the past decade[35-37]. With the development of next-generation sequencing (NGS) technology, RNA-seq are applied to many RNA-level studies, such as single-cell gene expression, RNA translation, and RNA structure. In addition, new and exciting applications such as spatial transcriptomics are also being developed. Combining with better computational analysis tools, RNA-seq can help us understand transcriptome more and more comprehensively[38].

In the RNA-seq experiment, first, mRNAs are extracted from the samples. Then the mRNAs are converted to cDNAs. The cDNAs are amplified by PCR and construct a sequencing library. Sequencing is performed on a high-throughput platform with 5–200 million reads per sample. The length of the reads is usually between 75–125 bps. After sequencing, the raw reads are aligned to the reference genome to determine the gene they belong to. For every sample, the expression of each gene is identified by the number of reads that gene has. Finally, the statistical analysis is conducted between samples to identify differentially expressed genes.

1.2 Integration analysis of multi-omics data

Analysis of multi-omics data along with clinical information is helping researchers derive valuable insights into the cellular functions of disease and generate hypotheses which can then be tested in functional studies[39]. Integration of these distinct multi-omics datasets will also allow researchers to systematically and holistically interrogate the complexity of biological systems[40]. When integrating the omics data in a sequential manner, the interactions between different regulatory mechanisms can be identified[41]. Integrated analyses assess the information flow from one "-ome" level to the a different level, and thus facilitate connecting the knowledge gap from genotype to phenotype[1]. For example, integrative analyses of ChIP-Seq and RNA-Seq data have shown that cancer-specific epigenetic modifications are associated with transcriptional changes in cancer driver genes[42-45]. The combination of Whole Genome Bisulfite Sequencing and RNA-seq revealed that the shifts in methylation lead to dysregulation of key genes' expression in various diseases[46-48]. The expression quantitative trait loci (eQTL) analysis integrates the genomic sequences and transcriptomic profiles to explains the fraction of the genetic variance in a gene expression phenotype[49-51]. Moreover, the rapid development of single cell sequencing technology moves omics studies into the level of individual cells. For example, scRNA-seq studies combined with scATAC-seq and CITE-seq revealed regulatory heterogeneity in cell development and inflammatory responses[52-55]. These recent studies have demonstrated that integrating multi-omics data can provide deep biological insights and that it has wide applicability in various biological scenarios.

Efficient and accurate analysis tools are essential in integrating and analyzing large amounts of data generated from multi-omics studies. Machine learning, including

deep learning models, may be of value in such integrative approaches by projecting distinct multi-omics datasets into the same “space” and perform downstream analyses with the new latent features[56-60]. Some methods of integration are motivated by predicting clinical information and using incomplete -omics data as not all patients undergo complete -omics profiling at all levels. These integration methods are usually based on supervised learning models, such as Support Vector Machines[61-63], ensemble-based approaches[64] and deep neural networks[65-69]. Recently, single cell multi-omics analysis tools are developed to integrate different single cell levels omics datasets. These methods share many underlying concepts with the integration methods developed for bulk cell data. The widely used Seurat approach performs dataset integration and batch removal using canonical correlation analysis[70]. LIGER employs integrative nonnegative matrix factorization (iNMF) for integration[71]. MOFA is using a matrix decomposition method that generates a loading matrix and a weight matrix[72].

The existing methods offered a variety of options for multi-omics data analyses; however, key challenges remain in this field. First, the large size of data sets leads to computationally intensive analyses, and the different formats between omics data generate major obstacles in standardizing preprocessing steps such as data filtering, data normalization, batch correction, and quality control. Second, there are many publicly available databases for omics data[73-75]. However, there is a lack of solid approaches to prioritize the knowledge from existing databases and transfer the knowledge to different omics data analyses in order to reveal deeper biological insights. Also, integrating the omics data from bulk and single cell levels remains a challenging problem. Third, because

of the difficulties in model interpretation, it is extremely challenging to ascertain biologically meaningful relations between distinct omics data.

1.3 Thesis outline

This dissertation focuses on the application and development of machine learning and deep learning models on the multi-omics data to reveal biological insights. The projects presented in this dissertation are organized as follows:

In Chapter 2, we discuss the application of machine learning in the identification of lung endothelial cell subpopulations during lung injury and regeneration by single cell RNA sequencing. The analysis approach based on unsupervised learning is applied to the scRNA-seq data. Dimension reduction, clustering, and trajectory building algorithms are used in the approach. The results show that there are distinct endothelial cell subpopulations in the lung at different time points, including baseline and post injury. One major subpopulation predominantly expresses immune genes, and another one predominantly developmental genes. In the late stage of lung injury, a subpopulation with high expression on cell cycle genes is identified. This chapter is based on the publication:

Lianghui Zhang*, **Shang Gao***, Jalees Rehman. “Single-cell Transcriptomic Profiling Identifies Dynamic Inflammatory and Regenerative Endothelial Cell Subpopulations Following Vascular Injury”. **JCI Insight.** 2022: 2379-3708 (* co first authors)

In Chapter 3, we develop a Bayesian hierarchical model (BITFAM) to infer transcription factor activity by integrating single cell RNA-seq (scRNA-seq) data and existing biological data on transcription factor binding sites. BITFAM is applied on multiple

single cell RNA-seq datasets from healthy mouse organs or blood cell development. The inferred transcription factor activities from BITFAM correspond to the known functions of the cell types. This inferred transcription factor activity profile can also be used to identify cell heterogeneity as well as other downstream analysis. The inferred regulatory strengths between transcription factors and target genes also match the biological functions of the transcription factor. This chapter is based on the publication:

Shang Gao, Yang Dai, Jalees Rehman. "A Bayesian inference transcription factor activity model for the analysis of single-cell transcriptomes." **Genome Res** 2021, 31: 1296-1311.

In Chapter 4, we develop a deep learning frame (iSEGnet) to predict the gene expression by DNA sequence and epigenetics modifications on the gene regulatory regions. The optimal regions that make the best prediction are learned by input different combinations of regulatory regions. Then, the attribution of the influence of DNA sequence and epigenetic modifications on gene expression is assessed by an Integrated Gradient. For each epigenetic modification on regulatory regions of every gene, the most important regions with high attributions are identified by this approach. We find that the high attribution regions overlap with known transcription factor ChIP-seq data. This suggests that these regions predominantly impact gene expression by affecting TF binding. Finally. iSEGnet is applied to two cancer multi-omics data. The results from the model show that the potential application of our model in identification of the gene expression associated epigenetics modifications and regulatory regions as well as the

regulatory mechanism that drive the cancer development. This chapter is based on the publication:

Shang Gao, Jalees Rehman, Yang Dai. “Assessing comparative importance of DNA sequence and epigenetic modifications on gene expression using a deep convolutional neural network”. **Comput Struct Biotechnol J** 20, 3814-3823 (2022)

1.4 Significance of the Thesis Work

Developing and applying machine learning models to analyze multi-omic data could help uncover meaningful biological insights about underlying regulatory mechanisms. In the single cell RNA-seq analysis of lung endothelial cells, the cellular heterogeneity of ECs during lung injury and recovery for the first time tracks the dynamics of lung endothelial injury and repair. This is made possible by the longitudinal scRNA-seq analysis with 6 time points which provides a comprehensive and high temporal resolution landscape of single cell gene expression profile shifts.

The BITFAM model integrates scRNA-seq data and transcription factor ChIP-seq data. In the existing methods of transcription factor activity inference, the target genes of transcription factors are selected based on the correlation of gene expression. The ChIP-seq transcription factor binding information leverages prior knowledge about the potential targets of transcription factors. Armed with this prior knowledge, we developed a novel Bayesian hierarchical model that could infer the transcription factor activities in individual cells from single cell RNA-seq data as well as the regulatory strength between a transcription factor and its potential target genes. This was the first approach to integrate ChiP-seq data with single cell RNA-seq data and will likely impact the development of additional integrative approaches.

In iSEGnet, we integrate genomic sequence information, multiple epigenetic modification profiles, and transcriptomic data. Our framework uses genomic sequences and epigenetic modifications as two sources of inputs to predict gene expression (Transcripts per million counts, TPM). Then, we apply the integrated gradient to assess the relative attribution of DNA sequence and epigenetics profiles in distinct regions on gene expression. We demonstrate that this attribution has an important biological significance because it identifies the most active cis-regulatory regions. This provides a new approach to apply a deep learning model in multi-omics data analysis.

Chapter 2 Single cell RNA-seq data analysis to identify the cell heterogeneity during lung injury and regeneration

Adapted from Zhang, L., Gao, S., et al. (2022) Single-cell Transcriptomic Profiling Identifies Dynamic Inflammatory and Regenerative Endothelial Cell Subpopulations Following Vascular Injury. JCI Insight. 2022: 2379-3708.

2.1 Introduction

Endothelial cells (ECs) located in blood vessels play important roles in many functions such as immune cell transmigration across the vascular barrier, regulating thrombosis and hemostasis, controlling metabolites transporting to the tissue and expansion or regeneration of the vascular system after injury[76]. The signature genes expressed in endothelial cells in specific tissue have been identified in multiple recent studies of the vascular endothelium. In the lung endothelium, the genes with functions in immune responses and inflammatory are significantly highly expressed compared to the expressions in endothelial cells in the brain and heart[77]. The endothelium has critical functions in the dysfunction of organ specific vascular disease[76]. This leads to essential questions: are there distinct subpopulations of ECs in the immune response and regeneration post-injury? Can the treatment of specific endothelial cells repress inflammatory response or activate endothelial regeneration?

Single-cell RNA sequencing (scRNA-seq) is an effective experimental method to identify the heterogeneity of cells in transcriptomics level and provide gene expression profiles for individual cell[78]. Many scRNA-seq studies have been conducted to reveal subpopulations of cells in healthy tissue and during the inflammation and recovery in injury, e.g., identification of macrophage subpopulations during neuroinflammation in the

brain[79], identification of macrophage subpopulations in myocardial infarction, and the identification of macrophage subpopulations during fibrosis in the lung[80, 81]. scRNA-seq analysis has already been applied to identify endothelial subpopulations in different organs such as the brain[82], aorta[83], and lymphatic vessels[84]. In lung endothelium, single cell RNA-seq analysis identified an EC subpopulation with a signature gene carbonic anhydrase 4 (Car4). Car4 expression relates to the expression of VEGFA in alveolar type I epithelial cells[85]. This Car4 lung EC subpopulation has been defined as “aerocytes” as the expression of genes related to gas exchange increases, whereas other endothelial cells were defined as general capillary endothelial cells (gCap).

This study explored the distinct lung EC subpopulations in the inflammatory response with an acute lung inflammatory injury mouse model. The bacterial endotoxin lipopolysaccharide (LPS)[86] is used to induce the injury and inflammation of the lung endothelial cells. After the injury, an adaptive regeneration of endothelial cells is regulated by the developmental transcription factors such as SRY-box transcription factor 17 (Sox17)[87]. However, whether the endothelial inflammatory response and regeneration response are conducted by the same subpopulations of lung ECs or distinct subpopulations of lung ECs remains unknown. LPS-induced injury is well-suited for this temporal study because the model is calibrated for temporal quantification. In addition, another single cell RNA-seq study with LPS has revealed the heterogeneity of alveolar epithelial cells (AECs)[88].

In this study, we identified distinct lung EC subpopulations in healthy mice and post-injury and regeneration with single cell RNA-seq data. We found that at baseline, there is an EC subpopulation with inflammatory response functions (immuneEC) and an

EC subpopulation with developmental functions (devEC). The immuneEC had a higher response to the LPS treatment compared to devEC. Additionally, there is an EC subpopulation with proliferation and cell cycle functions (proEC) in the late stage of post-injury regeneration. Trajectory analysis on the endothelial cell subpopulations demonstrated that the proliferative EC subpopulation emerges from the developmental EC subpopulation. These results suggest that EC plasticity and EC phenotypic change are critical in supporting the regeneration of vascular. Chronic inflammation in the lungs may result from these adaptive systems failing.

2.2 Methods

2.2.1 Mouse models of inflammatory lung injury

The mouse experiments were conducted based on NIH guidelines for the Care and Use of Laboratory Animals and approved by the IACUC of the University of Illinois. The crossbred of tdTomatofl/fl mice and Cdh5-CreERT2 mice were used to generate tdTomatofl/fl:Cdh5-CreERT2 mice which is an inducible endothelial cell lineage. Several generations of crossbred tdTomatofl/fl or Cdh5-CreERT2 transgenic mice and C57BL/6J mice were generated. Tamoxifen (0.1mg/g, i.p., Sigma-Aldrich, Cat#: T5648) in 100µl Corn Oil (Sigma-Aldrich, Cat#: C8267) was injected to the adult mice (two to four month old, tdTomatofl/fl:Cdh5-CreERT2 mice), three times (once a day), to active the Cre recombinase. One month later, the mice were injected with sublethal dose of Lipopolysaccharide (LPS) (12.5 mg/kg, i.p., Sigma-Aldrich, Cat#: L2030).

2.2.2 Single cell capture and library preparation

Following the instructions from 10x Genomics, the tdTomato+/DAPI- Cells were put into the microfluidics chip packaged with barcoded oligo-dT-containing gel beads by

10x Genomics Chromium controller. At each time point, more than 80% of cells should be alive and 5,000 cells at each time point were the sequencing targets. We built the single-cell cDNA libraries by v2 kit following the instructions provided by 10x Genomics. At each time cDNA quality was high. The cDNAs from different time points were multiplexed for sequencing on NovaSeq 6000 S4 (~100k reads per cell and per lane 2x150bp Paired-Read) in two lanes with a sequencing depth of ~200,000 reads per cell.

2.2.3 Processing of scRNA-seq data

Cellranger (v3.0.0) was applied on raw reads fastq files to generate the raw counts table for each time point. Default parameters were used in Cellranger. After quality control, there were 35,973 cells obtained for the downstream analysis. The numbers of cells in each time point are shown in table 1.

2.2.4 Analysis of scRNA-seq data

R package Seurat (v2.3)[70] was used to perform scRNA-seq data analysis and visualization. We analyzed the data at each time separately. NormalizeData() and ScaleData() functions from Seurat was used to perform Data normalization and scaling. The parameters in these two functions were default parameters. FindVariableGenes() function was used to identify the most variably expressed genes. We set the parameters in FindVariableGenes() with x.low.cutoff as 0.0125, x.high.cutoff as 3 and y.cutoff as 0.5. RunPCA() function was used to perform principal component analysis as dimension reduction method with the most variably expressed genes. FindClusters() function was used to identified the subpopulations with the top 20 principle components. In the FindClusters(), we set the parameters k.param as 10 and resolution as 0.4 to get a better clustering result. For the visualization of scRNA-seq data, we used the uniform manifold

approximation (UMAP) to visualize the similarity of transcriptomic profiles between cells in a 2D plot. RunUMAP() function was used to perform UMAP with top 20 principle components. FindMarkers() function was used to find the positive markers in each cluster identified by clustering analysis. The non-parametric Wilcoxon rank sum test was used to identify differentially expressed genes between subpopulations. Multiple testing corrected was conducted by Benjamini-Hochberg method

2.2.5 GO enrichment analysis

The enrichment analysis with hypergeometric test was conducted on the marker genes of each subpopulation by the DAVID web server to reveal the signature functions of the subpopulations. Biological Process (BP) level 4 terms were used in this analysis. The testing p-values were corrected by Benjamini-Hochberg method. The statistically significant threshold was set to 0.05.

2.2.6 Analysis of transcription factor activities by SCENIC

SCENIC was used to identify the transcription factors activities in each subpopulation[89]. We used default parameters in SCENIC. First, the most variably expressed genes which are also co-expressed with transcription factors were identified as target genes of transcription factors. A random forest model was used in this step. Then, the most important TF target genes were chosen by Genie3Weights. In each subpopulation, AUCell() function in SCENIC was used to reveal transcription factors activities in each cell based on the expressions of their target genes. In the heatmap, for each subpopulation, the transcription factors activities were the mean of the transcription factors activities in each cell belongs to the subpopulation. In order to make the activities

comparable between transcription factors, The activity scores were rescaled to [0, 1] across the subpopulations

2.2.7 Trajectory building

R package Monocle 3[90, 91] was used to identify the cell trajectory. The most variably expressed genes were involved in this analysis. DDRTree was used in trajectory building on a low dimensional space which was generated by PCA.

2.3 Results

tdTomatofl/fl:Cdh5-CreERT2 mice were used in this study to label the endothelial cells. tdTomato was only expressed in the endothelial cells. tdTomato protein is induced by tamoxifen and displays read florescent. Lung ECs was isolated by fluorescence-activated cell sorting (FACS) at multiple time points, including baseline, 6 hours, 1 day, 2 days, 3 days, and 7 days post LPS injury (Figure 1A). Dead cells were removed by DAPI nuclear labeling. Finally, the cell which are tdTomato+ DAPI- were used for single cell RNA sequencing. 35,973 individual endothelial cells were obtained in total. At each time point, there are 5,000 to 8,000 cells. For each cell, 1,100 to 1,900 genes were expressed (Table 1). The landscape of the similarity of gene expression profiles between all 35,973 cells were shown in the UMAP plot (Figure 1B). The expressions of hematopoietic lineage-specific gene Cd45 and endothelial-specific gene Cdh5 (VE-cadherin) were used to evaluate the specificity of cells in the dataset. In Figures 2A-B, we could find that endothelial cell marker gene Cdh5 was expressed in 99% of the isolated cells. We also assessed Prox1 and Nr2f2 expression in the cells. Prox1 and Nr2f2 are the lymphatic endothelial cell markers. We found a lymphatic endothelial subpopulation and remove it from the downstream analyses.

2.3.1 Lung endothelial cell subpopulations at baseline

There were 8,191 endothelial cells in healthy mice involved in the cell clustering analysis. We first did PCA on the most variably expressed genes expression profiles. Then, the Louvain's algorithm was used on the top 20 principal components to discover distinct subpopulations of endothelial cells. We used Seurat to implement this approach [70]. 8 clusters were identified based on the transcriptomic similarities between cells from the clustering approach (Figure 1C). We performed differential expression analysis on each of the subpopulation to identify their marker genes. Then gene ontology (GO) enrichment analysis was conducted on the marker genes of each subpopulation. We found there are distinct biological functions that are enriched by the marker genes from different EC subpopulations. In Figures 1D-F, we show the GO terms with different biological processes which are the signature functions of the largest 3 EC subpopulations. The signature functions of other subpopulations are shown in Figures 2A-E.

In the 3 largest EC subpopulations, the signature functions of cluster 1 were enriched for inflammatory response and immune regulation. The signature functions of cluster 2 were related to vascular development, and cluster 3 were characterized by pro-migratory functions (Figures 1D-F). The expression of the most significant marker genes of these three subpopulations are shown in the heatmap (Figure 1G). For cluster 1, the marker genes are important genes in regulating adaptive and innate immune responses. For example, H2-Aa (histocompatibility 2, class II antigen A, alpha), and H2-Ab1 (histocompatibility 2, class II antigen A, beta 1) (Figures 1G), were among the most significant marker genes of cluster 1. They are key genes in major histocompatibility complex (MHC) class II for processing and presenting the antigen [92, 93]. For cluster 2, the

marker genes played an essential role in endothelial cells development and differentiation. For example, Junb (Jun B proto-oncogene), Egr1 (early growth response 1), and Sox17 (sex determining region Y – box17) (Figures 1G) are highly expressed in cluster 2 specifically. These genes are important to regulate cell development[94]. For cluster 2, some marker genes are Ednrb (Endothelin receptor type B), Emp2 (Epithelial membrane protein 2), and Itgb5 (Integrin beta 5) (Figures 1G). These genes have crucial functions for cell migration, cell adhesion and actin filament organization[95, 96]. Additionally, Car4 was identified as a marker gene of cluster 3. Based on the results from previous study, Car4 is the marker of one lung EC subpopulation which is defined as “aerocyte” [97].

Immunofluorescence staining was used to validate whether the marker genes identified by single cell RNA-seq data are expressed only in a subset of ECs. H2-Ab and Sox17 are selected to do immunofluorescence staining. They have important functions regarding to the function of cluster 1 and cluster 2. We found that the lung endothelial cells which expressed CD31 (red) were expressing either H2-Ab (white) or Sox17 (green) (Figures 1H-I). The fraction of endothelial cells that expressed both H2-Ab and Sox17 is small (Figure 1I). This demonstrates that there are two distinct subpopulations of endothelial cells that have immune functions or developmental functions.

Therefore, using clustering and differential expression analysis on the single cell RNA-seq data we revealed lung endothelial subpopulations at baseline. There were 3 major subpopulations. One cluster highly expressed immune regulating genes, such as H2-Ab1 and H2-Aa. We defined this cluster as immuneEC. One cluster had endothelial development functions with marker genes such as Sox17 and Junb. We defined this cluster as devEC. Another major subpopulation of endothelial cells exhibited cell

migration and cell adhesion. They were defined as aerocyte according to the previous study [85, 97]. For each major subpopulation, there are 857, 774, and 313 differentially expressed genes, respectively, although more than 12000 genes are found expressed in the cells from these 3 subpopulations.

We also analyzed the single cell RNA-seq data from healthy human lungs. There are 18,253 lung ECs selected from the human data[98]. We found 5 clusters in the human lung ECs (Figure 3A) using the same clustering approach as our mice single cell data analysis. Among these 5 clusters, we also performed GO enrichment analysis on the marker genes of subpopulations. We found that two major EC subpopulations had signature functions of immune response or cell development, respectively. We also defined them as immuneEC (cluster 1) and devEC (cluster 2) (Figure 3B, C).

In Figure 3D, the most significantly differentially expressed genes in distinct EC subpopulations were shown in the heatmap. The expression pattern of marker genes in immuneEC subpopulations and devEC subpopulations are visualized by uniform manifold approximation (UMAP). Here, we selected HLA-E, CD74 and B2M for immuneEC, Sox7, ENG and TEK for devEC (Figure 3E, F). These results indicated that human lung endothelial cells also had distinct subpopulations with signatures functions at baseline similar to the endothelial cells in healthy mouse lungs.

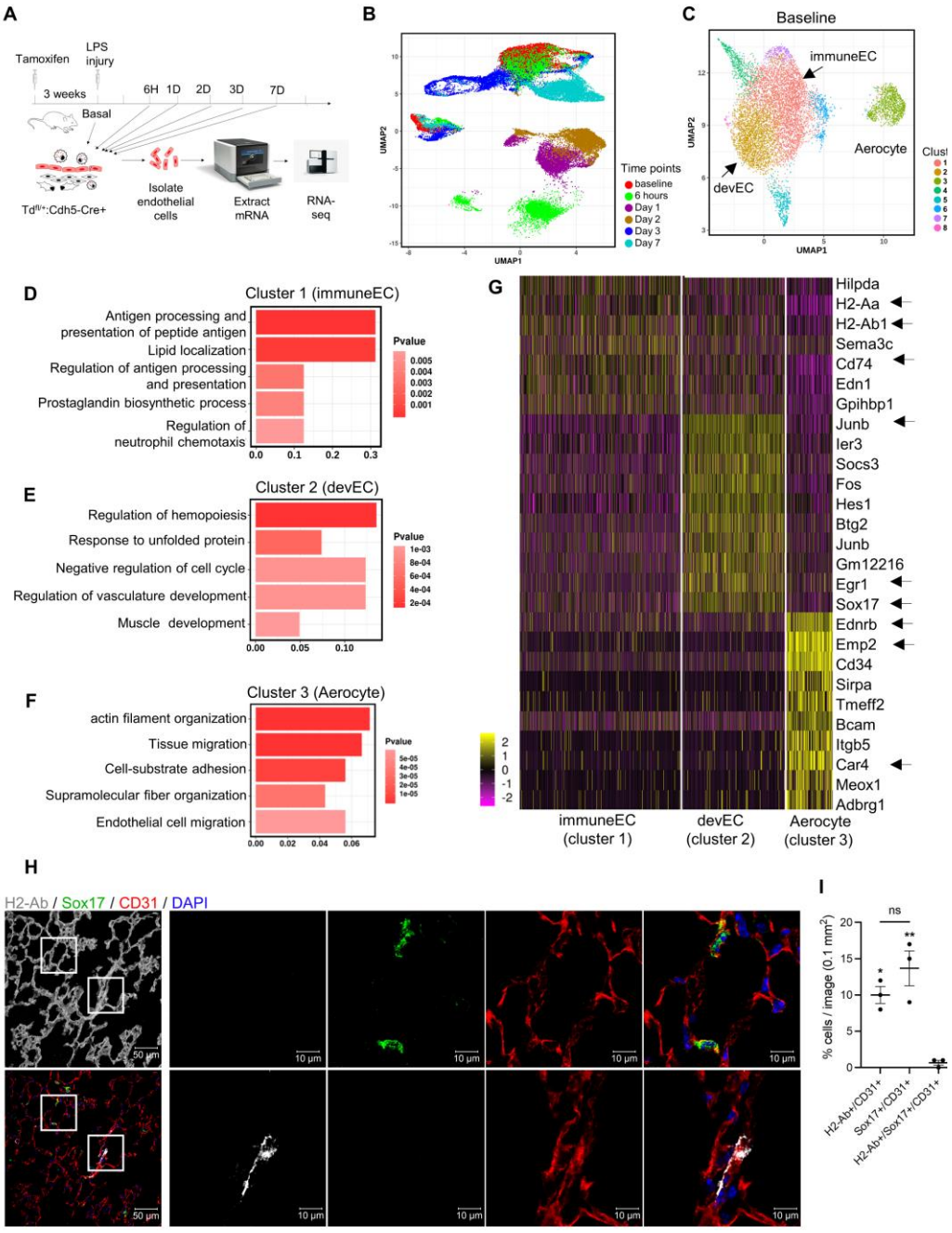

Figure 1. Overview of scRNA-seq of lung EC and analysis on the healthy lung EC subpopulations.

(**A**) lung EC experiment schema. (**B**) UMAP of all cells in the study. Cells are colored by the time point. (**C**) UMAP of 8,191 ECs at baseline. Cells are colored by the Seurat clusters. (**D**)(**E**)(**F**) The signature functions of the marker genes in cluster 1, 2 and 3. (**G**) The heatmap of the most significant maker genes for cluster 1, 2 and 3. (**H**) Immunofluorescence staining results for H2-Ab (Grey), Sox17 (Green) and CD31 (Red). (**I**) Quantification of endothelial cells with different proteins.

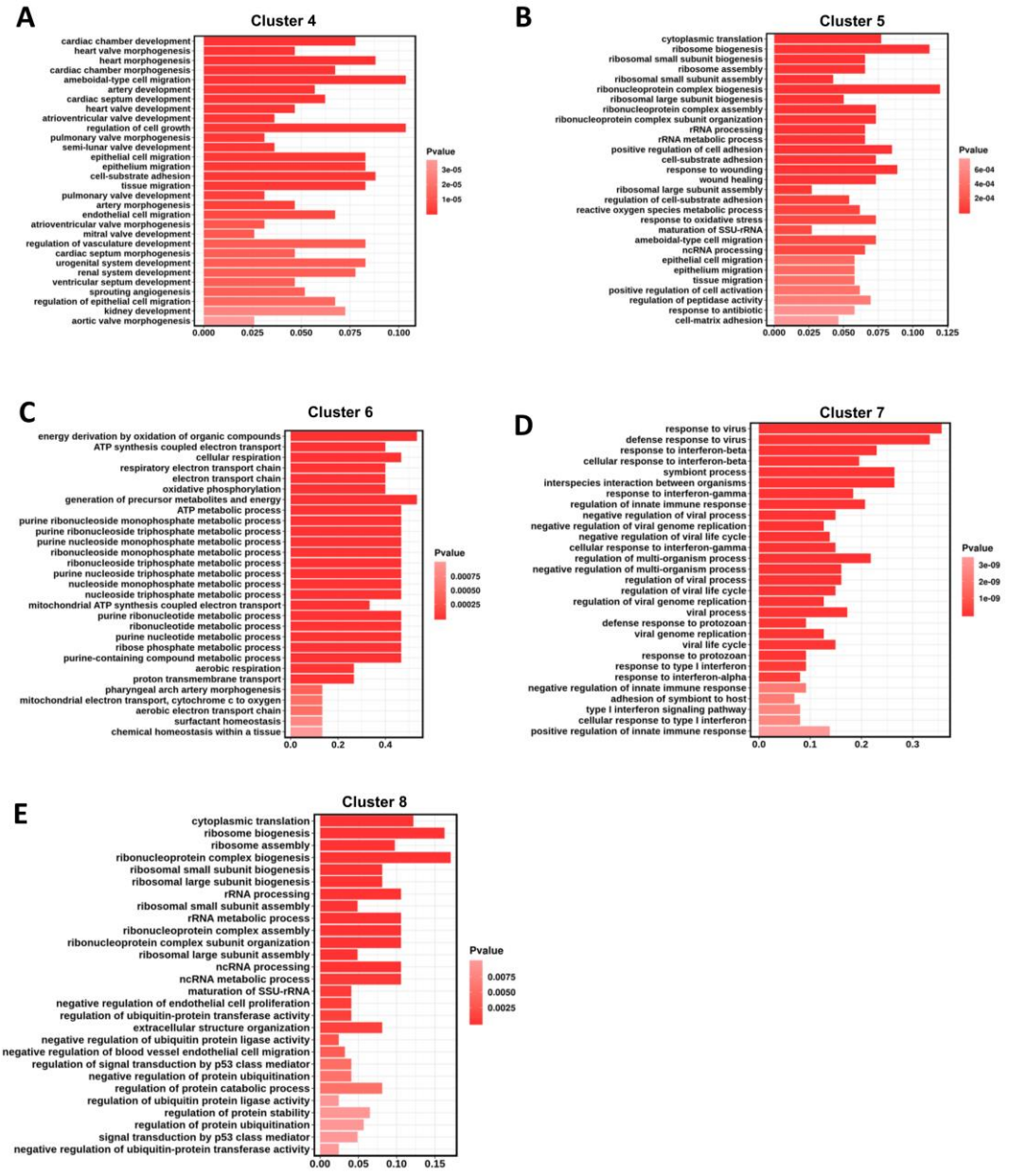

Figure 2. The signature GO terms that enriched on the marker genes of EC cluster 4, 5, 6, 7 and 8 at baseline. The color bar shows the Pvalue of enrichment testing.

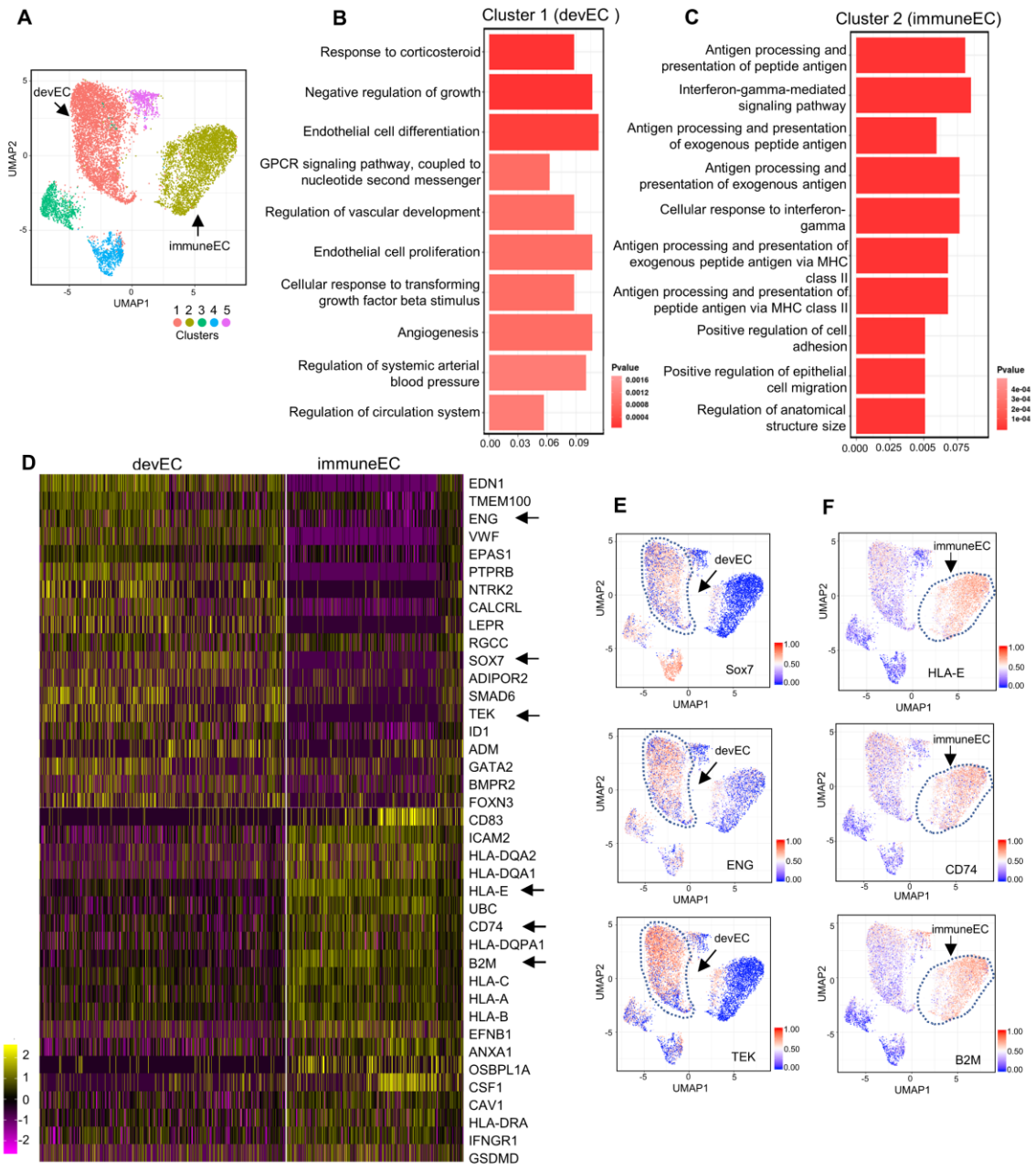

Figure 3. Identification of human lung microvascular EC subpopulations.

(**A**) UMAP of 18,253 human healthy lung ECs. Cells are colored by the Seurat clusters. (**B**)(**C**) The signature functions of the marker genes in cluster 1 and 2. (**D**) The heatmap of the most significant maker genes for cluster 1 and 2. (**E**)(**F**) UMAP of 18,253 human healthy lung ECs. Cells are colored by scaled the expression of the selected marker genes.

2.3.2 The EC subpopulations at 6 hours post LPS injury

Next, we explored the shifts of transcriptomic profiles in immuneECs and devECs in the early stage of LPS injury. When treated with LPS, after 6 hours there were significant differences in the gene expression of endothelial cells. We performed the same analysis on the single cell RNA-seq data obtained 6 hours post LPS injury. We found 8 clusters in the ECs. The UMAP plot showed that at this time point there are much more distinct subpopulations (Figure 4A) compared to the endothelial subpopulations at baseline. We did differential expression analysis and GO enrichment analysis on these subpopulations to identify the marker genes and signature functions. We found that the marker genes in cluster 1 had functions of immune and inflammatory response (Figure 4B). The marker genes in cluster 2 were enriched in the functions with vascular development (Figure 4C). This finding was consistent with the result we observed in the healthy stage. However, the transcriptomic differences are much larger between immuneEC and devEC after the LPS treatment.

In Figure 4D, we showed the most significantly differentially expressed genes in distinct EC subpopulations at 6 hours post LPS injury with a heatmap (Figure 4D). Similar to the baseline, the expression pattern of marker genes in immuneEC subpopulations and devEC subpopulations are visualized by uniform manifold approximation (UMAPs). We showed the expression distributions of marker genes, Irf7, Icam1 and Il4ra for immuneEC and Sox17, Ace and Kdr for devEC (Figure 4E, F).

RNA fluorescence in situ hybridization (RNA FISH) and immunofluorescence were used to validate whether the marker genes identified by single cell RNA-seq data are expressed only in a subset of ECs at 6 hours post LPS injury (Figures 4G-H). Irf7 and

Sox17, marker genes of cluster 1 and cluster 2, were selected in this validation (figure). The results demonstrated that the number of cells expressing Irf7 is much larger than the number of cells expressing Sox17 (Figures 4H). Importantly, at this time point, the expression of Irf7 and Sox17 in lung ECs are mutually exclusive at both protein and mRNA levels. So, we found that there were distinct subpopulations in lung ECs which had inflammatory functions (immuneEC) or developmental functions (devEC) during the early phase of immune response.

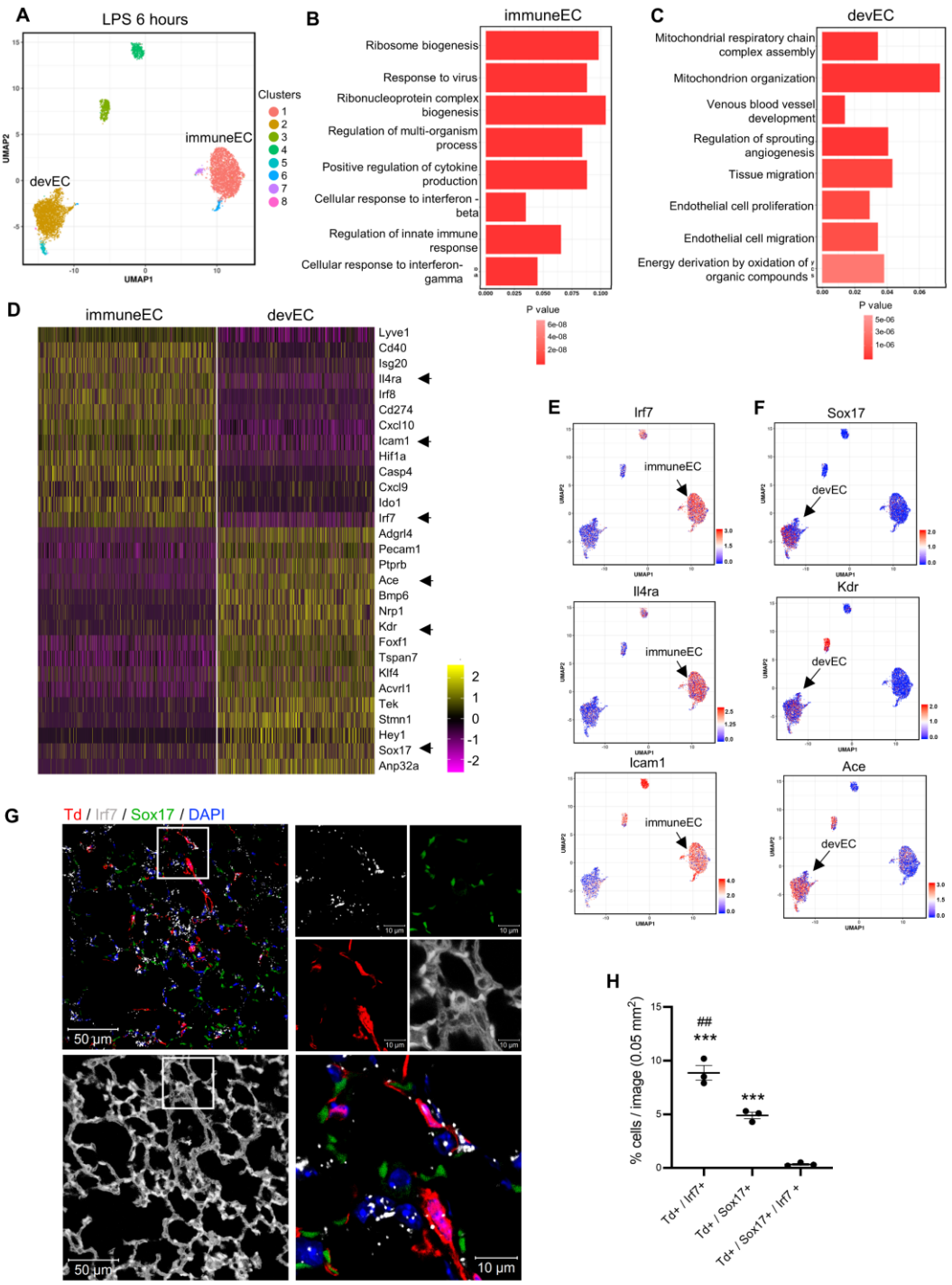

Figure 4. scRNA-seq identified lung EC subpopulations at 6 hours post LPS-injury.

(**A**) UMAP of 6,527 cells at 6 hours post LPS-injury. Cells are colored by the Seurat clusters. (**B**) (**C**) The signature functions of the marker genes in cluster 1 and 2, which have immune response function and developmental function, respectively. (**D**) The heatmap of the most significant maker genes for immuneEC and devEC. (**E**)(**F**) UMAP of 6,527 cells at 6 hours post LPS-injury. Cells are colored by scaled the expression of the selected marker genes. (**G**) Immunofluorescence staining results for H2-Ab (Grey), Sox17 (Green) and tdTomato (Red). (**H**) Quantification of endothelial cells with different proteins.

2.3.3 The EC subpopulations at 24 hours post LPS injury

Then, we investigated the changes of gene expression profiles in the immuneECs and devECs 24 hours post LPS injury. The same analysis was performed on these data. We observed the immuneEC and devEC subpopulations at 24 hours (Figures 5A-E). RNA fluorescence in situ hybridization (RNA FISH) and immunofluorescence were applied to validate the quality of the marker genes. We found that the Sox17 and Irf7 were the in-situ marker genes of developmental EC subpopulation and the immune EC subpopulations (Figures 5F). The result suggested that the number of cells expressed Irf7 are much larger the number of cells expressed Sox17 (Figures 5G). This finding was consistent with the result at 6 hours post LPS injury. And the fraction of cells expressed both Irf7 and Sox17 was very small.

In our previous study, we found that the regeneration of endothelial cells occurs on day 2 post LPS injury[87]. The single cell RAN-seq data at this time point showed that there were distinct subpopulations in lung ECs. Based on the DE analysis and GO enrichment analysis, immuneEC and devEC subpopulations were also identified from the data (Figures 6A-C). In Figure 6D, we showed the most significant marker genes in immuneEC and devEC subpopulations were at 2 days post-LPS injury. In the devEC, Acvrl1, Ace and Sox17 were highly expressed whereas Cd24, Cebpb and Nfkbia were highly expressed in immuneEC. The expression profiles of these genes demonstrated that resolution of inflammation had begun in immuneEC, whereas Cd24, Cebpb and Nfkbia are key genes that negatively regulate inflammation[99-101].

The distinct transcriptomic signatures of immuneEC and devEC were shown in a volcano plot (Figure 6E). Immunofluorescence was used to validate the marker genes of

immuneEC and devEC. The result showed that number of cells expressed Sox17 are much larger the number of cells expressed Irf7 (Figures 6F-G) which is reversal compared to the observation in 24 hours post LPS injury. This suggested that major programing in the endothelial cells shifted from immune response to regeneration.

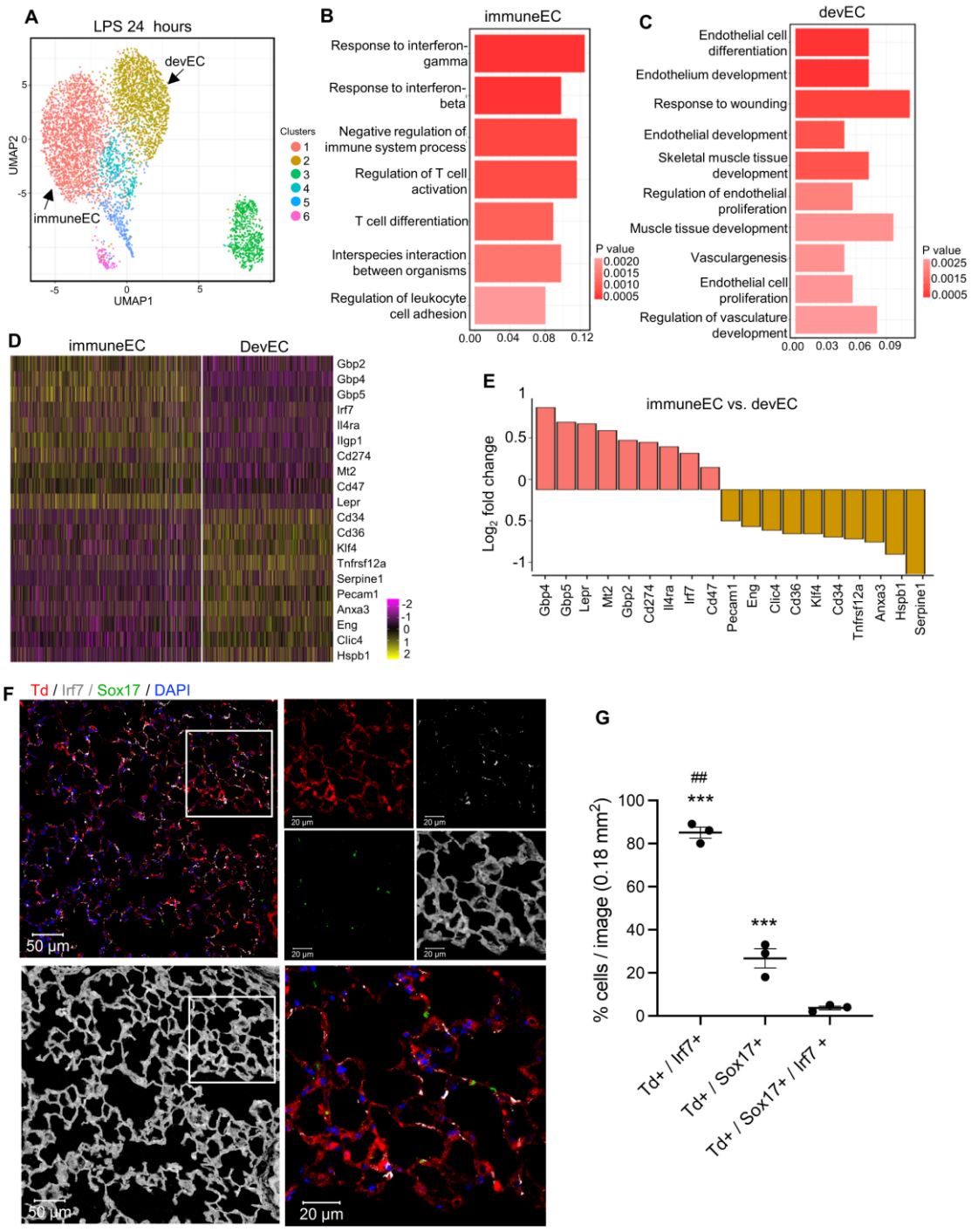

Figure 5. scRNA-seq identified lung EC subpopulations at 24 hours post LPS-injury.

(**A**) UMAP of 5,158 cells at 24 hours post LPS-injury. Cells are colored by the Seurat clusters. (**B**) (**C**) The signature functions of the marker genes in cluster 1 and 2, which have immune response function and developmental function, respectively. (**D**) The heatmap of the most significant maker genes for immuneEC and devEC. (**E**) The log flod-changes between the top differentially expression genes between immuneEC and devEC. (**F**) Immunofluorescence staining results for Irf7 (Grey), Sox17 (Green) and tdTomato (Red). (**G**) Quantification of endothelial cells with different proteins.

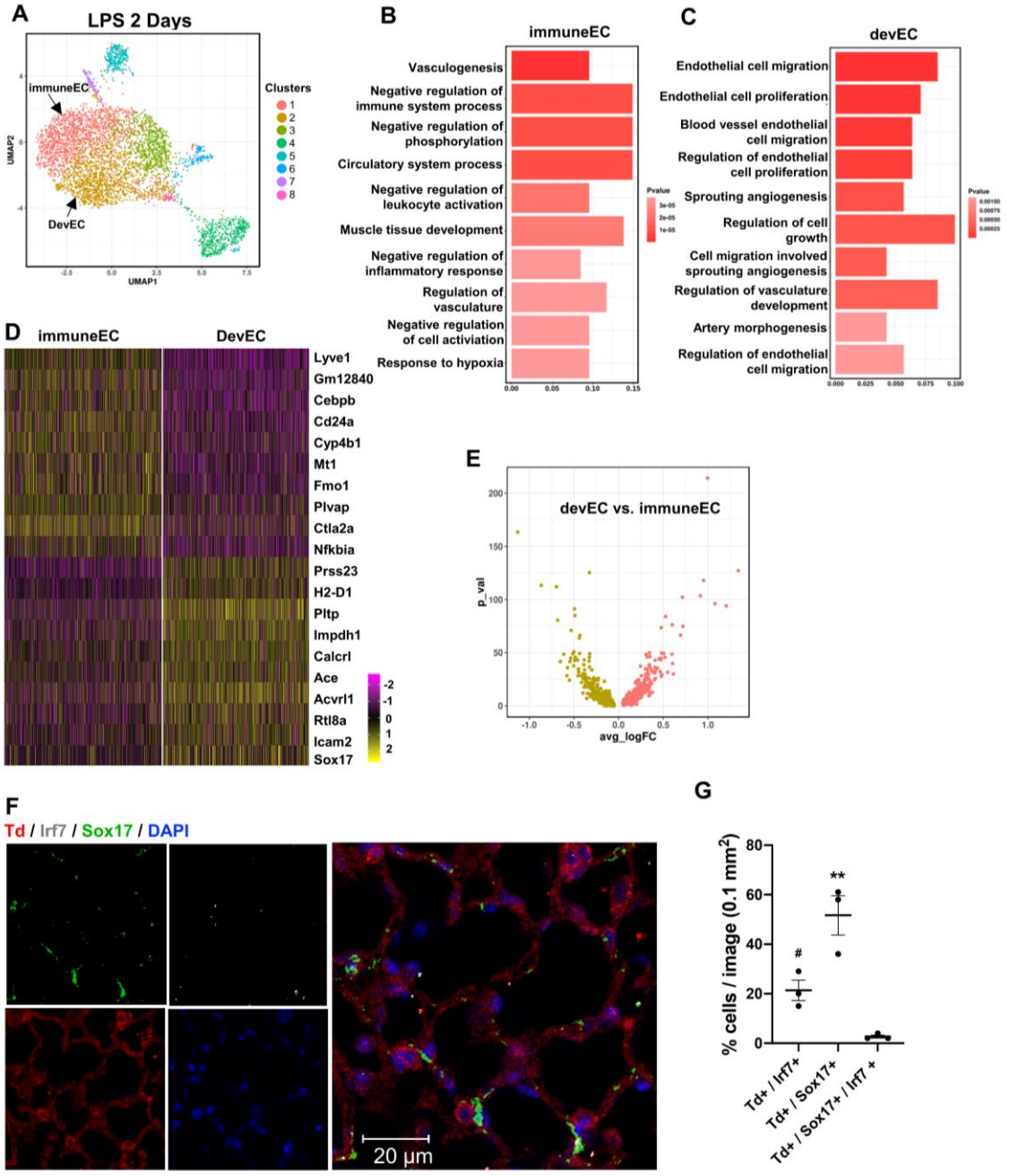

Figure 6. scRNA-seq identified lung EC subpopulations at 48 hours post LPS-injury.

(**A**) UMAP of 4,608 cells at 48 hours post LPS-injury. Cells are colored by the Seurat clusters. (**B**) (**C**) The signature functions of the marker genes in cluster 1 and 2, which have immune response function and developmental function, respectively. (**D**) The heatmap of the most significant maker genes for immuneEC and devEC. (**E**) The volcano plot of genes between immuneEC and devEC. (**F**) Immunofluorescence staining results for Irf7 (Grey), Sox17 (Green) and tdTomato (Red). (**G**) Quantification of endothelial cells with different proteins.

2.3.4 The proliferating EC subpopulation at 3 days post LPS injury

We next analyzed the single cell RNA-seq data at day 3 post-injury. By clustering analysis, we again identified subpopulations in lung ECs (Figure 7A). In the results of GO analysis, we found a novel EC subpopulation that highly expressed proliferative genes. In this subpopulation, the marker genes were enriched for functions of DNA replication, nuclear division and cell cycle (Figure 7B). In Figure 7C, we showed the expression profiles of E2f1, Ccna2, Tk1, Cdk1 and Ccne1, which are key genes related to cell cycle. Based on these findings, we defined this subpopulations as proliferative EC (proEC).

We used RNA-FISH to validate the expression profiles of Ccne1 which was a cell cycle gene and highly expressed in the proliferative EC subpopulation. The RNA-FISH images in lung showed that the number of endothelial cells with Ccne1 expression increased significantly at 3 days post LPS injury (Figure 7D). This result was consistent with our previous study that there was an upregulation of Cyclin E1 at day 3 after vascular injury. Conversely, at other time points, there was no increase in cells expressing Ccne1 based on the RNA-FISH analysis (Figure 7E). This suggested that proliferative EC subpopulations appear at this specific time point during the regeneration after lung injury.

We next sought to identify the origin of the proEC population using Monocle3 transcriptomic trajectory modeling which identifies the most proximate cell populations by generating a pseudo-time trajectory in which cells are ordered according to dynamic changes in their gene expression profiles[90, 91]. The pseudo-time trajectory of proEC at 3 days post LPS and devEC at early timepoints showed that proEC at day 3 gradually traced back to devEC at baseline through devEC of LPS day 2, LPS day 1, LPS 6 hours as temporal ancestors of previous time points (Figure 7F). These findings suggest that

the proEC subpopulation, which emerged at 3 days post LPS, was most likely derived from devEC at baseline.

We performed the same analysis on the single cell RNA-seq data from day 7 post LPS injury where the lung endothelial cells have recovered from the injury[87]. The gene expression profiles of lung endothelial cells at 7 days post LPS injury were analogous to the cells at baseline (Figure 8A-B). There were subpopulations that have inflammatory response functions and subpopulations that have cell development functions. The maker genes of immuneEC and devEC were presented in the violin plots (Figure 9). The high expression of Irf7, Cd47, and Ifit3 in immuneEC suggested that the cells gradually recovered from inflammation, whereas the expression of Sox17, Ace1, and Eng indicated that the ECs at 7 days post injury were returning to the baseline.

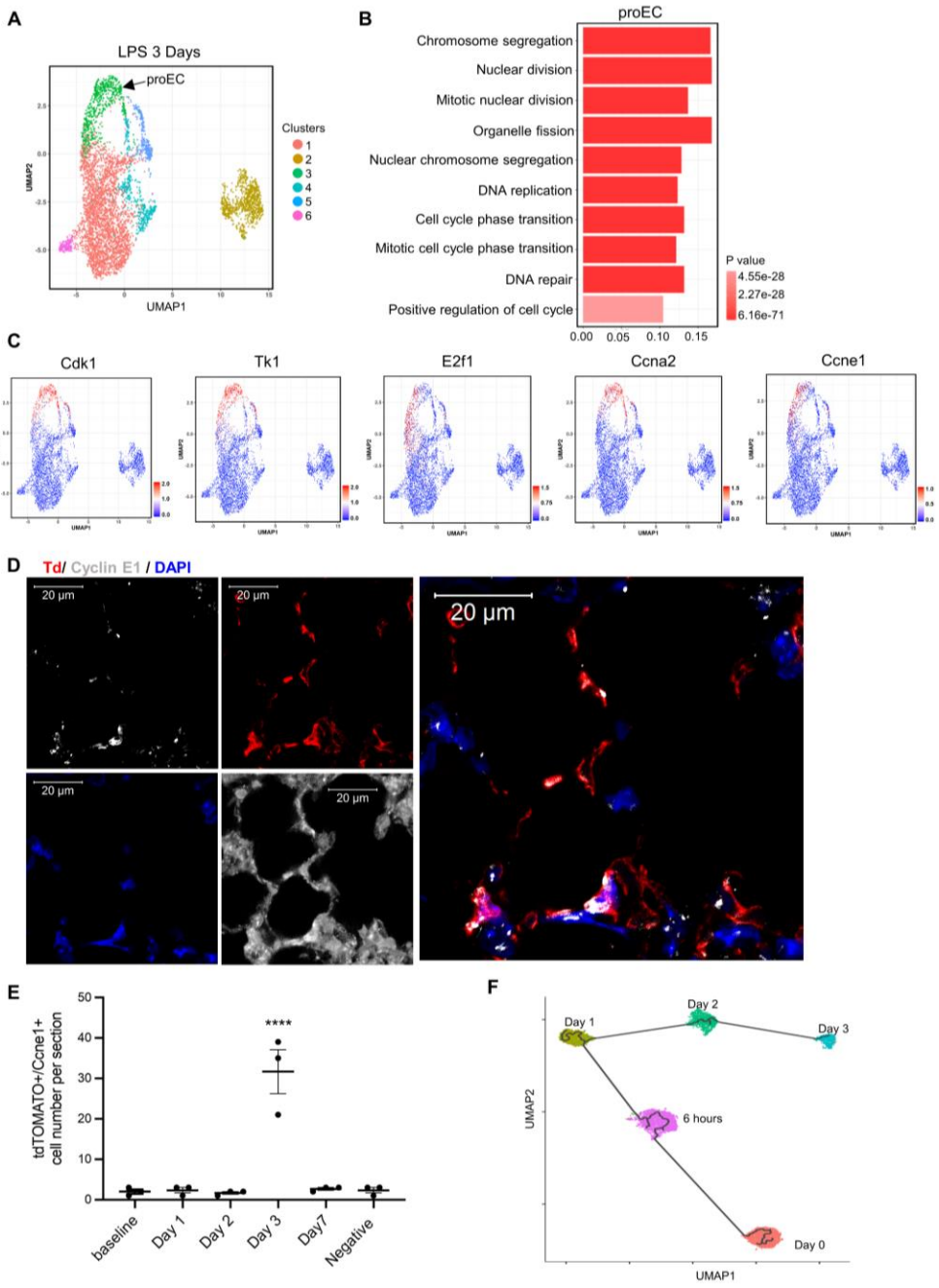

Figure 7. scRNA-seq identified lung EC subpopulations at 72 hours post LPS-injury.

(**A**) UMAP of 5,318 cells at 72 hours post LPS-injury. Cells are colored by the Seurat clusters. (**B**) The signature functions of the marker genes in cluster 3, which have proliferation and regeneration functions. (**C**) UMAP of 5,318 cells at 72 hours post LPS-injury. Cells are colored by scaled the expression of the selected marker genes. (**D**) Immunofluorescence staining results for Ccne1 (Grey) and tdTomato (Red). (**E**) Quantification of endothelial cells have Ccne1 protein in different time points. (**F**) Trajectory analysis between devEC at baseline, 6 hours, 1 day, 2 days, and proEC from 3 days.

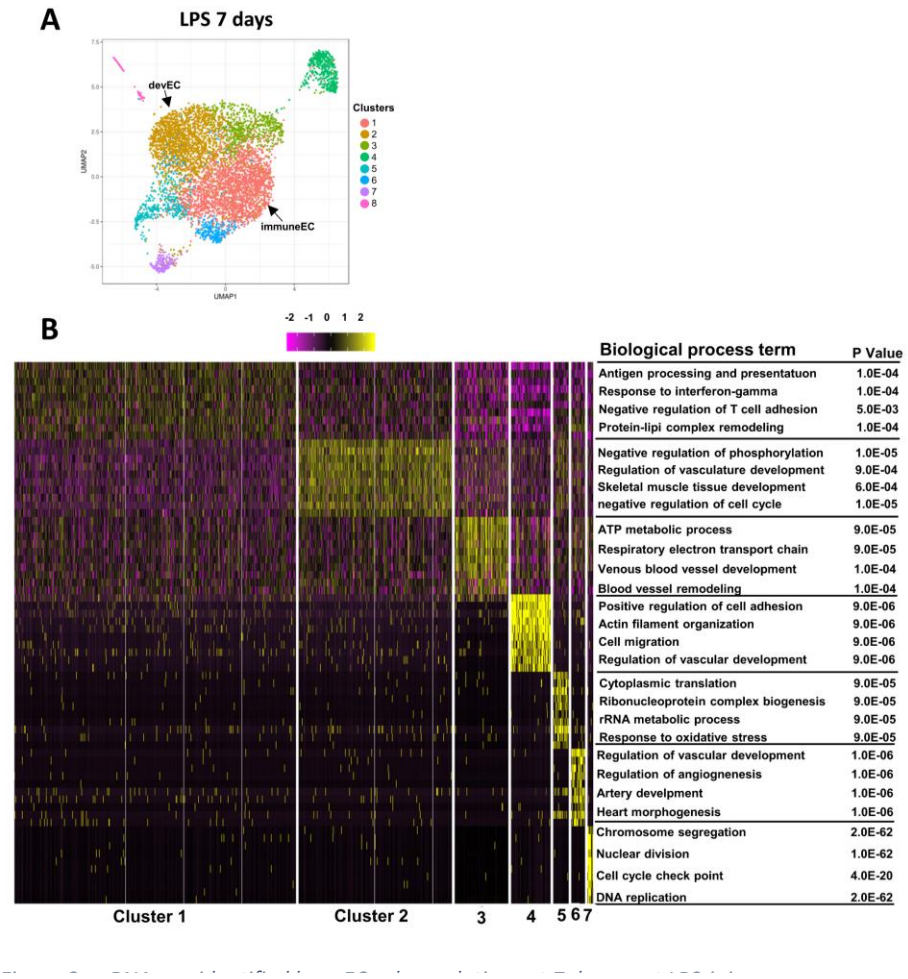

Figure 8. scRNA-seq identified lung EC subpopulations at 7 days post LPS-injury.

(**A**) UMAP of 6,171 cells at 7 days post LPS-injury. Cells are colored by the Seurat clusters. (**B**) The signature functions of the marker genes in all clusters at 7 days post LPS-injury

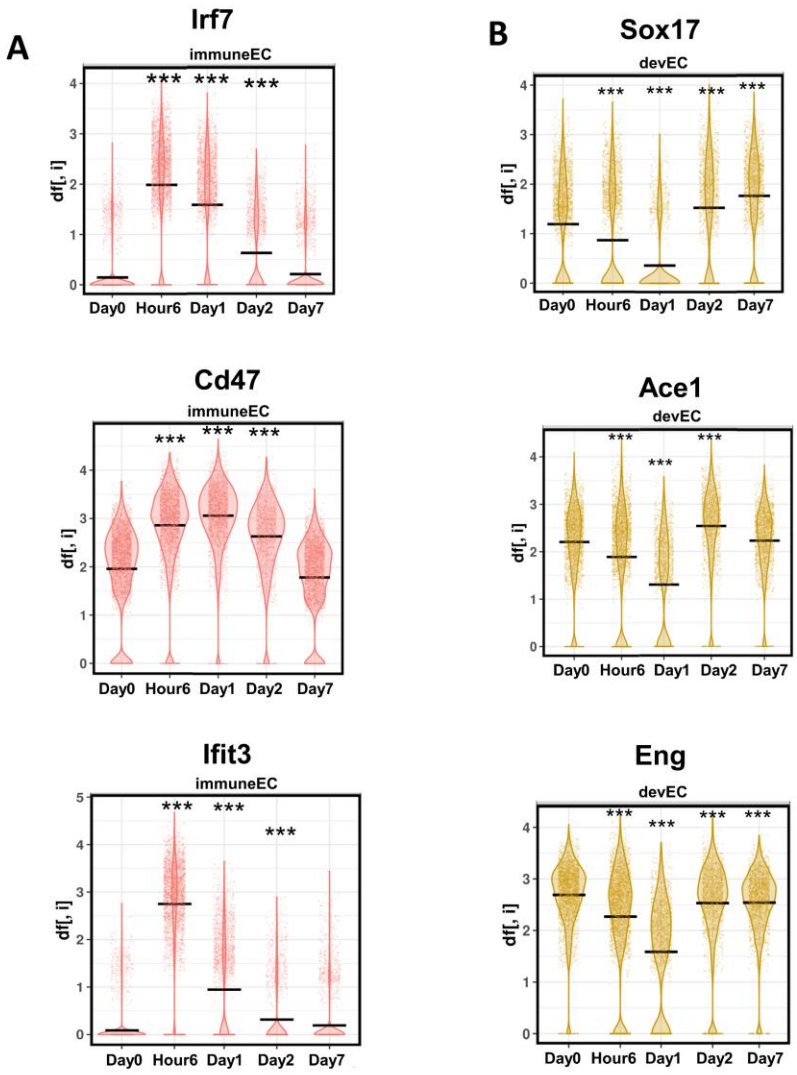

Figure 9. The violin of marker genes in immuneEC and devEC subpopulations from baseline and different time points post LPS-injury.

2.3.5 Sox17 highly expresses in devEC subpopulations in EC across different timepoints

Next, we used SCENIC (single-cell regulatory network inference and clustering) model to infer transcription factor activities in the ECs at each time point. By doing so, we established TF-gene regulatory networks to understand the regulatory mechanisms between the distinct EC subpopulations. SCENIC is a model to learn the transcription factors activities in single cell levels by the expression of transcription factor target genes. They identify the target genes of transcription factors by co-expression and the presenting of transcription factor binding motifs in the promoter regions of genes[89]. At LPS 6 hours, the inferred transcription factor activities in immuneEC and devEC subpopulations corresponded to the functions of these subpopulations that were revealed by GO enrichment analysis. In devEC subpopulation, the activated transcription factors are known to have developmental functions, while the activated transcription factors have immune functions in immuneEC subpopulation (Figure 10A). The activities profiles of key transcription factors were shown in the UMAP and the predicted target genes of Sox17 was shown in Figure 10B. We could see that Sox17 was inferred to be highly activate in the devEC subpopulation. This result corresponded to the findings from Sox17 expression level and demonstrated that Sox17 is an essential factor that suppresses inflammation in developmental endothelial cells.

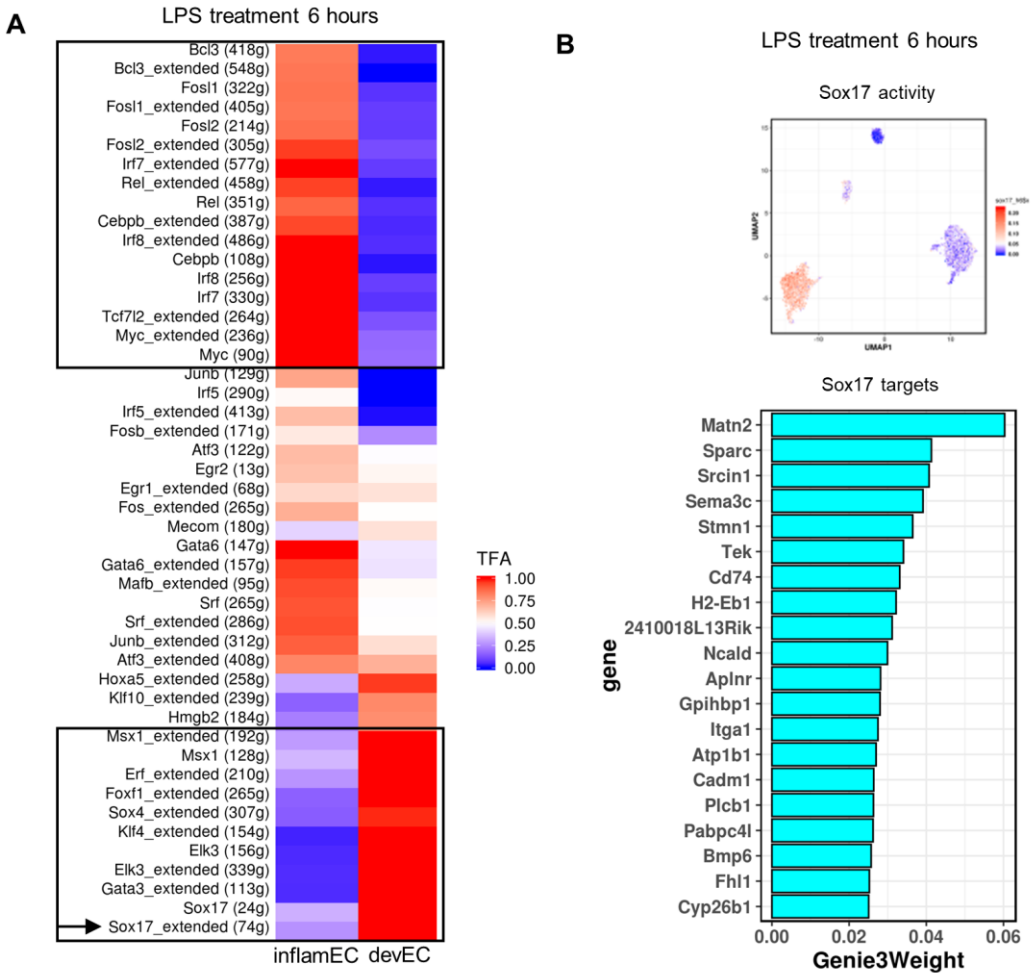

Figure 10. Transcription factors analysis at 6 hours post LPS injury with SCENIC.

(**A**) The inferred transcription factors activities at 6 hours post LPS injury by SCENIC. The heatmap shows the mean activation of transcription factors in distinct subpopulations. (**B**) The inferred activity of Sox17 at 6 hours post LPS injury and the inferred target genes of Sox17 ordered by Genie3Weight.

2.4 Summary

Many studies have shown that there is cellular heterogeneity in endothelial cells (ECs) in blood vessels. However, how distinct endothelial subpopulations are and what function the subpopulations have in the lung remains unknown. In this project, we performed single cell RNA-sequencing of 35,973 endothelial cells in the lung at baseline and multiple time points after LPS injury and regeneration. The R package Seurat was used in the single cell RNA-seq analysis. We identified the clusters in the lung ECs at different time point. And by performing differential expression analysis and GO enrichment analysis on each subpopulation, we found that there are distinct subpopulations with specific functions at each time point. A major subpopulation that highly expressed inflammatory response genes was defined as the immuneEC subpopulations. Another major subpopulation that highly expressed cell developmental genes such as Sox17 was defined as the devEC subpopulation. We also found a subpopulation that expressed proliferation genes in the late stage of LPS injury. We defined it as proEC and they emerged from the devEC in the previous time points based on the trajectory analysis. Therefore, a “division of labor” exists within the lung endothelium. During the lung injury and regeneration, some ECs have inflammatory response propensity while others have regeneration and proliferation functions. These findings lay a foundation for the specific targeting of lung endothelial subpopulations to reduce inflammatory lung damage and enhance lung endothelial regeneration.

Chapter 3 The Bayesian Inference Transcription Factor Activities Model

Adapted from Gao, S., Dai, Y. and Rehman, J. (2021). A Bayesian inference transcription factor activity model for the analysis of single-cell transcriptomes. Genome Res.

3.1 Introduction

Single-cell RNA sequencing (scRNA-seq) is an effective experimental method to identify the heterogeneity of cells in transcriptomics level and provide gene expression profiles for individual cell[102]. It has been used to identify the subpopulations cells in distinct tissues[103] as well as the stages in the development of progenitors and the underlying the regulation of cell fate[104-108]. The heterogeneity of cells within a given tissue is usually identified by evaluating the similarities between cells by their gene expression profiles. Many methods and tools have been developed to examine cell heterogeneity in single cell RNA-seq data analysis[109]. Most existing methods share a common approach. The first step is to transform the single cell gene expression data into a lower dimensional space. By doing this, the major discrepancies between cells on the transcriptome level will be preserved[110-119]. However, in these approaches, the biological context is not considered in the identification of the distinctions between subpopulations of cells in the lower dimensional space[120]. Thus, the cellular heterogeneity that is recognized from this lower dimensional space does not necessarily reveal the subpopulations of cells with distinct biological functions. Moreover, the

existing methods do not provide a direct answer about the functions of the subpopulations or their regulatory mechanisms.

Transcription factors are the key regulators of gene expression. The activity of transcription factors are indicators of cell subpopulation function. However, transcription factor activities are not available by experimental methods. We hypothesized that the transcription factor activity information is encoded in the gene expression dataWe developed a Bayesian hierarchical model that integrates known transcription factor ChIP-seq data with single cell RNA-seq data to infer the transcription factors activities in single cells. The results of inference can be used for downstream analysis such as classifying cell subpopulations as well as identifying the preferred target genes of transcription factors in datasets.

3.2 Methods

3.2.1 The Bayesian Inference Transcription Factor Activity Model (BITFAM)

BITFAM is Bayesian factor analysis model [121] that decomposes the normalized single cell RNA-seq count matrix $Y = WZ + \phi$. BITFAM factorized the normalized scRNA-seq data into two matrices, the transcription factor target gene regulation strengths matrix (W) and the inferred transcription factor activities matrix (Z). Matrix W is the inferred regulatory strength between target genes and transcription factors. Based on the transcription factor ChIP-seq data, different prior distributions are assigned to the elements in matrix W. The prior distributions are normal distributions with distinct variances.

The weights between transcription factor k and the gene n follow a normal distribution as prior distribution. For W_{nk}:

$$W_{nk} \sim \begin{cases} N(W_{nk}|0, 1/\delta_k) & \text{if gene } n \text{ is targeted by transcription factor } k \\ N(W_{nk}|0, 0.001) & \text{otherwise} \end{cases}$$

for transcription factor $k = 1, \dots, K$ and gene $n = 1, \dots, N$. Automatic relevance determination (ARD) is used to model the inverse of variance, δ_k, i. e.,

$$\delta_k \sim Gamma(1e^{-3}, 1e^{-3}).$$

The purpose of using automatic relevance determination (ARD)[122] here is to model the regulatory strength between transcription factors and target genes automatically. Even the relation of them is shown by the ChIP-seq data, it is still possible that the gene is not a target of transcription factor in the analytic dataset. Because the prior knowledge is a combination of multiple ChIP-seq datasets.

Matrix Z is the inferred transcription factor activities for every cell. Beta distribution is assigned to the activity of transcription factor k in cell m ($m = 1, \dots, M$) as the prior distribution, i.e.,

$$Z_{km} \sim Beta(0.5, 0.5).$$

A normal distribution with variance ϵ is used to model the residual noise ϕ_{nm}. $\phi_{nm} \sim N(0, \varepsilon)$; $\varepsilon \sim Gamma(1, 1)$ for gene n.

The likelihood of BITFAM is:

$$Y|(W, Z, \phi) \sim Normal(WZ,\ \phi).$$

3.2.2 Parameter inference

We use variational methods[123, 124] to approximate the posterior distribution in our Bayesian hierarchical model. The variational methods will use distributions from the exponential family to approximate the posterior distributions. The inference procedure will be converted to an optimization problem. By doing so, it achieves a faster running time,

scalability to a large number of samples, and high accuracy. The final matrix W and Z are determined by the average of 300 random samples from the posterior distributions.

An R[125] package Rstan (Version 2.18.2) is used to implement the inference of BITFAM. Automatic Differentiation Variational Inference (ADVI)[126] is used in variational inference in Rstan. ADVI uses Monte Carlo integration to approximate the evidence lower bound (ELBO) which is the variational objective function. The optimization of the ELBO uses Stochastic gradient ascent. When the mean change of ELBO is under 0.01, the algorithm will stop and return the final posterior distributions.

3.2.3 Processing and analysis of the scRNA-seq data sets

The Smart-Seq2 protocol was used to generate The *Tabula Muris* data from mouse lung, heart and brain. There are 5447, 4321 and 6315 in the three organs, respectively. The raw counts were downloaded from Tabula muris Figshare websites. The expressions of well-known marker genes are used to label the cell types in each organ[103]. The MARS-seq platform was used to generate the blood cell development data. There are 10,368 cells in the dataset from NCBI GEO (accession GSE72857). 2,729 cells are myeloid progenitor cells (CMP, GMP and MEP) which are used in this study.

R package Seurat[70] was used to normalize the gene expression data. the normalization method “LogNormalize” was selected. The most variably expressed genes were identified by FindVariableGenes() in Seurat.

3.2.4 The identification of transcription factor target genes

The Gene Transcription Regulation Database (GTRD v19.04)[27] was used as the database to obtain the ChIP-seq data. In the GTRD database, the developer processed the ChIP-seq data from human and mouse in the same pipeline. They provide the transcription factor binding information in different cell types and conditions. The meta cluster intervals integrated the TF binding signal peak from different experiment and peak calling methods for each TF. Then we defined the target genes of transcription factor as the genes that overlap with transcription factor ChIP-seq peak intervals and gene promoter. For default settings, we use 2000 bps upstream and 200 bps downstream of the TSS as the gene promoter region. The location of the TSS was obtained from UCSC Table Browser [127].

In each dataset, we filtered the transcription factors by the expression. We only looked at the transcription factors that are expressed in the cells. If the number of overlaps between the most variably expressed genes and known target genes of the transcription factor was less than 10, we removed that transcription factor from the inference.

3.2.5 Identification of the cell type specific transcription factors by random forest model

We built a random forest model to find the signature TFs for each cell types or clusters in single cell datasets. To train the random forest model, we labeled the cells with 1 and 0. If the cell belongs to the cell type or cluster we are interested in, it is labeled as 1, otherwise it is labeled as 0. In the training of the random forest model, we used the inferred transcription factors activities as the features. We investigated the most

significant TFs for each cell type by ranking the feature importance score. The R package randomForest is used for the training and feature selection.

3.2.6 Louvain's algorithm

Louvain's algorithm is a clustering method based on graph. It is developed to detect communities by maximizing a community modularity score. The modularity refers to the density of nodes within one community[128]. In order to build a graph to apply Louvain's algorithm, we use R package destiny [129] to construct a fully connected graph that all cells are connected. The edge weights between cells are the diffusion distances on the inferred transcription factor activity profiles. Then we only keep the edges that are among the top 20% minimum weight edges. We applied Louvain's algorithm on this new graph. The R package igraph is used for graph building and implementation of Louvain's algorithm.

3.2.7 Metrics for evaluation of clustering

3.2.7.1 Rand index (RI) and Adjusted Rand index (ARI)

The Rand Index (RI) is defined as

$$RI = \frac{a + b}{\binom{n}{2}}$$

where a is how many times that one element belongs to the same group under two different classifications and b is how many times that one element is in different groups under two classifications [130]. The RI, which ranges from 0 to 1, is the frequency of same classification across all elements. The Rand Index is 1, when the two classifications are the same.

The ARI is defined as

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}$$

where n_{ij} is the number of the entry i, j in contingency matrix. a_i is the sum of the contingency matrix row i, b_j is the sum of contingency matrix column j. The RI is corrected by ARI for chance[130].

3.2.7.1 Normalized Mutual Information (NMI)

The Normalized Mutual Information (NMI) is defined as

$$NMI = \frac{2 \times I(X;Y)}{[H(X) + H(Y)]}$$

where X is the classification of all elements, Y is the clustering result. $H(.)$ is the entropy of classifications, $H(.) = \sum_i (-p_i)\log(p_i)$, p_i is the probability of one element is grouped into cluster i. $I(X;Y)$ is the mutual information defined as $I(X;Y) = H(X) - H(X|Y)$.

3.2.8 Jaccard index

The Jaccard index is a value which is used to measure how similar two groups are. $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, where A and B are different groups.

3.2.9 Diffusion pseudo time (DPT)

Diffusion distances between cells were used to assign Diffusion pseudo-time (DPT) to every cell. The profiles of reduced dimensions, such as PCA on gene expression or inferred transcription factor activities from BITFAM, were used to compute the Diffusion distances. We set one of CMP cell as starting point and two cells from GMP and MEP were treated as endpoints of the development branches. Pseudo-time was scaled to [0,

1] and [0, -1] respectively to vasulize the trajectory more clearly. Diffusion pseudo-time (DPT) was calculated by the R package destiny[129].

3.2.10 Other existing tools

SC3, SIMLR and Seurat were available to download at Bioconductor (Version 3.9). Default parameters were used to run the Seurat. For the clustering, we use the top 20 principal components as the input. Default parameters were used to run SIMLR and the function SMLR::SIMLR_Estimate_Number_of_Clusters was used to detect the number of clusters. Default parameters were used to run SC3 and the function SC3::sc3_estimate_k was used to detect the number of clusters. Default parameters were used in Monocle3.

3.2.11 Gene Ontology enrichment analysis

We used the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8[131] to do Gene Ontology enrichment analysis on the top weighted target genes of transcription factors. The Biological Process (BP) (level 5) was applied as the gene annotation terms to do the enrichment analysis. The GO terms with adjust p-value < 0.005 were identified the signature functions of the genes.

3.2.12 Benchmarking with AUROC on CRISPRi data

The CRISPRi data is available on GEO database (accession GSE127202). It has 141 perturbation experiments with 50 transcription factors deletion. We applied BITFAM on the CRISPRi dataset and computed the logFC of inferred transcription factor activities between control cells and perturbed cells. For example, transcription factor X is one of the targets of CRISPR, we computed the logFC of inferred activities between cells X was perturbed and cells X was not perturbed. We also labeled the cells with 1 and 0 by the

biological ground truth that is known from the CRISPRi experiment. The perturbed cells are positive class with label 1, and the non-perturbed cells are negative class with label 0. We could calculate he AUROC by changing the threshold of logFC to define in one cell whether transcription factor X is inferred as deletion. Using this approach, we could compute an overall AUROC to evaluate the inference performance. The R package yardstick (version 0.0.3) was used to perform the AUROC analysis.

3.2.13 Other transcription factor activities methods on CRISPRi data

To identify the co-expression between target genes and transcription factors, we selected the top 500 largest absolute correlation values. To identify the transcription factor binding motifs in the gene promoters, regions 2000bps upstream and 200bps downstream around the TSS were selected as promoter regions. Then FIMO [132] was used to search known transcription factor motifs on these promoter regions. The transcription factor motifs were obtained from HOCOMOCO[133] and JASPAR[134]. The genes with binding motifs (P-value < 0.0001) were used as transcription factors target genes. To identify transcription factor target genes with ChIP-eat method, which integrate the ChIP-seq peaks and computational models for the prediction of transcription factor binding, we used the BED files downloaded from UniBind website at http://unibind.uio.no for every transcription factor.

3.2.14 Software Availability

BITFAM (version 1.2.0) is implemented based on R [125]. The BITFAM source code and package are publicly available at GitHub: https://github.com/jaleesr/BITFAM.

3.3 Results

3.3.1 BITFAM Overview

Bayesian Inference Transcription Factor Activity Model (BITFAM) is developed to infer transcription factor activity by integrating scRNA-seq data and existing biological data on transcription factor binding sites. In the model, we obtained the potential target genes of transcription factors from a publicly available database GTRD. This database has curated 17485 transcription factors ChIP-seq experiments and provide the target genes of transcription factors by overlapping the ChIP-seq signal and the promoter of the genes[27]. A Bayesian factor analysis model is applied to the normalized single cell RNA-seq counts table to do factorization under the guide of prior knowledge from known ChIP-seq data transcription factor target genes[135]. Bayesian factor analysis has been used to reveal heterogeneity from transcriptome data by integrating with gene functions and pathways as factors[136-138]. BITFAM requests a log-transformed normalized gene expression matrix (Y) as input and decompose the matrix into two matrices W and Z. For matrix Y, the rows correspond to genes (G, number of genes) and the columns correspond to each cell (M, number of cells). The transcription factor activities are generated from single cell RNA-seq data in the matrix Z. The inferred factor loadings are obtained in the matrix W. Unobserved random noise term is modeled with Φ.

In the motivation of BITFAM, we wanted to do a biological meaningful decomposition on the normalized gene expression matrix Y (Figure 11). W and Z should represent specific biological meaning. In order to achieve this goal, we integrated the transcription factor ChIP-seq data and generated prior knowledge about the relations between transcription factors and their target genes. This information is a binary matrix

with the same size of W. The columns of it correspond to transcription factors and the rows represent genes. If the gene is the known target of one transcription factor in the GTRD ChIP-seq database, the corresponding entry in the prior knowledge matrix is 1. Otherwise, the entry is 0. Then we incorporated this binary matrix in the inference. The known relations between genes and transcription factors determine the prior distributions of elements in W in the Bayesian inference model. By doing so, we could infer the transcription factor and gene regulatory strength in the specific single cell RNA-seq data even the ChIP-seq experiments are from cell types which are very different.

The matrix $\mathbb{Z}$ are the inferred transcription factor activities in the single cell RNA-seq data. BITFAM can be used as a dimension reduction approach as the total number of genes is much larger than the number of TFs. The matrix $\mathbb{W}$ are as the regulatory weights between the target genes and transcription factors. The matrices $\mathbb{Z}$ and $\mathbb{W}$ could be applied on different downstream analyses, such as 1) identifying the transcription factors activities in each single cell to uncover regulatory mechanisms, 2) revealing the preferred target genes of transcription factors in the biological context where the single cell RNA-seq is generated, and 3) clustering and building trajectory on the cells with inferred transcription factors activities.

In order to evaluate and validate BITFAM, we applied it on several single cell RNA-seq data. The Tabula Muris datasets contain mouse single cell RNA-seq data from all major organs under health state. Each organ consists of multiple cell types[103]. A blood cell development single cell RNA-seq dataset has two differentiation trajectories from common myeloid progenitors (CMP) to granulocyte-macrophage progenitors (GMP) or megakaryocyte–erythroid progenitor (MEP)[139]. A CRISPRi single cell RNA-seq data has CRISPR-deletions on 50 transcription factors [140]. The properties of these datasets are

shown in the Table 2. In the first two datasets, cells are labeled with cell types by the expression of the cell type specific marker genes. This label could be treated as knowledge to evaluate the functional relevance of the inference results from BITFAM. In the CRISPRi dataset, if one transcription factor is deleted or knocked down in a cell, the activity of the corresponding transcription factor should be zero. Therefore, the accuracy of the inferred transcription factor activities from BITFAM can be validated with the CRISPRi dataset which can be treated as biological ground truth. We are using a transcription factor ChIP-seq database GTRD to obtain the transcription factor and target gene information. GTRD has comprehensive ChIP-seq data identified from the experiments for human and mouse[27] (http://gtrd.biouml.org).

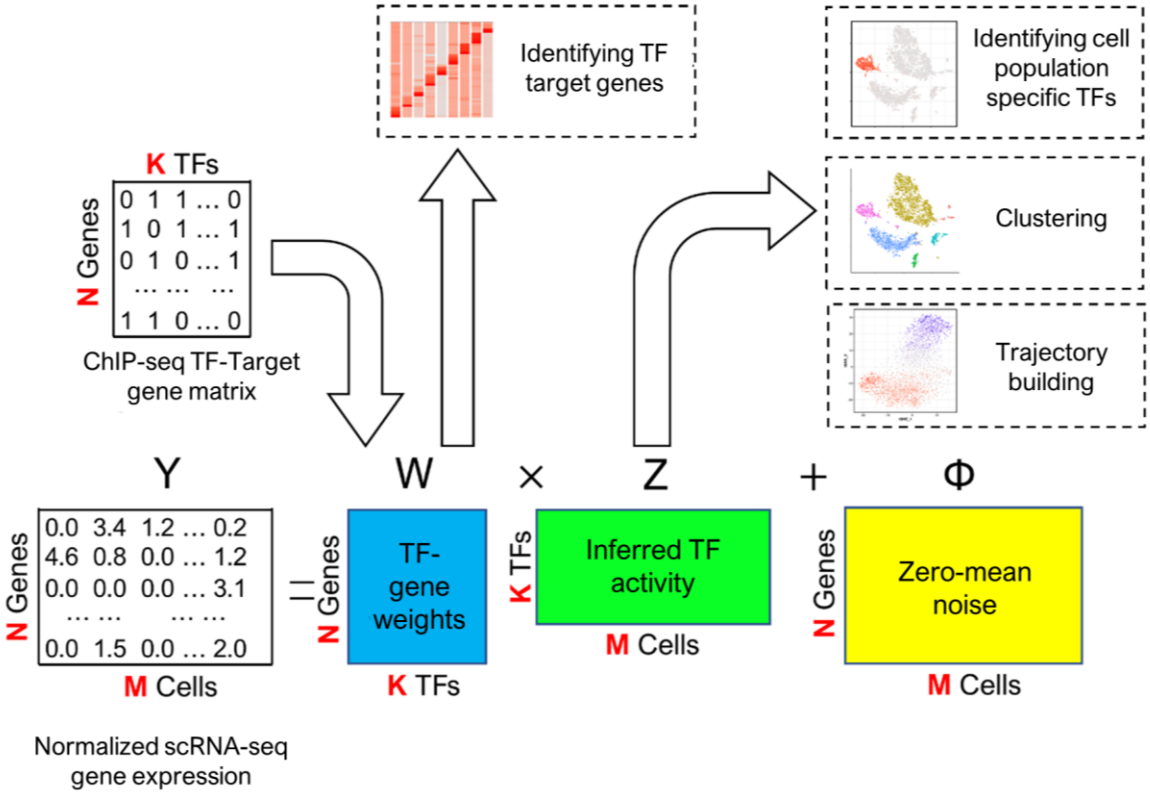

Figure 11. The schema of the Bayesian Inference Transcription Factor Activity model (BITFAM).

3.3.2 The inferred transcription factor activities conform with known biological functions of cells

We applied BITFAM on two single cell RNA-seq datasets: The Tabula Muris lung dataset[103] and the blood cell development dataset[139]. The Tabula Muris lung data represents a discrete biological scenario. The blood cell development dataset represents a continuous biological scenario. We examined whether BITFAM could infer the biologically meaningful transcription factor activities for each cell types.

There are 16 distinct cell types including endothelial cells, stromal cells, B cells, T cells and macrophages in Tabula Muris dataset. The overall similarities between each cell type on gene expression profiles are visualized in a t-SNE plot colored with cell types (Figure 12 A). Cells are labeled with cell types by the expression of the cell type specific marker genes. There are three criteria for choosing the transcription factors in the inference. First, the transcription factor must be variably expressed across all cells. Second, the transcription factor must have at least one ChIP-seq experiment in the GTRD database. Third, in the most variably expressed genes, the transcription factor must have at least 10 target genes. After selection with these criteria, the Tabula Muris lung dataset has 106 transcription factors in the inference with BITFAM (Figure 13). This is just for the default settings. If there are other transcription factors of interest, we could add these transcription factors into the learning. The number of transcription factors did not significantly influence the downstream analyses, such as the visualization of cells with inferred transcription factor activities. This indicates robustness of the model when learning different numbers of transcription factors (Figure 14).

The inferred transcription factor activities can be used in multiple downstream analyses. One of the applications of inferring transcription factor activities is to identify the specific transcription factor for each cell type. We use random forest models for classification of every cell type with the inferred transcription factor activities from BITFAM. In the prediction target of the random forest model, we label the cell with 1 and 0 based on whether the cell is from a specific cell type. By exploring the important features in the random forest model, we could identify the specific transcription factors for each cell type. We repeat this approach for each cell type and get the landscape of relevant transcription factors across all cells. Other methods, such as ROAUC and statistical testing, can also be used to identify the transcription factors for each cell type. The inferred transcription factor activities in the Tabula Muris lung dataset are shown in the heatmap (Figure 12B). In biologically labeled cell types, a clear profile of transcription factor activation emerges.

In the BITFAM results from the Tabula Muris lung dataset, TAL1 is highly activated in endothelial cells (Figure 12C). PAX5 and EBF1 were the transcription factors that were highly activated in B cells. MAFB was the transcription factor that was highly activated in alveolar macrophages (Figure 12D, E). However, only 25% of the endothelial cells had detectable Tal1 mRNA in the single cell RNA-seq data. In B cells and macrophages, the mRNA levels of Pax5 and Mafb were lower. These data suggest that the inferred transcription factors activities do not necessarily match the expression of transcription factors. The limitation of single cell RNA sequencing coverage may not fully detect the mRNAs of important transcription factors in single cells. And at the same time, this might also indicate that BITFAM can infer the activity of transcription factors that undergo post-translational modification. The inference results from BITFAM correspond to biological

knowledge. TAL1 activates the endothelial-specific gene enhancers with important E-box binding elements [141]. It is an essential factor to induce the expression of endothelial genes in vivo[142]. Pax5 encodes the B cell specific activator protein (BSAP). It is a specific regulator of the B cell [143]. MAFB is a key regulator of macrophages[144].

In the BITFAM result from the blood cell development dataset (Figure 12I), the inferred activity of GATA1 shows that it is highly activated in Megakaryocyte–erythroid progenitors (MEPs) (Figure 12J). The inferred activity of CEBPA and STAT5A shows that they are highly activated in Granulocyte-macrophage progenitors (GMPs) (Figure 12K, L). The inferred activities of Gata1 and Cebpa corresponded to the mRNA level of Gata1 and Cebpa in the single cell RNA-seq data. However, the mRNA levels of Stat5a did not match the inferred activity. In the procedure of blood cell development and differentiation, GATA1 is an important factor in erythroid cell development[145]. CEBPA and STAT5A are essential factors in granulocyte-macrophage progenitor development[146, 147]. These results demonstrated that the inferred transcription factor activities from BITFAM match with their known biological functions.

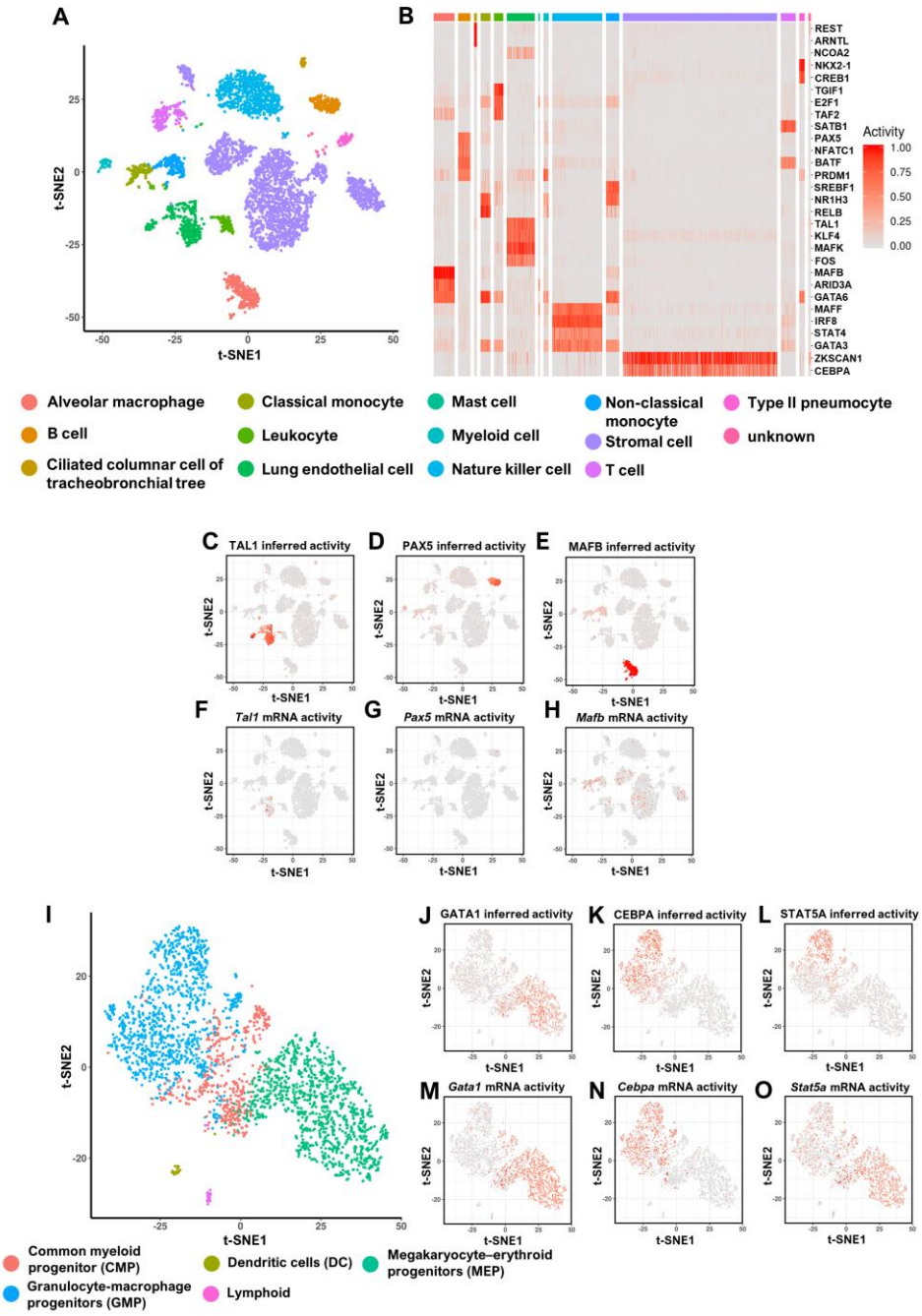

Figure 12. BITFAM inferred the transcription factors activities in individual cell in different dataset.

(**A**) The Tabula Muris lung scRNA-seq data t-SNE plot. (**B**) The heatmap of inferred transcription factors activities that are specifically activated in different cell types. (**C**)(**D**)(**E**) Inferred transcription factors activities of TAL1, PAX5, MAFB. (**F**)(**G**)(**H**) mRNA levels of TAL1, PAX5, MAFB. (**I**) The blood cell development data t-SNE plot. (**J**)(**K**)(**L**) Inferred transcription factors activities of GATA1, CEBPA, STAT5A. (**F**)(**G**)(**H**) mRNA levels of GATA1, CEBPA, STAT5A.

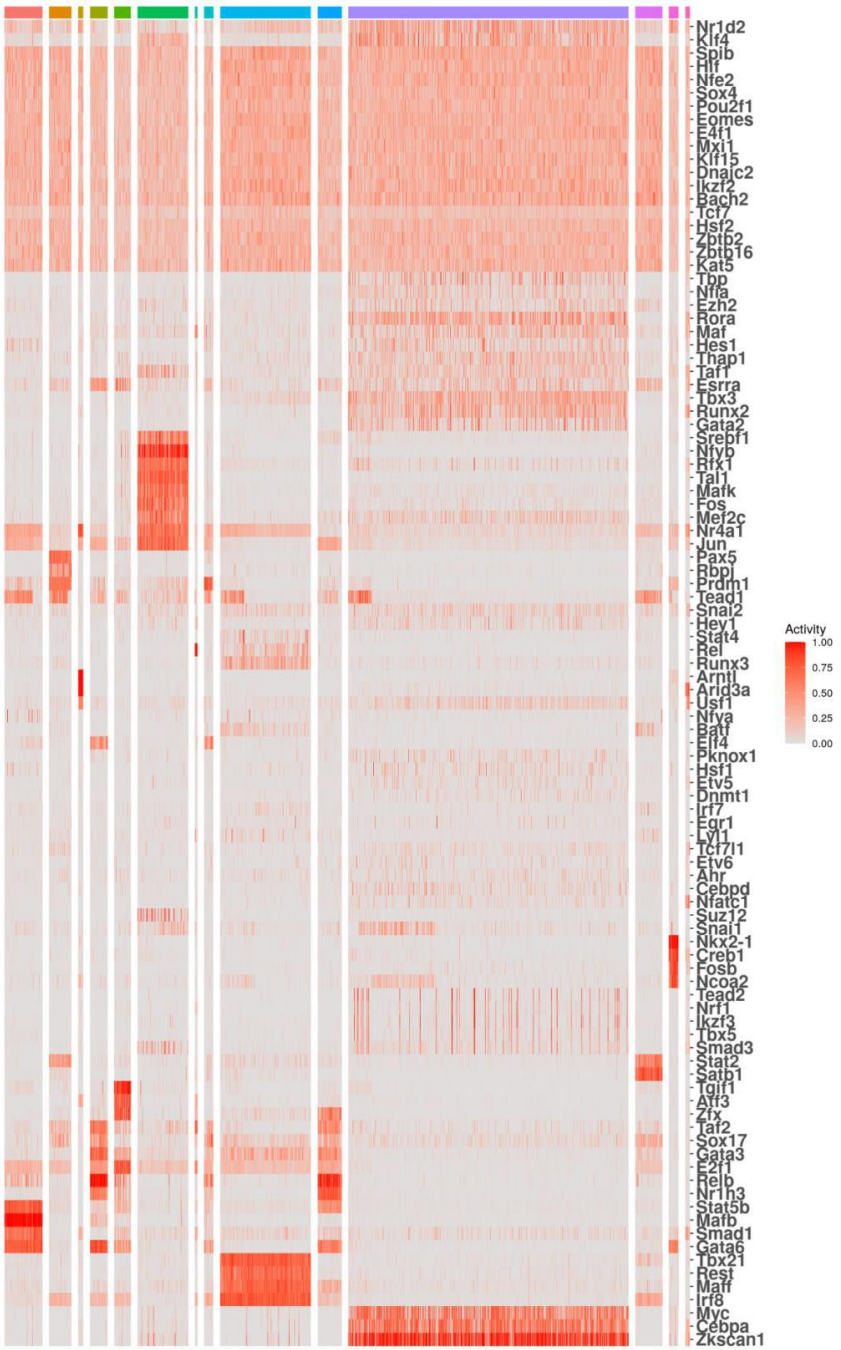

Figure 13. The heatmap of inferred activities of all transcription factors in lung dataset.

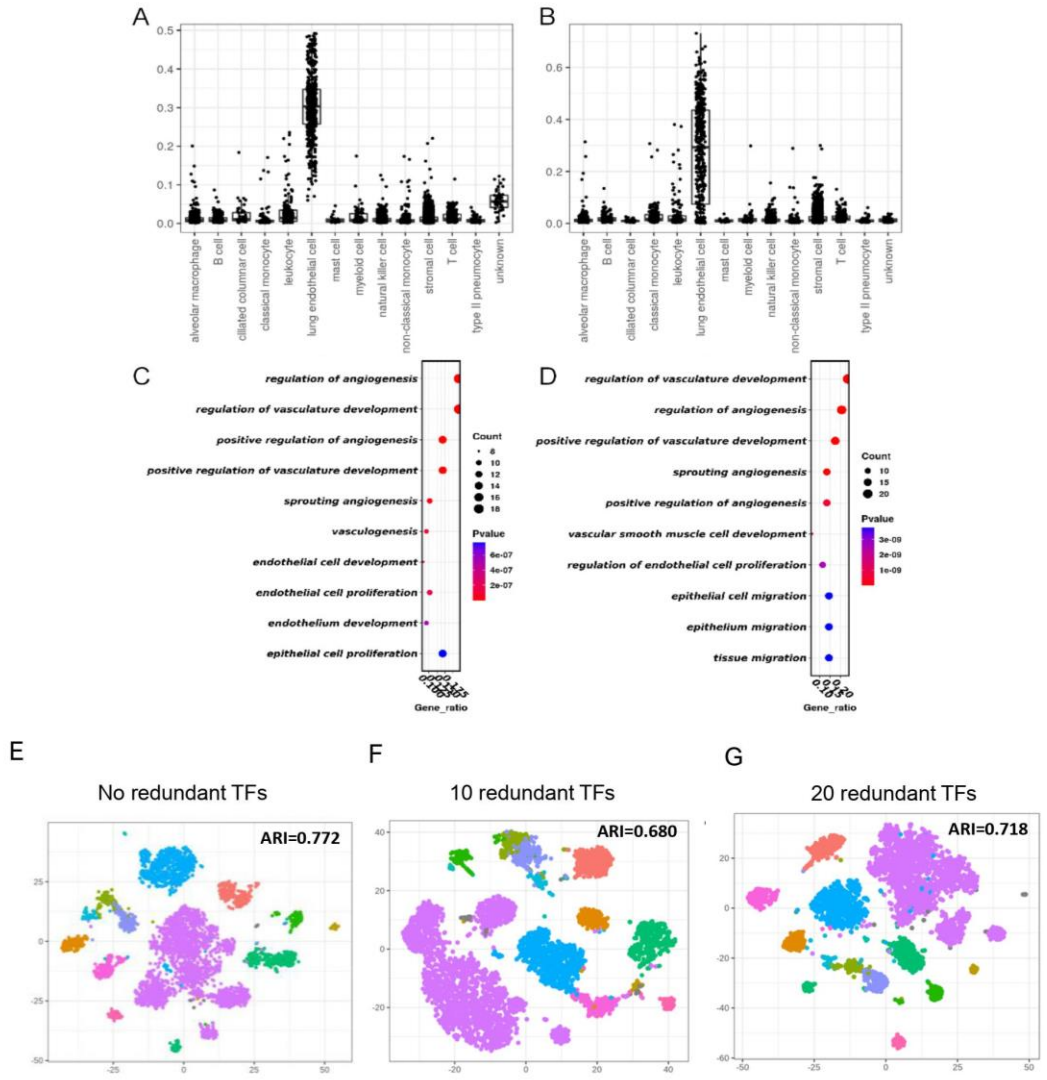

Figure 14. The validation of model robustness in Tabula Muris lung dataset.

(**A**) (**B**) The boxplot of inferred activities of TAL1. (**C**) (**D**) The signature functions of inferred top weighted genes of TAL1. (**E**) (**F**) (**G**) The t-SNE plot of inferred transcription factor activities with no redundant TFs, 10 redundant TFs and 20 redundant TFs.

3.3.3 BITFAM identifies the preferred target genes of transcription in the single cell data.

Next, we explore the biological significance of the inferred transcription factors and gene weights matrix W. In the GTRD database, they integrate the transcription factor binding information from multiple biological conditions and cell types. So, for some transcription factors, they may have thousands of potential target genes. For any given cell type, it is very likely that one transcription factor only regulates a small part of the potential target genes. And the regulatory strength of one transcription factor and different target genes might vary. BITFAM could infer the transcription factor and target gene relations in the given single cell RNA-seq data. A rank of the preferred target genes for given transcription factors can be identified by BITFAM. The inferred regulatory strength is generated based on the mean of the posterior variational distribution of W (Figure 15A).

In the Tabula Muris lung dataset, We checked the top-weighted target genes inferred by BITFAM and picked the transcription factors that were learned to be highly activated in endothelial cells, alveolar macrophages and B cells (Figure 15B). The top 100 genes with the highest positive weights were used in Gene Ontology (GO) enrichment analysis, and the results were contrasted with those of the GO analysis with all the variably expressed experimental target genes. For instance, common biological functions such as protein modification and nucleic acid metabolism are the functions that are enriched by the TAL1 experimental target genes (Figure 15C). The TAL1 top-weighted target genes, on the other hand, were enriched in the endothelial cell specific functions like angiogenesis and vascular development (Figure 15D). Clec14a, a recently

identified important regulator of blood vessel development[148], is one of the gene among top 20 weighted genes of TAL1 (Figure 15E). This is in line with the biologically proven function of TAL1 in vascular development[142]. The enrichment analysis of all PAX5 experimental target genes revealed that they were enriched for general biological processes like nucleic acid metabolism and macromolecule modification (Figure 15F). This result did not match the knowledge about the transcription factor PAX5 which has high level of activity in B cells[143]. But when we used the most weighted PAX5 inferred target genes, we discovered that these genes have functions about B cell-related immune functions and B cell activation (Figure 15G). This result revealed crucial insights about the biological function of PAX5 which had been established by Pax5 deletion studies, independently and experimentally[149]. As shown in Figure 15, the top 20 preferred target genes of PAX5 are crucial genes for B cell function. Alternatively, we performed the same enrichment analysis on the top 100 expressed experimental target genes and 100 randomly selected experimental target genes. The analysis found that the pathways were involved in cellular function in general and did no pathways were specific to B cells. This result indicates that the top weighted target genes learned by BITFAM may offer more functional insights than a purely random selection of experimental potential target genes (Figure 16).

These findings proved BITFAM's capacity to identify transcription factor target genes that are particular to the biological activities of those transcription factors, and they would infer the preferred target genes of transcription factors in a specific dataset.

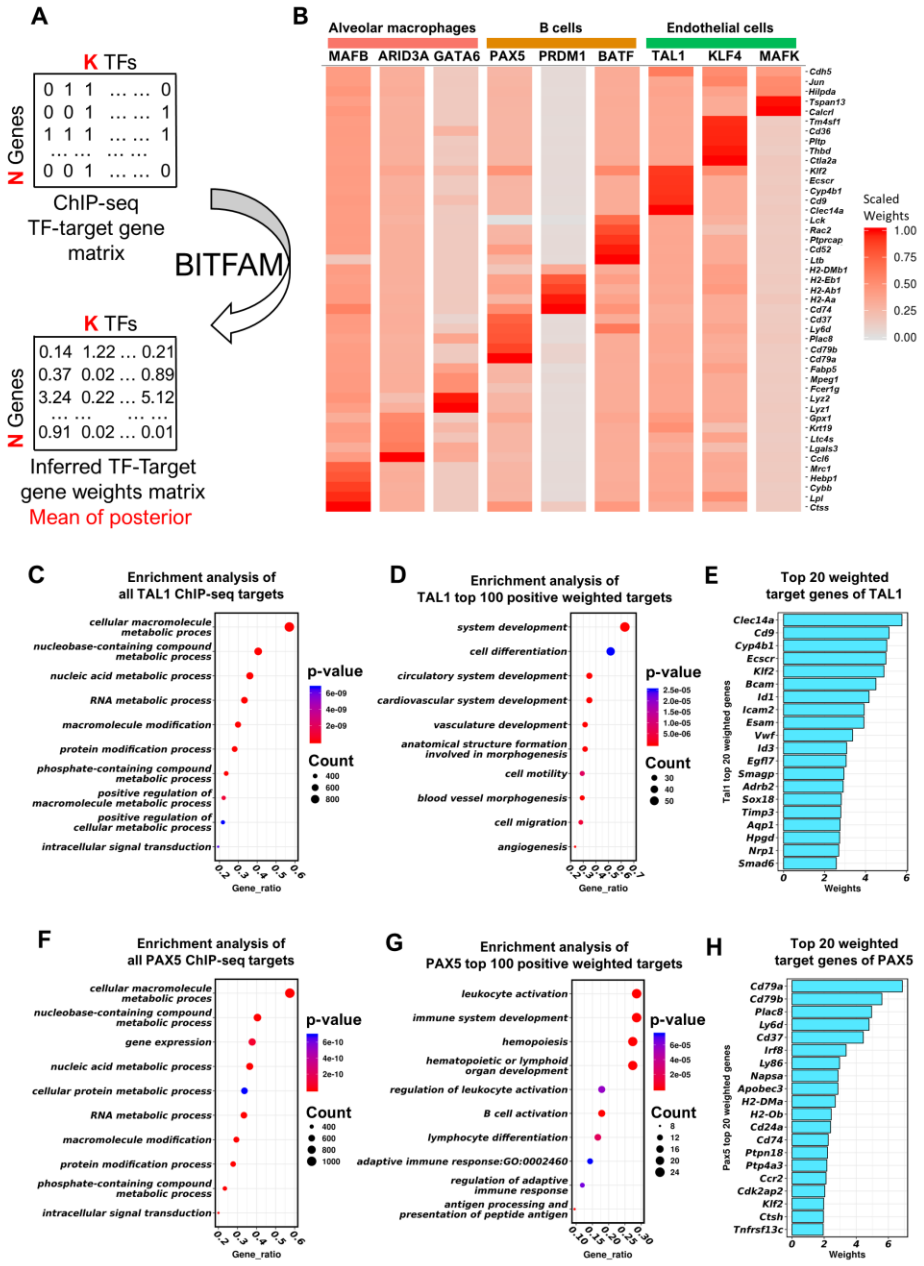

Figure 15. The TF-gene weights inferred by BITFAM.

(**A**)The weights inferred by BITFAM are learned based on prior knowledge of ChIP-seq data. (**B**) The heatmap of TF-genes weights for selected transcription factors for specific cell types. (**C**) The top signature functions of TAL1 ChIP-seq target genes. (**D**) The top signature functions of TAL1 inferred target genes by BITFAM. (**E**) The top weighted TAL1 target genes. (**F**) The top signature functions of PAX5 ChIP-seq target genes. (**G**) The top signature functions of PAX5inferred target genes by BITFAM. (**H**) The top 20 weighted target genes of PAX5.

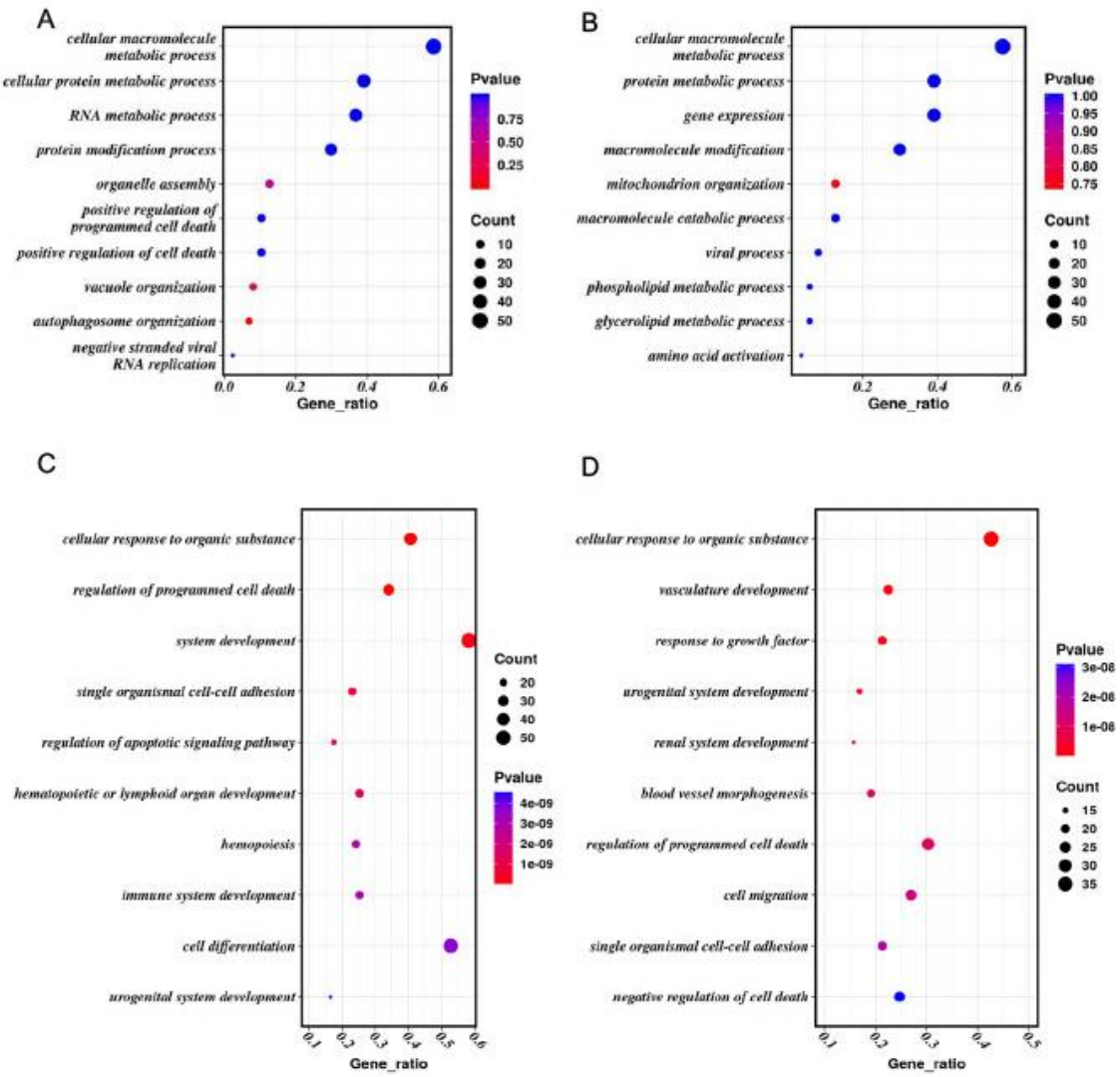

Figure 16. The top signature functions of TAL1 ChIP-seq target genes.

(**A**) (**B**) The top signature functions of TAL1 of 100 randomly selected target genes. (**C**) (**D**) The top signature functions of TAL1 of 100 mosted expressed target genes.

3.3.4 Downstream analysis with the inferred transcription factor activity profiles by BITFAM

Next, we looked at how downstream analyses may make use of inferred transcription factor activities and how activity patterns would indicate heterogeneity between different subpopulations. In order to evaluate the distances between cells with the inferred transcription factor activities and identify cell subpopulations, Louvain's algorithm is applied on the inferred transcription factor activities. We found six cell clusters in the Tabula Muris heart data (Figure 17A) and compared to physiologically distinct cell subpopulations such as endothelial cells, fibroblasts, endocardial cells and cardiac muscle cells (Figure 17B). We found that the clusters grouping by the inferred transcription factor activities overlapped with the cell types that are labeled by the expression of the cell type marker genes. This showed that the inferred transcription factor activity does indeed represent different biological roles of cell subpopulations. For each cluster identified by inferred transcription factors activities, we further investigate the most active transcription factors with the random forest model mentioned before. TAL1 was one of the inferred signature TFs in group 3 and 5 (Figure 17A, Figure 18A). IRF8 was one of the inferred signature transcription factors in group 6 (Figure 17A, Figure 18B). These findings agreed with the cell types (specified by expression of cell type marker genes), which showed that the cells in clusters 3, 5 and 6 were endothelial cells, endocardial cells, and leukocytes respectively.

On the Tabula Muris brain data set, the same analysis was done. 8 clusters were found, and their biologically defined cell labels were compared to the clustering result (Figure 17C, D). The clusters identified by inferred transcription factor activities again

overlapped with the biological labeled cell types. NEUROD1 was inferred to be highly activated in cluster 2 (Figure 18C). ASCL1 was one of the signature transcriptions factors specific to cluster 3 (Figure 17C, Figure 18D). The cell types indicated that the cells in cluster 3 were astrocytes and the cells in cluster 2 were neurons, and these results perfectly matched the biological cell labels (Figure 17C, D). According to these results, BITFAM creates biologically meaningful heterogeneity for individual cells and recognizes cell heterogeneity by various inferred transcription factor activity.

We examined the inferred transcription factor activities discovered from the single cell RNA-seq data investigation of hematopoietic differentiation to see if they can be applied for the visualization and trajectory construction of continuous biological scenario. The differentiation trajectories of common myeloid progenitors (CMPs) towards either megakaryocyte-erythroid progenitors (MEPs) or granulocyte-macrophage progenitors (GMPs) were visible when the estimated transcription factor activity were visualized using UMAP (Figure 17E). In order to create a pseudo-time order and create a differentiation trajectory, we also used the conventional diffusion pseudo-time (DPT) technique on the inferred transcription factor activity (Haghverdi et al. 2015; Angerer et al. 2016). The BITFAM-DPT technique allocated the cells to two directions when using the common myeloid progenitors (CMPs) as start points (Figure 17F), demonstrating the value of BITFAM in creating temporal trajectories.

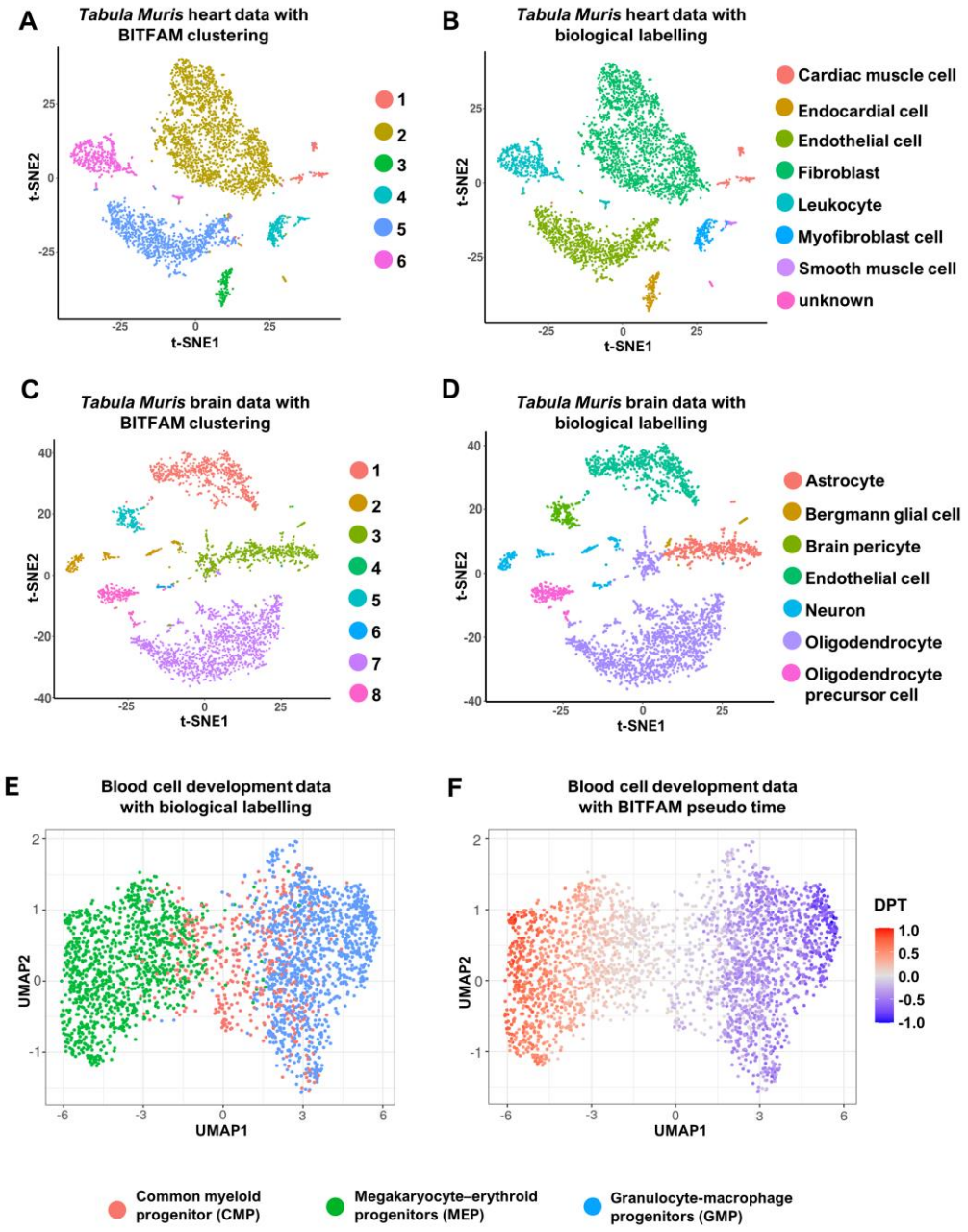

Figure 17. Downstream analysis with inferred transcription factors activities.

(**A**) t-SNE plot of the single cell RNA-seq data from Tabula Muris heart datasets. (**B**) The clustering result of Louvain's algorithm on inferred transcription factors activities. (**C**) t-SNE plot of the single cell RNA-seq data from Tabula Muris brain datasets. (**D**) The clustering results of Louvain's algorithm on inferred transcription factors activities.

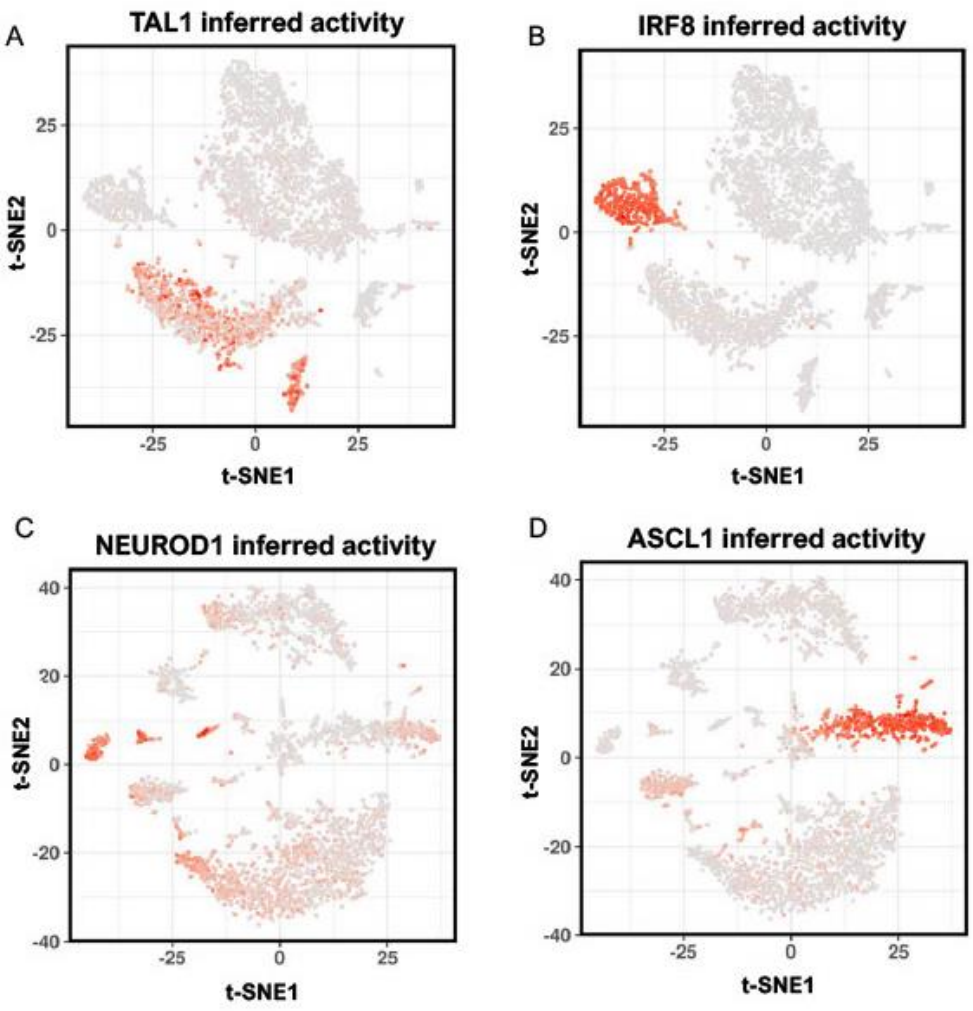

Figure 18. The signature transcription factor activities in heart and brain.

3.3.5 Using different transcription factor target genes for BITFAM

We used a single cell RNA-seq dataset of CRISPRi to test the performance of BITFAM. In this study, fifty transcription factors are deleted or knocked down by CRISPR interferon screening [140]. In each cell, one of the fifty transcription factors is deleted. So, the activity of that transcription factor in the cell should be lower than other cells. This made the dataset well-suited for evaluating BITFAM performance. On the CRISPRi dataset, we implemented BITFAM, and we assessed the performance using the AUROC-based approach.

BITFAM used prior knowledge of transcription factor target genes identified by ChIP-seq to infer transcription factor activity. We used randomly selected genes as the prior knowledge about target genes of transcription factors instead of the experimentally identified target genes. Then this randomly generated target gene list is applied in the inference of BITFAM in the Tabula Muris lung and CRISPRi datasets. By doing so, we could assess the importance using ChIP-seq as the prior knowledge.The AUROC for randomly chosen TF targets decreases from 0.575 to 0.482 in the CRISPRi data (Figure 19A), demonstrating the importance of using ChIP-seq identified target genes. The PAX5 inferred activity in the BITFAM with ChIP-seq target genes in the Tabula Muris lung datasets nicely matches the knowledge about the function of PAX5 in B cell; however, when the target genes were randomly shuffled, the PAX5 inferred activity was no longer restricted to B cells (Figure 19B, C). This again demonstrates the importance of using ChIP-seq identified target genes.

Next, we investigated the effect of using different approaches to get the target genes of transcription factors as prior knowledge. We tried multiple ways to predict the

target genes of transcription factors. These methods included examining the binding motifs of transcription factors in the gene promoters, the co-expression between genes and transcription factors, and other existing methods that integrated binding motifs with ChIP-seq data (ChIP-eat) [150]. Using these approaches, we generated other target gene sets for transcription factors. We then evaluated the performance of BITFAM using these new transcription factor target gene sets in the CRISPRi dataset. Compared to the default BITFAM with ChIP-seq data identified target genes, these methods had a lower AUROC (ChIP-seq target: 0.575, ChIP-eat identified genes: 0.563, co-expression genes: 0.511, genes with binding motifs: 0.496) (Figure 19D). The transcription factors can also bind on the distal regions (regulatory enhancers) to regulate the expression of target genes. So, we also used the ChIP-seq signals that locate in the whole chromosome as the binding regions for target genes. In CRISPRi dataset with the extending regulatory regions, the AUROC is 0.504 which was lower than using the promoter regions in the [-2000, +200] around TSS (AUROC: 0.575) (Figure 20).

Then, we wanted to understand the impact of the source of ChIP-seq data on BITFAM result. We used another public available ChIP-seq database, REMAP [151]. The result showed that BITFAM with GTRD ChIP-seq data performed better than using the REMAP database as the target genes of transcription factors (Figure 21). We then investigated how the cell type that the ChIP-seq data were generated impact the inference on the scRNA-seq data from cell types that are not relevant. Thus, we manually selected some transcription factors ChIP-seq experiments from certain cell types and generated new target gene lists for these transcription factors. Then, we applied BITFAM on the single cell RNA-seq dataset with this new knowledge (Figure 21). We discovered that

even when we used non-hematopoietic ChIP-seq datasets as background knowledge, BITFAM inference results of CEBPA activity still corresponded to known biological functions. This indicated that for a TF like CEBPA, for which many ChIP-seq datasets are avialable, the cell types of the ChIP-seq data did not influence the inference result too much (Figure 22A, B). However, when we filtered non-B cell ChIP-seq datasets for the TF PAX5, which is a transcription factor that is specific to B cells, we found that the model will not assign B cells with high PAX5 activity (Figure 22C, D). For the transcription factor NEUROD1, 19 ChIP-seq datasets are available from various cell types and conditions. There were no cell types related to brain neurons available in the CHIP-seq database. Nonetheless, BITFAM inferred a high activity of NEUROD1 in brain neurons. However, the BITFAM inference of NEUROD1 activity changed the profile when remving the pituitary tumor line (Figure 22E,F). These findings imply that ChIP-seq data collected from those cell types or closely related cell types might enhance accuracy of the inferred transcription factors activities for certain TFs.

We investigated the impact of the range of promoter regions used in identifying the transcription factor target genes. We are using the 2000 bps upstream and 200 bps downstream around the transcription start site (TSS) as the default setting of gene promoter regions. Then, we evaluated the performance of multiple settings of promoter regions. 1) [-500, +50] around TSS, 2) [-1000, +100] round TSS, 3) [-5000, +5000] around TSS, 4) [-10kb, +10kb] around TSS. Using the CRISPRi dataset, we found the best performance when setting the promoter region as [-2000, +200] and [-1000, +100] ([-

1000, +100]: 0.581, [-2000, +200]: 0.575, [-5000, +5000]: 0.536, [-500, +50]: 0.501, [-10kb, +10kb]: 0.512) (Figure 18E).

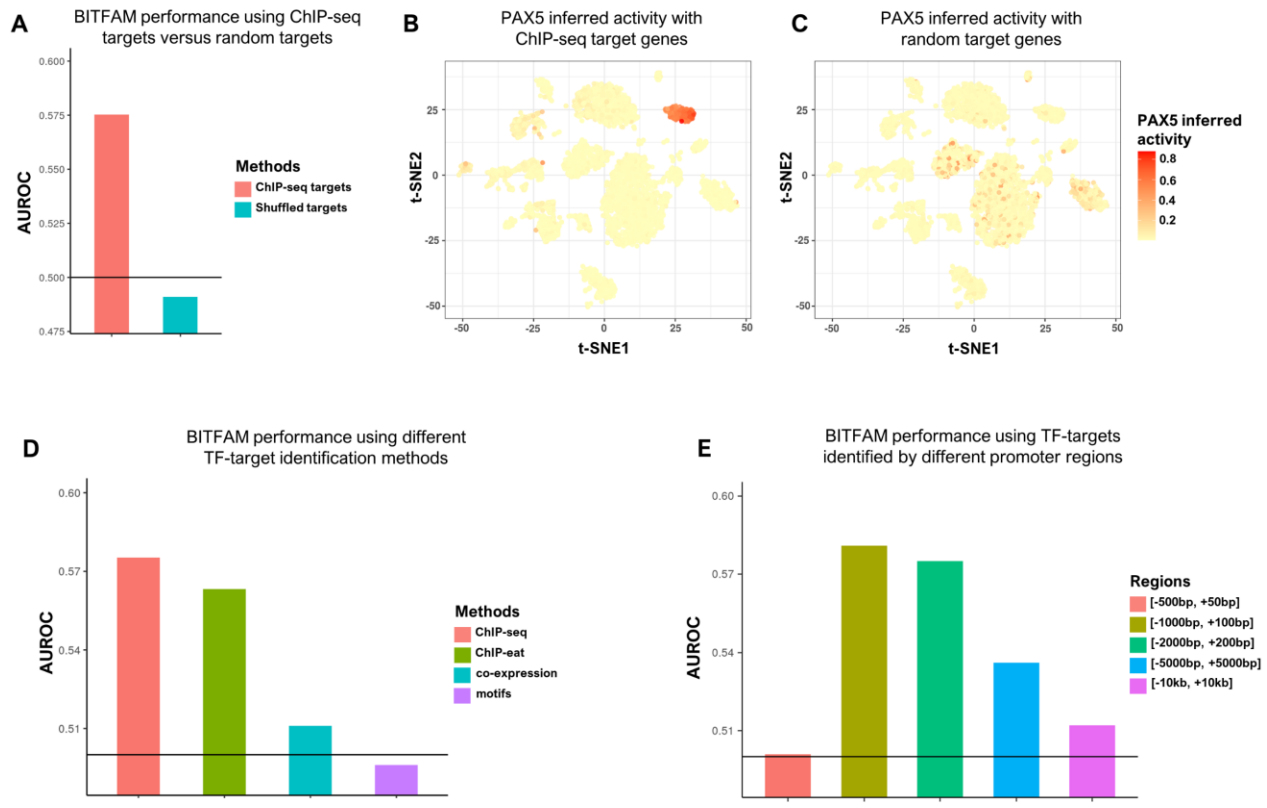

Figure 19. The validation of BITFAM with different prior knowledge.

(**A**)(**B**) The AUROC of transcription factor activities by random target genes. (**B**)(**C**) The PAX5 inferred activities in lung with different ChIP-seq and random targets. (**D**) AUROC of BITFAM with different TF target genes identification approaches (CRISPRi scRNA-seq). (**E**) AUROC of BITFAM with different promoter regions (CRISPRi scRNA-seq).

Figure 20. BITFAM performance using distal ChIP-seq signals.

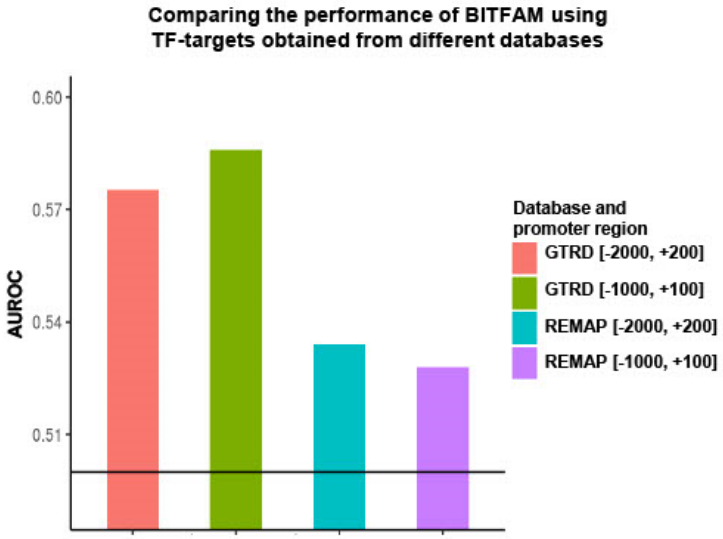

Figure 21. Comparison of distinct ChIP-seq input databases.

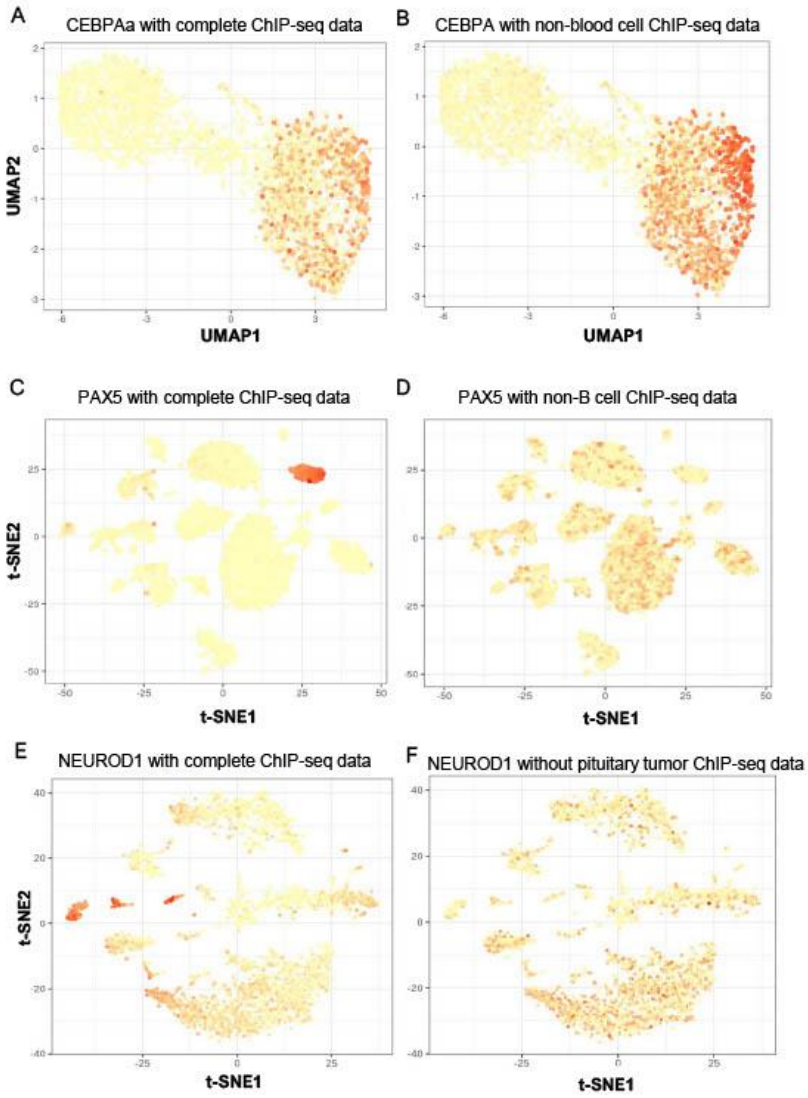

Figure 22. The inferred activities before and after partially removing ChIP-seq datasets.

3.3.6 Performance of BITFAM and existing methods on inference of transcription factor activities and subpopulation identification

We compared BITFAM with other methods to validate the performance of our model. First, we evaluated the performance of BITFAM and SCENIC[89]. We applied both methods to the CRISPRi dataset. For SCENIC, we tried two settings: 1) using only motifs within [-10kb, +10kb] around TSS, 2) using motifs and ChIP-seq data with [-10kb, +10kb] around TSS. The AUROCs of SCENIC were 0.537 and 0.557, respectively. For BITFAM, we tried three promoter regions: 1) [-2000, +200] around TSS, 2) [-1000, +100] around TSS, 3) 5kb-10kb regions. The AUROCs in optimal regions of BITFAM ([-2000, +200] with AUROC 0.575 and [-1000, +100] with AUROC 0.581) are silightly better than SCENIC (Figure 23A) . We also compared the running time of BITFAM and SCENIC on the CRISPRi dataset and Tabula Muris dataset. BITFAM spent 2.81 hours and 2.54 hours in these two datasets. They were significantly faster than SCENIC on the same datasets (10.23 hours and 11.78 hours) and the computer with AMD 3900XT 12-core CPU (Figure 23B).

Next, we evaluated transcription factors target genes that were identified by BITFAM and SCENIC. We compared the top weighted genes that are inferred by both methods for the same transcription factors. By checking the Jaccard index (Jaccard index > 0.1), the target genes learned from these two methods had few overlaps (Figure 24). However, the inferred transcription factors activities learned by SCENIC and BITFAM had some overlap (Figure 25). GO enrichment analysis was performed on the top 100 positive weighted genes. Compared to the results from the same GO analysis on the target genes learned by SCENIC. The target genes identified by BITFAM were more enriched for the

known functions of the TFs. For instance, an essential transcription factor, RELB, which regulates immune response and cell survival[152]. Its target genes identified by SCENIC had functions of RNA and DNA metabolic processing (Figure 26A). However, the most important target genes of RELB learned from our model had functions about immune cell activation, proliferation and apoptosis.

We also tested the clustering performance between using inferred transcription factor activity profile and the gene expression profile. The inferred transcription factor activity profiles were used as input to Louvain's clustering algorithm[128]. The clustering results then were compared with other approaches such as Seurat[70], SIMLR[153], and SC3[154]. These methods used single cell gene expression profiles as input to cluster cells. Adjusted rand index, rand index, and normalized mutual information were used to evaluate the clustering quality. In the Tabula Muris datasets, BITFAM-based clustering approach had better ARI, NMI and RI comparing to other approaches (Figure 23C). This indicates that the inferred transcription factor activity could be an alternative way to assess the cell heterogeneity. The clustering with inferred TF activities divides cells into subpopulations that match biological knowledge. This implies that the transcription factor activities inferred by BITFAM are relevant to determining cell function.

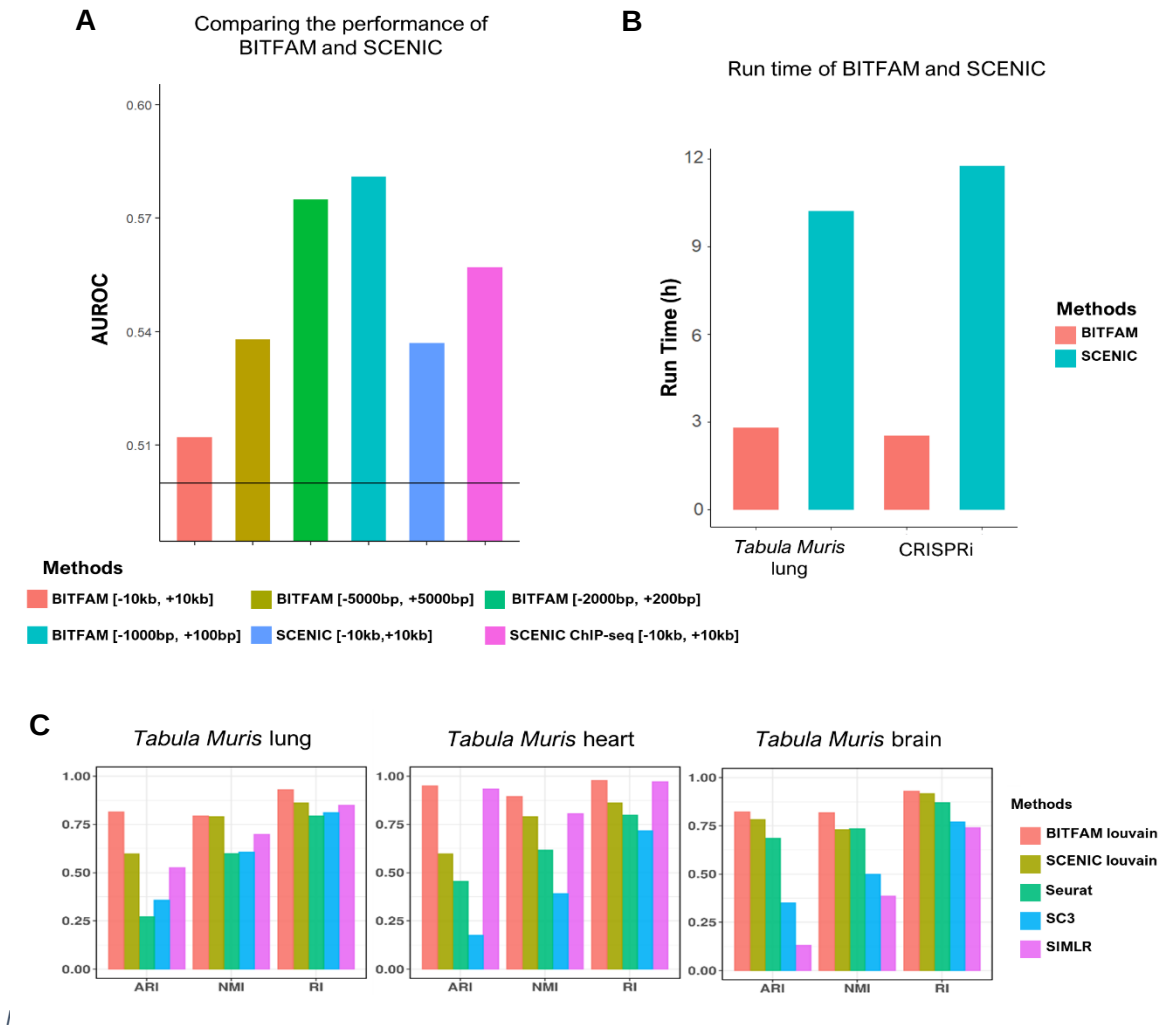

Figure 23. Comparison of BITFAM with other methods.

(**A**) Comparison of BITFAM with SCENIC on the different settings. (**B**) The running time of BITFAM and SCENIC on two datasets. (**C**) Clustering performance comparison between BITFAM and other approaches.

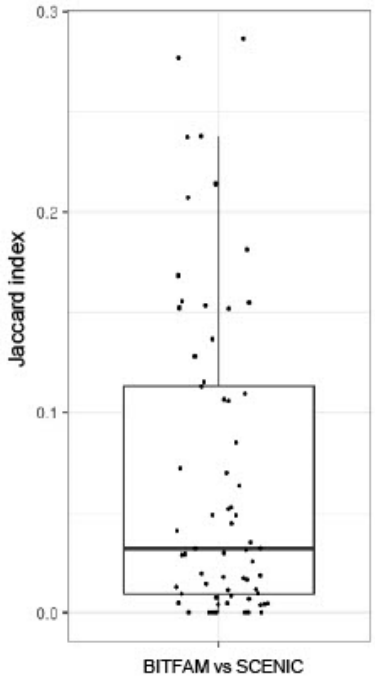

Figure 24. The overlap of target genes identified by BITFAM and SCENIC

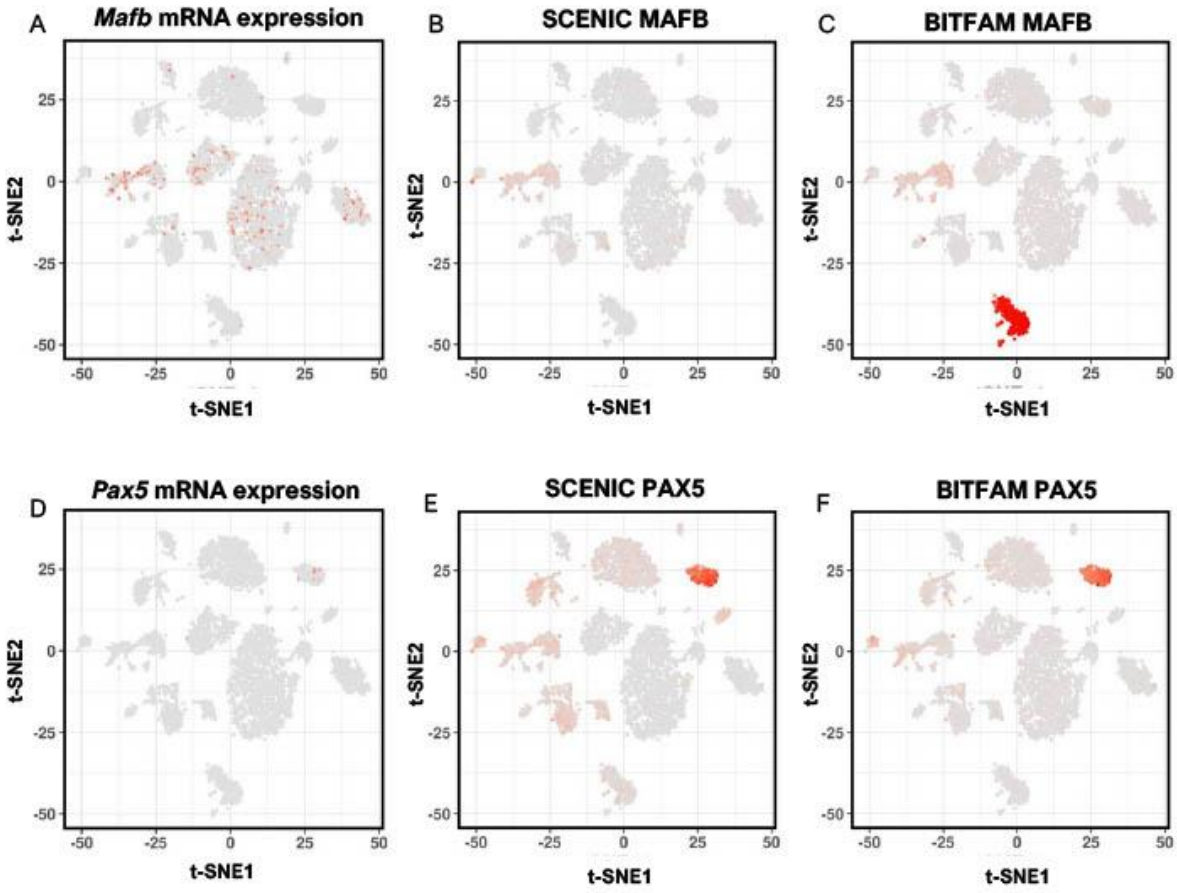

Figure 25. The activities of MAFB and PAX5 inferred by SCENIC and BITFAM.

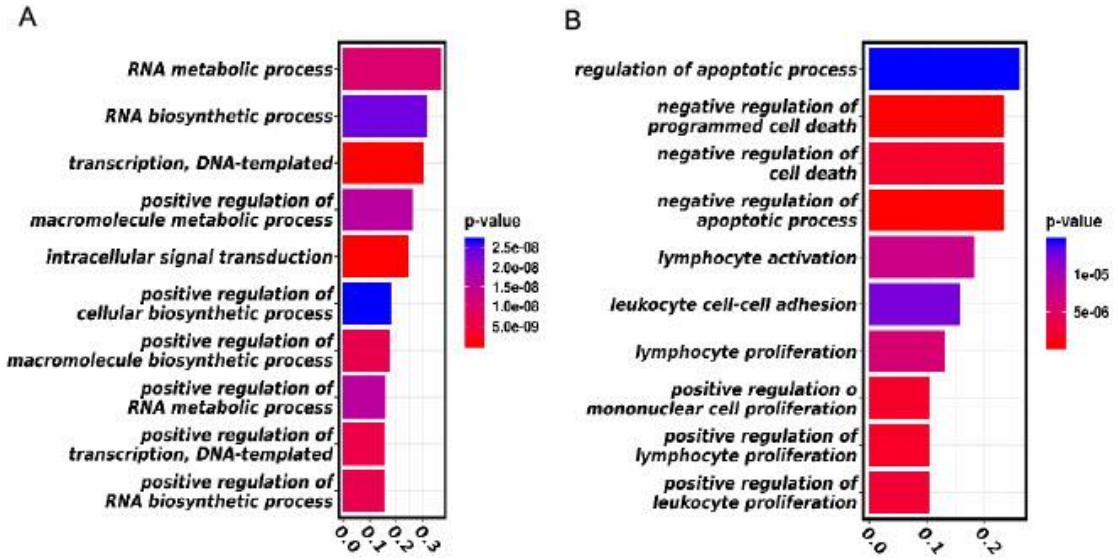

Figure 26. Enrichment analysis of SCENIC target genes and BITFAM target genes for the transcription factor RELB.

3.4 Summary

Single-cell RNA sequencing (scRNA-seq) is an important experimental technique to identify heterogeneity in transcriptomic level. There are two major challenges in scRNA-seq data analysis. One is integrating the transcriptome data with different types of biological data such as chromatin accessibility and transcription factor binding. Another one is recognizing hidden biological functions and regulatory mechanisms, such as gene regulatory networks and transcription factor activities. To address these challenges, we built a method to infer TF activities by integrating existing ChIP-seq data on transcription factor binding and scRNA-seq data. The Bayesian Inference Transcription Factor Activity Model (BITFAM) combines transcription factor ChIP-seq data with binding information and single cell RNA-seq data. We demonstrate that the inferred transcription factor activities from BITFAM correspond to the known functions of the cell types. This inferred transcription factor activity profile can also be used to identify cell heterogeneity as well as other downstream analysis. The inferred regulatory strengths between transcription factors and target genes also match the biological functions of the transcription factor.

Chapter 4 Assessing Importance of DNA Sequence and Epigenetics Profiles for Gene expression using a Deep Neural Network

Adapted from Gao, S., Rehman, J., and Dai, Y. (2022) "Assessing comparative importance of DNA sequence and epigenetic modifications on gene expression using a deep convolutional neural network". Comput Struct Biotechnol J 20, 3814-3823

4.1 Introduction

The accurate regulation of gene expression in one of the most important biological processes. It enables the coordination of complex transcriptional procedures. There are several mechanisms to regulate the transcription of genes. First, the regulatory regions of genes on chromatin must be accessible to be able to bind with transcription factors. The transcription factors will initiate or repress gene expression. DNase I hypersensitivity can be used to assess the accessibility of chromatin[155]. Second, the modifications on histone are another crucial regulator of gene expression because they determine the structure of chromatin. The histone modifications are usually associated with regulation of transcription on regulatory regions. H3K4 trimethylation (H3K4me3), for example, is often associated with chromatin remodeling through the NURF complex to activate transcription of nearby genes[156]. Then, DNA methylation is also considered as an important modification to regulate gene transcription. It usually relates to the close of chromatin. DNA methylation can prevent the transcription factors or DNA Polymerase II from binding to the regulatory regions like enhancers and promoters[30]. Lastly, genetic variants located on the regions that regulate gene expression also impact transcription. These variants may affect the occurrence of the modifications mentioned above by switching the interactivity between regulatory proteins and DNA sequence[157]. By next-generation sequencing (NGS) technologies, these different gene regulation events can

be investigated [24, 158]. The histone modification can be assessed by ChIP-seq with specific antibody. The DNA methylation can be quantified by whole-genome bisulfite sequencing (WGBS). DNA sequence can be obtained by whole genome sequencing (WGS). Building a model to integrate these data can illuminate the relations between different modifications and gene expression and identify novel gene regulatory mechanisms[159, 160].

To assess the relationship between histone modifications and gene expression, multiple computational tools have been established[161]. Several machine learning models were developed to predict transcription level of genes by the epigenetic modifications around and on the gene coding regions. Linear regression[161], random forest[162] and support vector machine (SVM)[163] were used in these tools. In these existing models, the regions surrounding transcription start site (TSS) of genes were separated into smaller bins. The epigenetic signals from NGS data are averaged in each bin to form the features in the model. However, the influence of histone modifications on gene expression could extend to a large region of genome[164] which includes multiple bins. The machine learning models mentioned above cannot fully assess the relations between nearby bins. Thus, the influence of histone modifications on gene expression could not be explicitly explored with these methods.

In recent years, deep learning has been widely applied in systems genomic to study the associations between biological regulations and predict the omics data[165]. One of the deep learning frameworks, convolutional neural networks (CNNs), have been utilized in many biological studies. For example, a CNN model is developed to predict the transcription factor binding sites with the epigenetics modification and DNA sequence[166]. Another CNN model is used to classify cell types from the single cell RNA-seq data [167].

One of reasons that convolutional neural networks could improve the performance of learning and prediction is that it is able to extract both local and global interactions within the features. In the goal of predicting gene expression, multiple CNN models have been developed and these methods are using different types of input data. Some models using DNA sequences alone have accomplished high prediction performance in mouse and human genomes[168-172]. However, they haven't involved other information such as epigenetic modifications which also impact gene expression. Thus, these methods cannot identify the regulatory mechanism of gene expression. Also, these methods predict the mean expression of each gene by aggregating multiple mRNA-seq data. This limits the prediction of gene expression in a specific biological condition which might be very distinct from the existing mRNA-seq data. Histone modification signals is used to predict gene expression as well. In the models DeepChrom and DeepDiff, a CNN model is designed to predict gene expression levels and identify the genes that are differentially expressed between conditions with histone modification on binned regions [164, 173]. Other methods use DNase I hypersensitive sites and DNA methylation signals to predict gene expression[164, 174]. It is necessary to develop a model that integrates both DNA sequence and epigenetic modifications to predict gene expression and identify the interaction of them on regulation of gene expression.

Here, we developed a novel Convolutional Neural Network framework to assess the comparative importance of DNA Sequence and Epigenetic modifications on Gene expression regulation using a deep convolutional neural network (iSEGnet)[175]. After training the neural network model, Integrated Gradients (IG) is used on iSEGnet to calculate the attribution scores of epigenetic modifications on the prediction of gene

expression. For each site on the genome, Integrated Gradients will provide a score. This attribution score indicates the site- specific relevance effect of the epigenetic modification on gene expression. We applied iSEGnet on the multi-omics data of six different cell lines/types. These datasets are obtained from the ENCODE project[24]. We compared the prediction performance of iSEGnet with other machine learning approaches, such as linear regression, random forest and SVM. We demonstrated that the iSEGnet models has a better performance than other models on gene expression prediction. We also used iSEGnet to reveal the positions in cis-regulatory regions which are essential for gene expression prediction by Integrated Gradients. By comparing these high attribution regions with transcription factor binding sites derived from ChIP-seq data, we found that these regions significantly overlap with active regulatory regions. Moreover, we use iSEGnet on cancer multi-omics data in order to uncover potential TFs that have specific regulatory functions in different conditions.

4.2 Materials & Methods

4.2.1 Datasets and preprocessing

In the iSEGnet model, there are three types of data including epigenetic modification signals, DNA sequences of genes and mRNA levels. We used gene expression and epigenetic modification data from six different cell lines/types. These datasets were obtained from the ENCODE project[24]. A549, HepG2, K562, large intestine, pancreas, and small intestine are cell types used in iSEGnet training and evaluation. The DNA sequences on the regulatory regions of genes were collected from hg38 reference genome in NCBI database.

We additionally obtained datasets of breast cancer multi-omics[176] and esophageal tumor[48] multi-omics data from Gene Expression Omnibus (GEO) (accession numbers: GSE118716 and GSE149612). In the breast cancer study, the histone modifications H3K4me1 and H3K4me3 ChIP-seq, DNA methylation whole-genome bisulfite sequencing, and gene expression RNA-seq data were generated from breast cancer cell line TAMR (drug-resistant, endocrine-resistant derivatives tamoxifen-resistant) and breast cancer cell line MCF7 (drug-sensitive, endocrine-sensitive). Each breast cancer cell line has one sample. DNA methylation whole-genome bisulfite sequencing data and gene expression RNA-seq data were provided in the esophageal tumor study. 9 patients were involved in the project with both normal and tumor tissues. We tested the application of iSEGnet on these two cancer-related datasets as case studies.

The histone modification ChIP-seq and DNase-seq data were preprocessed by the ENCODE project. Bowtie2 and MACS were used to generate the signal p-values on each binned regions of whole genome [158]. The signal p-values are stored in bigWig format. The p-value indicates whether the signal in a segment is significant different compared to the background. The length of the binned regions are 20bps. So, for each site in the 20bp length regions, we used the bin p-value as the p-value of site. Then we used the negative log-transform on the p-value and scaled it to [0,1].

For DNA methylation whole-genome bisulfite sequencing data, we used the reprocessed raw signal data in bigWig format from ENCODE project. In each CpG site, the value is in the range from 0 to 100 corresponding to the methylation percentage at that site. We scaled the data to [0, 1], and for other regions that are not able to be methylated, we set them as 0.

For DNA sequence data, one-hot coding was used to convert the DNA sequences from strings to binary-valued matrices. The matrices have the shape of $L \times 4$, where the L is the length of input gene regulatory regions and 4 is the type of nucleotides. In section 3.3.1, we will talk about the approach to select L.

For gene expression data, we used the normalized counts (transcripts per million counts, TPM) and removed the unexpressed genes (the count of gene mRNA is zero)

4.2.2 iSEGnet architecture

As shown in Figure 27, in iSEGnet, a deep learning framework based on convolutional neural networks is developed. Each gene is a sample to train the model. iSEGnet require two parts of inputs including the DNA sequences and the epigenetic modification signals on the gene regulatory regions. iSEGnet predicts the gene expression level (TPM) of the genes. Thus, RNA-seq gene expression is required as the prediction targets to train the model.

In iSEGnet, there are multiple convolution layers. Each convolutional layer has several convolution kernels with the same size. Every kernel in the convolutional layer extracts local information from a different aspect. In the first convolutional layer, the information on the gene regulatory regions were extracted. The second convolutional layer extracts the high-level features from the outputs of first convolutional layer. Kernel size on the second convolutional layer decreases compared to the kernels in first convolutional layer. The activation function of each neuron in the network is the rectified linear unit (ReLU), i.e.,

$$ReLU = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Dropout is used to avoid overfitting for model regularization.

During the training of iSEGnet, for the input corresponding to the epigenetic modification, the data is provided into the first convolutional layer. The kernel size of this layer is $l \times m \times k_1$, where m is the number of epigenetic modification. For the input of the DNA sequence, the data is provided into the first convolutional layer. The kernel size of this layer is $l \times 4 \times k_1$. 4 is the fixed because the number of nucleotides. A max-pooling layer is used after the convolutional layer to further extract features from the most importance in a region with fixed size. Next, the second convolutional layer and max-pooling layer are added to the model. The kernel size of second convolutional layer is $l \times 1 \times k_2$. By doing so, two input data, the DNA sequence, and the epigenetic modifications, are in the same size ($L' \times 1 \times k_2$). Then, the outputs of the second layers are concatenated by columns. This new concatenated data was fed to the last convolutional layer (kernel size: $l \times 2 \times 64$). Lastly, the outputs from the last convolutional layer were flattened and inputted into a fully connected network. mRNA abundance level is the final output of the iSEGnet framework. TPM is used for the value of mRNA abundance. Hyper-parameters were. l, k_1, k_2 tunned with 5-fold cross validation. iSEGnet is built and trained with TensorFlow 2.0 [177, 178].

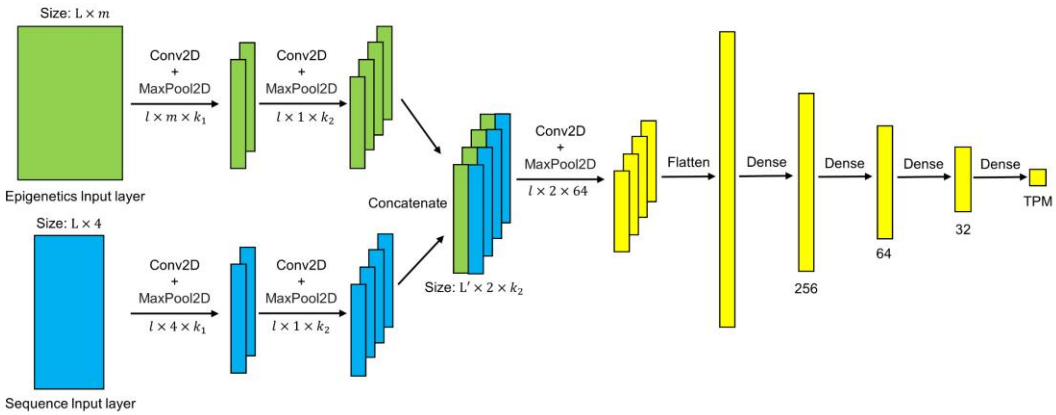

Figure 27. The overview of the iSEGnet framwork

4.2.3 Model training and testing

ADAM optimizer [179] was used to train iSEGnet and the loss function of iSEGnet is,

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2 + \lambda \sum \left|\left|W\right|\right|_2$$

where N is the number of genes in training data. Mean squared error (MSE) between the predicted expression y_i' and the observed gene expression y_i forms the first term. L2 regularization with a penalty parameter λ are the second term.

In the data of each cell type or cell line, 60% of the data are used as the training set. 20% of data are the validating set and 20% of data are the testing set. The network was trained with 200 epochs, and we set 100 as the batch size. To avoid overfitting, batch normalization was used in the model. Early stopping was used by evaluating the prediction accuracy on the validating set at each training iteration to prevent overfitting. Performance of the model was reported using the predictions accuracy on the test sets.

Hyper-parameters were tuned by 5-fold cross validation. The hyper-parameters are dropout rate, number of kernels, number of layers, the L2 regularization parameter λ, and kernel size. In each of the convolutional networks, there are two convolutional layers. The combination of number of kernels varied from 16, 32, 64, and 128 with different sizes: 20×7, 50×7, and 100×7. For L2 regularization parameter λ, 5 different values: 0.0001, 0.01, 0.1, and 1.0 were selected. 0.1, 0.3, and 0.5 were tested as the dropout rate. The model performances with the corresponding hyper-parameters are shown. From the analysis results, we found that the hyper-parameters with 64 neurons for the 1st and 128 neurons for the 2nd convolutional layers, λ value of 0.0001, dropout rate of 0.5, and kernel

size of 20x7 have the highest accuracy (Figure 28). Thus, in the rest of this study, we use this combination of hyperparameters in our model.

Another architecture of framework was also evaluated. We remove the last convolutional layer in iSEGnet and added the fully connected layers right after the second concatenation layer.

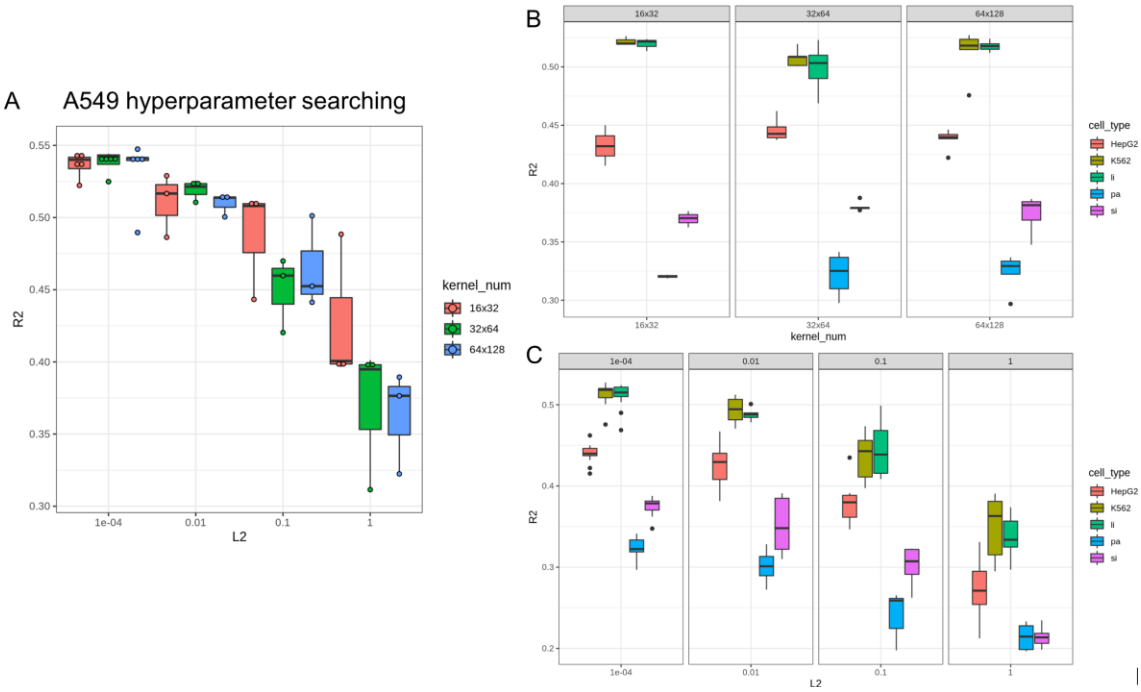

Figure 28. The performance of hyperparameter searching

4.2.4 Evaluation criteria

The coefficient of determination (R^2) of the test set for each dataset is used to evaluate the model performance. The coefficient of determination (R^2) is calculated by $R^2 = 1 - SSResidual/SSTotal$, where $SSResidual$ is the sum of residual squares of predictions and $SSTotal$ is the total sum of squares from the observations. Pearson's correlation coefficient between the predicted and observed expression is also used as an alternative measurement to assess the performance of models.

4.2.5 Comparison with other gene expression prediction methods

We compared iSEGnet with other machine learning models, including support vector machines and random forest. Python module, scikit-learn [180, 181], was used to build and train these models. *sklearn.svm.NuSVR* with rbf kernel, *sklearn.svm.SVR* with linear kernel and rbf kernel were used to test SVM. *sklearn.ensemble.RandomForestRegressor* with 100 and 200 decision trees used to test random forest. The grid search for hyperparameter tuning were conducted for support vector machines and random forest. 5-fold cross-validation are used in the tuning and the mean R^2 values were reported for every combination of hyperparameters. The same training and testing data were used in these machine learning models and iSEGnet.

4.2.6 Feature attribution identification

Integrated Gradients (IG)[182] was used to calculate the attribution score of epigenetics modifications on each site in the gene regulatory regions. By investigating the attribution score, we reveal the most relevant modifications and regions for gene expression prediction. Integrated Gradients interprets the predictions from a function F defined on a feature space X. In our case, the F is derived by the trained iSEGnet model.

The attribution of each feature on the prediction are calculated by a reference of that feature and $\boldsymbol{x}' \in X$ and its prediction $F(\boldsymbol{x}')$. The attribution for the jth feature $\boldsymbol{x}_j$ is defined in integrated gradient as

$$attr_j(\boldsymbol{x}) := (\boldsymbol{x}_j - \boldsymbol{x}'_j) \times \int_0^1 \alpha \, \partial F\big(\boldsymbol{x}' + \alpha \times (\boldsymbol{x} - \boldsymbol{x}')\big) \partial x_j \, d\alpha$$

where $\boldsymbol{x}'$ is the reference point of feature $\boldsymbol{x}$. $\alpha \in [0,1]$ is the parameter that control the steps between reference point and the observation. F is the learned prediction model by iSEGnet. $F(.)$ is the prediction of model with a specific input.

Function *alibi.explainers.IntegratedGradients* [183] with *n_steps=200* was used to perform Integrated Gradients.

4.2.7 Transcription factor ChIP-seq data for validation

We compared the transcription factor binding regions identified by ChIP-seq and the high attribution regions identified from Integrated Gradients of iSEGnet. By doing this, we could investigate whether the known transcription factor binding regions are overlapping with the high attribution regions. The ChIP-seq data were obtained from from the ENCODE project (ENCSR000DYC). Negative log-transformation was used on the p-values of the peaks as the MYC binding signal on specific regions.

4.2.8 KEGG enrichment analysis

To reveal the signature KEGG pathways[184] of the transcription factors binding to the high attribution regions, we conducted enrichment analysis with Enrichr[185-187]. The enrichment testing is based on Fisher's exact test. For the multi-test correction, we used Benjamini-Hochberg method with 0.05 as the adjusted p-values significant threshold.

4.3 Results

4.3.1 Identification of the best combination of regulatory regions as input to predict gene expression

There are four cis-regulatory regions that are important to coordinate gene expression, i.e, promoter, 5'-UTR, 3'-UTR, and terminator. We evaluated the combination of these four regions and the different lengths of these regions as the input to iSEGnet. We tried to understand the impact of these regions with different epigenetics modifications on gene expressions. The suitable length of these regions also remains not well-defined. We input the combinations of these regulatory regions with different lengths to iSEGnet, in order to investigate which regulatory regions as the input feature can reach the best prediction performance. We trained models with the different settings of regulatory regions as input. Then we found the combination with the best prediction performance on the test set (details see Methods). First, promoters and 5'-UTRs were used as inputs to iSEGnet. We tried different combinations of promoters and 5'-UTR with different lengths. 1000bp upstream of TSS and 500bp downstream of TSS ([-1000bp, +500bp] around TSS) was the region that reach the highest coefficient of determination (R^2) across different combinations and all six lines/cell types (Figure 29). Then this region about promoters and 5'-UTR was fixed. We combined it with the 3'-UTR and terminator with various combination of lengths. Model performance with all four regions increased 0.10-0.15 in R^2 values compared to the model with only promoters and 5'-UTRs in all six lines/cell types (Figure 29). This result highlights that the regulatory regions around transcription terminating site is also important to gene expression regulation. The performance of models was most consistent when we set 3'-UTR as 500bp upstream of

TTSs. The model had the best performance when we used the 500bp upstream of TTSs as 3'-UTR. Thus, [-1000bp, +500bp] around TSSs and [-500bp, +500bp] around TTSs were selected as the final input regions to train and use the model for the rest of the study (Figure 30). We also randomly shuffled the gene expression across genes to generate a random dataset to have a negative control data for the model performance. Then we trained the model in this dataset and evaluated the performance on the testing set. The R^2 values are very close to 0 when trained model with this random dataset in three cell lines. R^2 value was -0.001861 for A549 and -0.02004 for K562. This demonstrated that the prediction of iSEGnet model is not influenced by randomness.

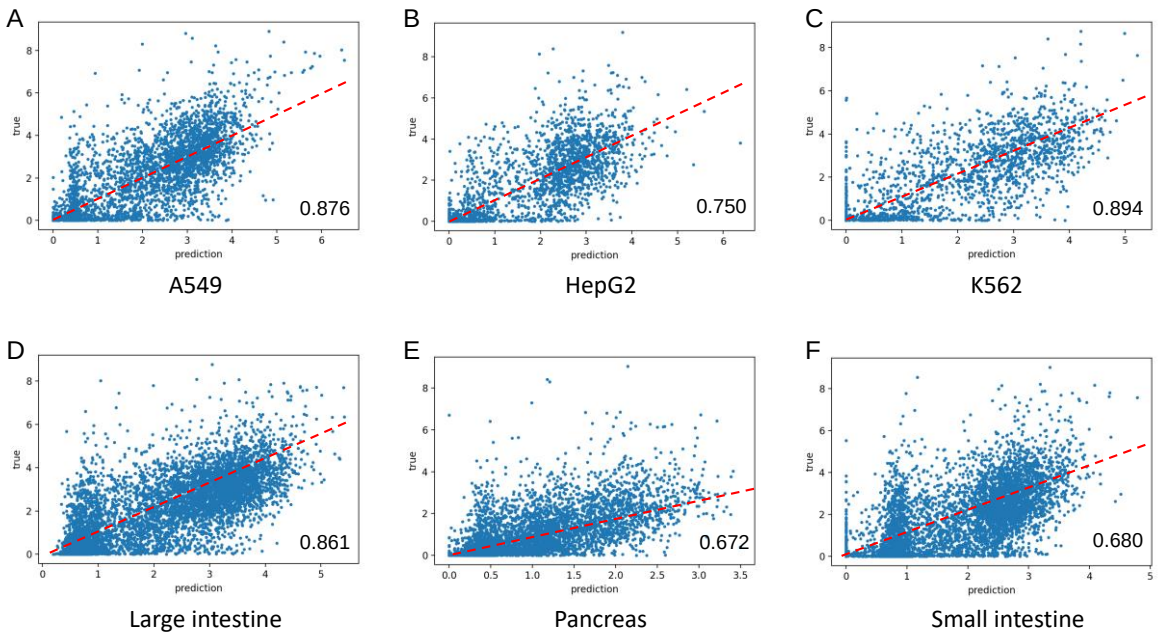

Figure 29. The Pearson's correlation of iSEGnet with optimal input regions.

(**A**) The R^2 values of the iSEGnet models with different TSS regions as input for the six cell lines/types. (**B**) The R^2 values of the iSEGnet models with different TSS regions as input for the six cell lines/types.

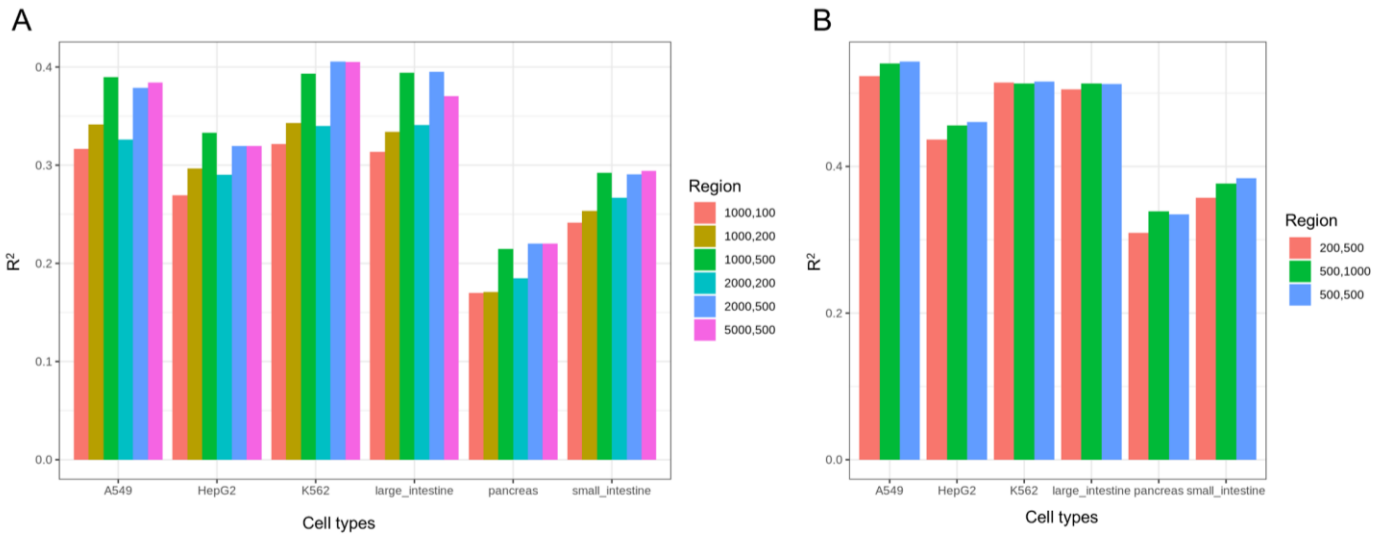

Figure 30. The performance of iSEGnet on different cell lines and cell types

4.3.2 The iSEGnet models has a better performance than other machine learning approaches

There are other machine learning models that are widely used in many different areas, such as support vector machine and random forest. So, we compared iSEGnet to these machine learning methods. In the cell lines data (A549, HepG2, and K562), iSEGnet had a better performance than the other machine learning models. The R^2 values and Pearson's correlations are 0.15-0.30 higher in the prediction of iSEGnet compare to the results from support vector machine and random forest (Figure 31). In primary cells, iSEGnet also had better performance than these methods on the data from large intestine dataset. For the data from pancreas and small intestine, iSEGnet and other machine learning methods had similar performance.

Next, we compared iSEGnet with DeepChrom. DeepChrom is a CNN model to predict gene expression with histone modification data. Based on the expression of the

genes, DeepChrom divided the genes into two groups. If the expression of a gene is higher than the median of all gene expression, it will be labeled as "high expression gene."

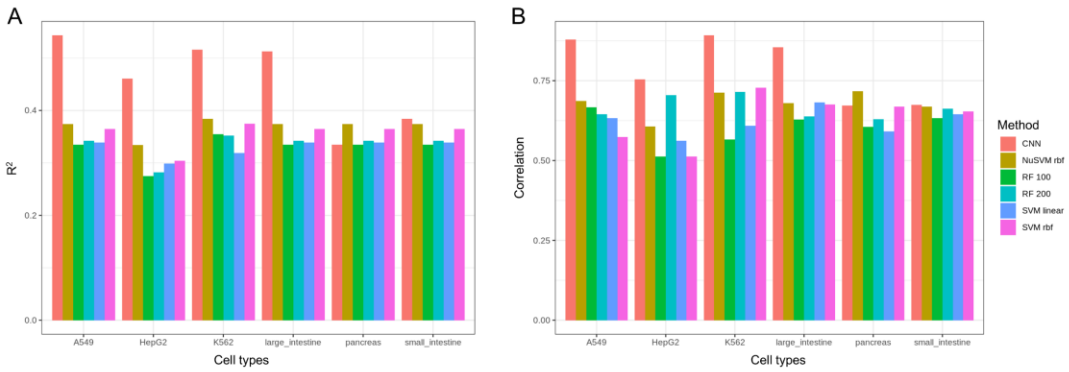

Figure 31. The performance of iSEGnet compared with other machine learning models

(**A**) The R^2 values of different models for the six cell lines/types. (**B**) The Pearson's correlations of different models for the six cell lines/types.

Otherwise, it will be labeled as "low expression gene." However, our model used normalized gene expression values as the prediction target which is a positive real number. In order to compare these two methods with fundamental different settings, iSEGnet has been modified to make the methods comparable. First, the genes are labeled to "high" and "low" with same approach in DeepChrom. Next, we changed the loss function of iSEGnet. The binary cross-entropy was used in modified version of iSEGnet and all other parts kept the same. In Figure 32, compared to DeepChrom, the performance of the classification results by iSGEnet has a better performance in all cell types.

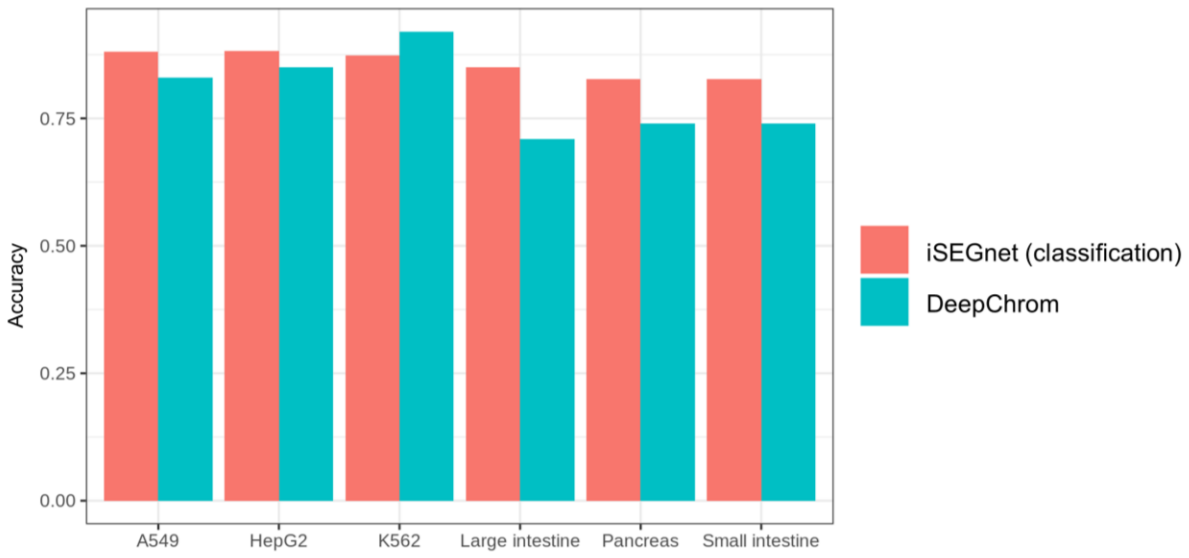

Figure 32. The performance of the binary version of iSEGnet and DeepChrom

4.3.3 Gene expression prediction are impacted by different epigenetics modifications

Next, we explored the impact of epigenetic modification on gene expression. First, we trained iSEGnet using one type of epigenetic modification only. Then we report the performance of the model with the prediction on testing data. We found that, in all cell types/cell lines, the predictions with epigenetic modifications are more accurate than the models with DNA sequences. Moreover, in different cell types, each of epigenetic modifications have different importance patterns (Figure 33). Using this method, we could evaluate the contributions of the DNA sequence and specific epigenetic modification to the performance. The models with only DNA sequences as input get the worst gene expression prediction accuracy (p-value 0.05 in paired t-test,) in all cell types. The performance of the models differed for various cell types when just one kind of epigenetic change was utilized as the input. The most significant epigenetic alterations for predicting gene expression, in general, were H3K36me3 and H3K4me3. Notably, the model that included every DNA sequence and epigenetic modifications have the best performance, highlighting the importance of a thorough and integrated strategy for the prediction of gene expression.

In A549, the prediction results from the model with both DNA sequence and epigenetics data and the model with epigenetics data only were compared in order to assess the impact of epigenetic modifications. The top 10% of genes which have smallest differences between the predicted gene expressions by these two models are selected. The expressions of these genes were minimally influenced by the DNA sequence. On the regions around TSS in these genes, higher DNase ChIP-seq signals were observed

compared to other genes, whereas lower DNA methylation levels were discovered on TSSs and 5'-UTRs (Kolmogorov–Smirnov test, p-value < 0.05) (Figure 33 A). We also trained the models with only DNA sequence data. Similarly, we compared this model with the model trained with both DNA sequence and epigenetics data. Then, the motifs enriched on the promoter regions of the genes with minimal differences on predicted expression level were identified and compared to other genes. These studies suggest that iSEGnet may help identify gene subsets that are more possible to be controlled by epigenetic modifications or DNA sequences, which will aid in the development of new hypotheses.

4.3.4 The epigenetic modification attributions on gene expression at each site of the gene expression regulatory regions

In order to find the input regulatory regions with the greatest prediction attributions for gene expression, we first identified the relative importance of different epigenetic modifications on gene expression. The Integrated gradient method was used on the trained models in all six cell lines/types. The attributions every epigenetic modification taking place on each site of input regions. We found that DNase I signals ("open chromatin") in the area around transcription starting site were crucial for mRNA level prediction in cell line A549 (Figure 33B). The mean prediction attributions for DNase I signals have a peak on the TSS regions across all genes.

This finding corresponds to the prior biological knowledge that the accessibility of the chromatin on the TSS regions which controlled by DNase I binding is important on the regulation of gene expression. H3K4me3 and H3K36me3, which are histone modifications, took place in 5’-UTRs impacted gene expression significantly. On the other

hand, H3K36me3 occurred at TTS regions had high prediction attributions on gene expression. In cell line K562, we observed similar patterns (Figure 33C). In the HepG2 cell line, the patterns of mean prediction attributions of epigenetic modifications on input regions are distinct from the patterns in the other cell lines (A549 and K562) (Figure 33 D). For example, in HepG2, the prediction attributions of histone modification H3K36me3 are much lower on TTS regions. These results suggested that the relative prediction attributions of certain epigenetic modifications vary depending on the cell type.

In order to show how the integrated gradient attributions looked like in one single gene, we used the attributions on the input region of MYC in the A549. MYC is an important oncogene. It has crucial functions in apoptosis and cell cycle progression[188]. For the DNase I signal which indicates the open chromatin, its prediction attribution locates on two high value regions. One was around the transcription starting site, and another was at the 500bp upstream of transcription starting site (Figure 34A). For the prediction attribution of H3K4me3 signals, we observed a high attribution region which located at 5'-UTR. This finding indicated that H3K4me3 presence on this region might have a high impact on gene expression (Figure 34C). The locations and values of the high attribution regions changed when comparing to the ChIP-seq signals (Figure 34B, D). These results demonstrated that in iSEGnet the prediction attributions computed by Integrated gradient are not simple reflections of observed signals. Instead, they reflect the importance of epigenetic modifications on the input regions for regulating gene expression.

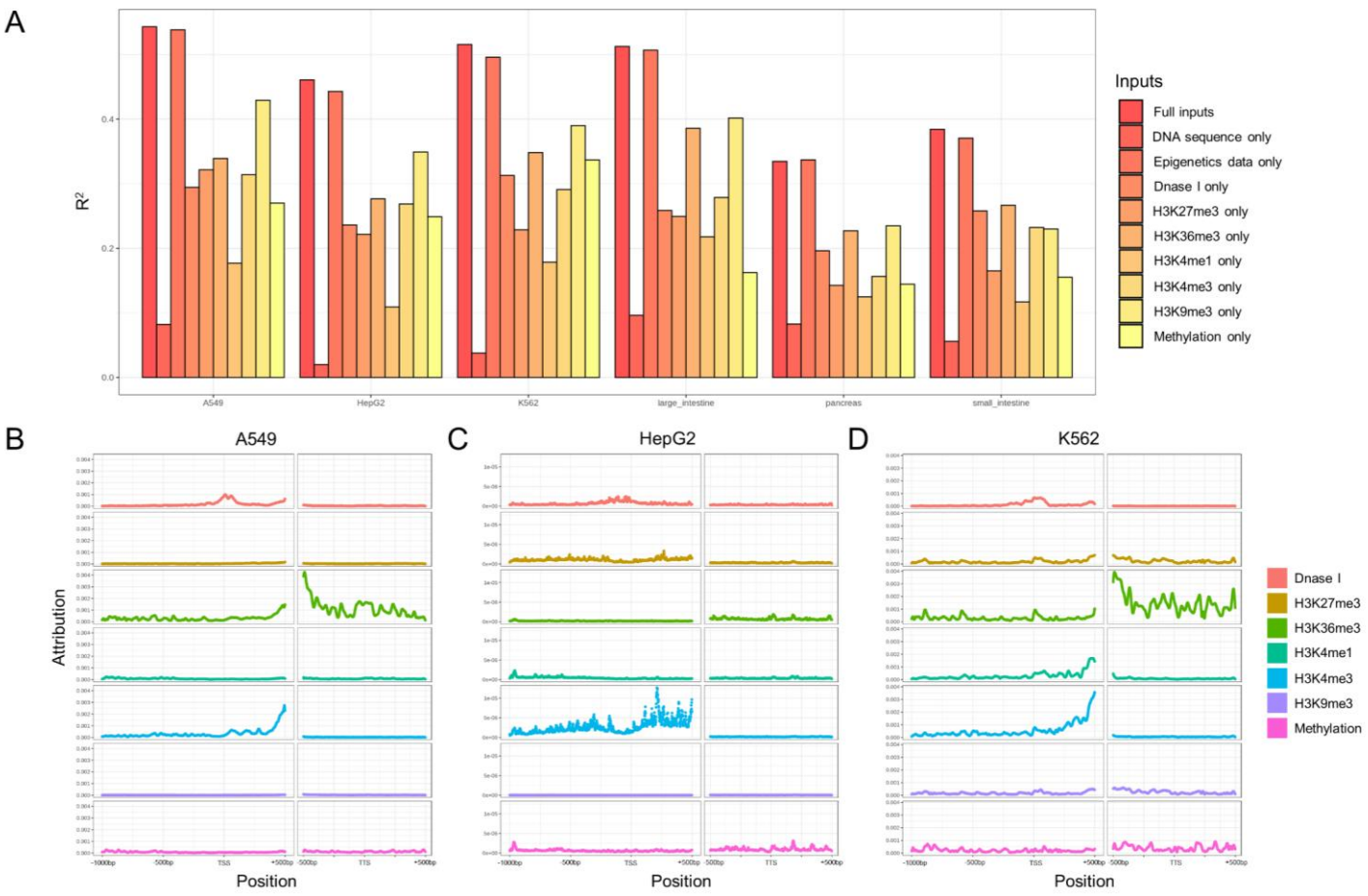

Figure 33. Identification of the attributions of epigenetic modification on gene expression.

(**A**) The R^2 values of iSEGnet with different epigenetics modification data as input for the six cell lines/types. (**B**) (**C**) (**D**) The mean integrated gradient attributions across all genes for six cell lines/types.

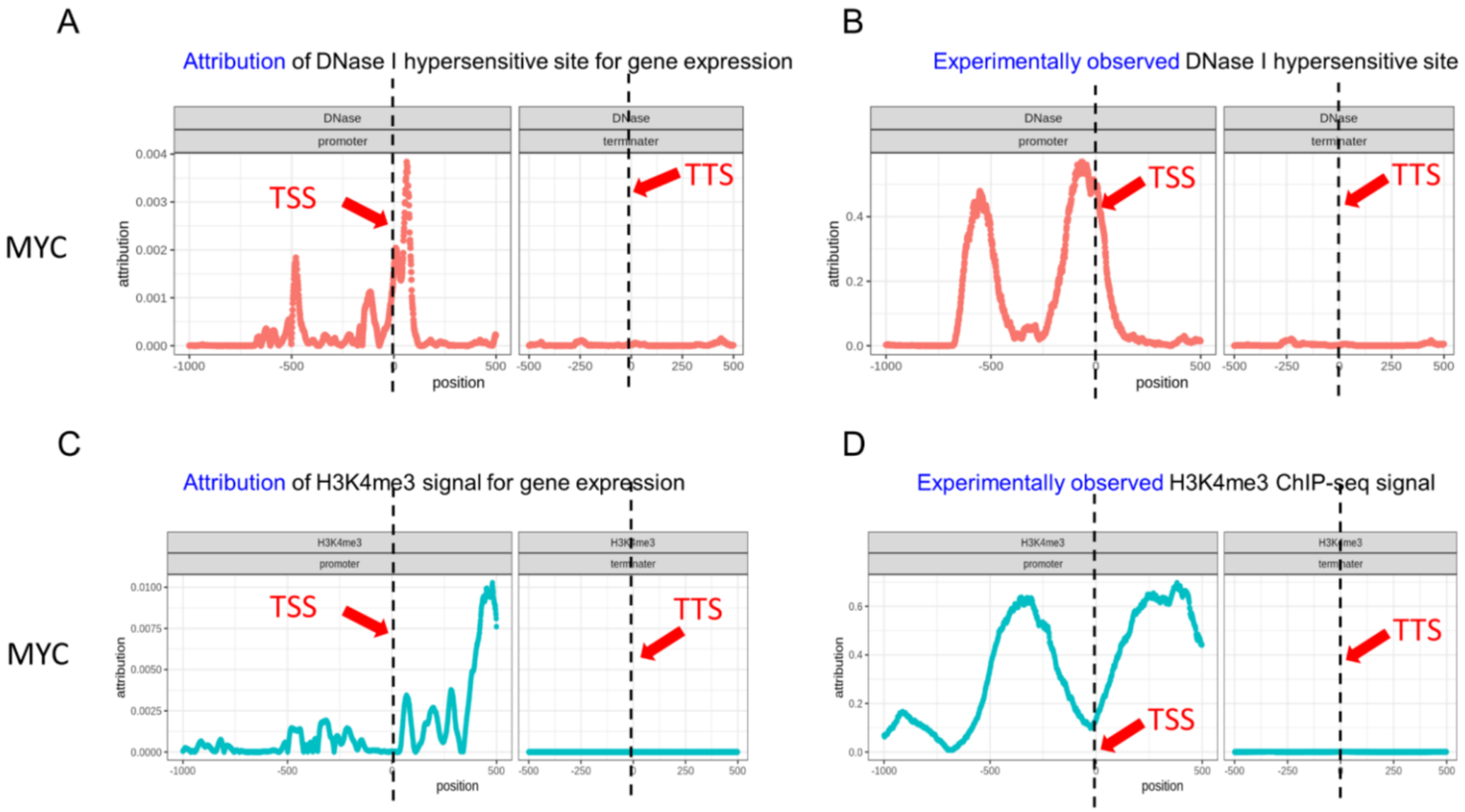

Figure 34. The attributions of DNase I hypersensitive site and H3K4me3 for MYC expression.

4.3.5 The high attribution regions and TF binding have overlaps

We next investigated the possible biological functions of the high prediction attribution regions. We surmised that the transcription factor binding functions might relate to these regions. Therefore, it could explain the reasons why the shifts of epigenetic modifications on these regions influenced the expression of genes nearby.

In order to test our idea, for each epigenetic modification, we first extracted the sequence 100 bps regions around the highest attribution site on each gene. Then we identified the TFs binding motifs that enriched on the extracted regions for every epigenetic modification respectively by AME[189] which belongs to the MEME suite[190]. The transcription factor binding motifs were obtained in the JASPAR database[191]. By doing so, we found the transcription factors that have the potential to bind on these high prediction attribution regions. Then, the same analyses had been conducted on each cell type. The heatmap in Figure 35A, displayed the TF binding motifs that were enriched on high prediction attribution regions across all cell types and epigenetics modifications. In the heatmap, the p-value of enrichment analysis was converted by negative log. We observed that there are transcription factors (e.g., FOXB1 and KLF10) shared by multiple cell types or epigenetic modifications. We also observed that some epigenetic modifications have specific transcription factors. For example, HOXC12, which is a member of homeobox family, was only found enriched for DNA methylation in the high attribution regions.

In the A549 cell line, we further investigated the high attribution regions that were enriched for MYC binding motifs. We analyzed the MYC ChIP-seq data of A549 (ENCODE project, GSM1003607) to confirm whether the high attribution regions with

MYC binding motifs are real MYC binding sites. A significant overlap between the peaks of MYC ChIP-seq signal and the high prediction attribution regions with MYC binding motifs was found (compared to randomly selected regions, p-value < 0.05). For example, on the NIPSNAP2 promoter region, there are attribution peaks of Dnase I hypersensitive sites and H3K4me3 overlapping with the MYC ChIP-seq signal peak sites (Figure 35B). On the promoter region of GUSB, the peak of MYC ChIP-seq signal overlapped with H3K4me1 and H3K9me3 high attribution peaks (Figure 35C). These results demonstrated that the high prediction attribution regions had the possibility to be bound by transcription factors and they might be the regulatory regions of gene expression.

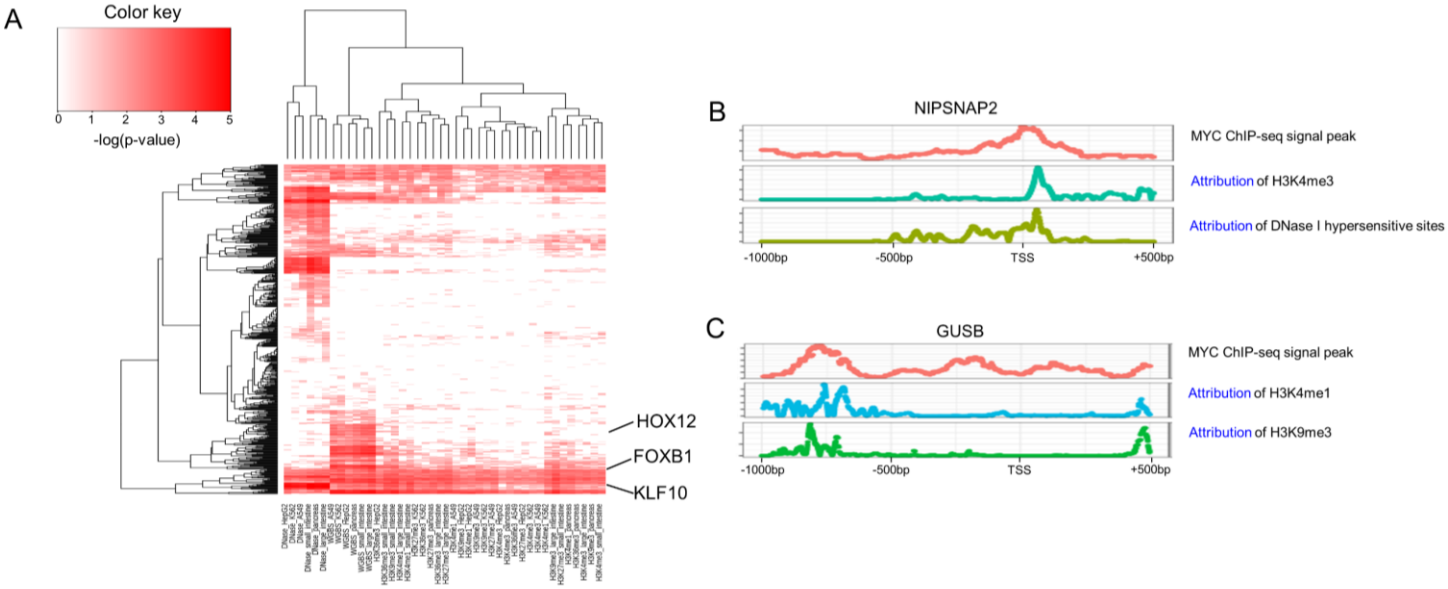

Figure 35. The TF binding motifs that enriched on the high attribution regions and ChIP-seq peak regions.

(**A**) The enrichment analysis of transcription factor binding motifs on high attribution regions. (**B**) (**C**) The epigenetic modification attributions and MYC ChIP-seq signals on the promoter regions of NIPSNAP2 and GUSB.

4.3.6 Case studies

Next, we explored the potential utility of iSEGnet on the data of human disease. iSEGnet was applied on two cancer datasets, one esophageal cancer [48] and one breast cancer [176]. In the esophageal cancer dataset, there are whole-genome bisulfite sequencing data and mRNA-seq data on both normal and tumor tissues from 9 patients. In the breast cancer dataset, there are H3K4me1 ChIP-seq data, H3K4me3 ChIP-seq data, whole-genome bisulfite sequencing data and mRNA-seq data. Two breast cancer cell lines, MCF7 (drug-sensitive) and TAMR (drug-resistant) are involved in this study. Only one sample was accessible for each epigenetic modification.

The esophageal cancer dataset had samples from nine patients. So, we first tested the influence of multiple replicates on the model prediction performance. We trained several models with different number of patients. The model prediction performance, R^2 values increased from 0.38 to 0.82 when we used more replicates in the training. When 5 patients were included in the training, the prediction performance became stable even if more samples were used (Figure 37A). This result indicated that multiple biological replicates in the dataset will benefit the performance of iSEGnet.

In the esophageal cancer dataset, there are normal and tumor tissues. So, we asked the following questions: Are there distinct high prediction attribution regions identified from the iSEGnet models for the differentially expressed genes between two tissues? Moreover, could we reveal the regulatory mechanisms of the expression differences based on the prediction attribution of the regulatory regions? For example, for a specific region of a differentially expressed gene, the regions with significantly different attribution scores between two conditions might indicate that these regions have essential

regulatory functions on the expression of this gene. Based on this assumption, we calculated the prediction attributions for input regions of each differentially expressed gene from normal and tumor data models. A site was defined as a differential attribution site if the mean difference of attributions between the two conditions passed a threshold. The 90% quantile of the difference of attributions on all the input regions was used as this threshold. Then, on every differential attribution site, the transcription factor binding motifs was identified by FIMO [] from the MEME suite []. 74 transcription factors binding motifs were discovered. We also identified the transcription factors which bind to the differentially methylated regions by the same approach. 45 transcription factors were found overlapping between the transcription factors on differential attribution region and the differentially methylated regions (Figure 37B). On the differentially methylated regions, the potential binding transcription factors were enriched for non-cancer pathways (Figure 37C). On the differential attribution regions, the potential binding transcription factors were enriched for cancer related KEGG pathways, e.g., Pathway in cancer (hsa05200) and Transcriptional dysregulation in cancer (hsa05202) (Figure 37D). Between transcription factors binding to differential attribution regions and transcription factors binding to differentially methylated regions, the overlapping transcription factors were also enriched for cancer-related pathways (Figure 37E). These results demonstrate that important regions involved in the gene expression dysregulation in cancer are uncovered by the high prediction attribution regions identified from iSEGnet.

In the breast cancer dataset, there is only one sample in each cell line. The prediction performance of iSEGnet on these two cell lines were in the range of [0.28, 0.32] (R^2 values) and [0.50, 0.63] (Pearson's correlations), respectively (Figure 38). The same

approach was used to reveal the high attribution regions for differentially expressed genes (fold-change based method) between the drug-resistant cell line and drug-sensitive cell lines. The transcription factors binding to these regions were enriched for cancer related KEGG pathways. We also found that there is no significant difference between the observed epigenetic signals on the differential attribution regions. For example, one of the differentially expressed genes, STXBP6, had high observed H3K4me1 signals in both cell lines at the region 500bs upstream of TSS. However, on the same region, the attributions identified from iSEGnet were differentially higher in the drug-sensitive cell line (Figure 38A). Similarly, another differentially expressed gene, BEX2, had differentially high attributions at 300bp to 500bp downstream of the TSS, whereas at the same region there was no significant difference observed in the H3K4me1 ChIP-seq signals (Figure 38B, C). These results demonstrated that iSEGnet can reveal regulatory regions even when there is no significant difference between observed epigenetic signals of two cell lines.

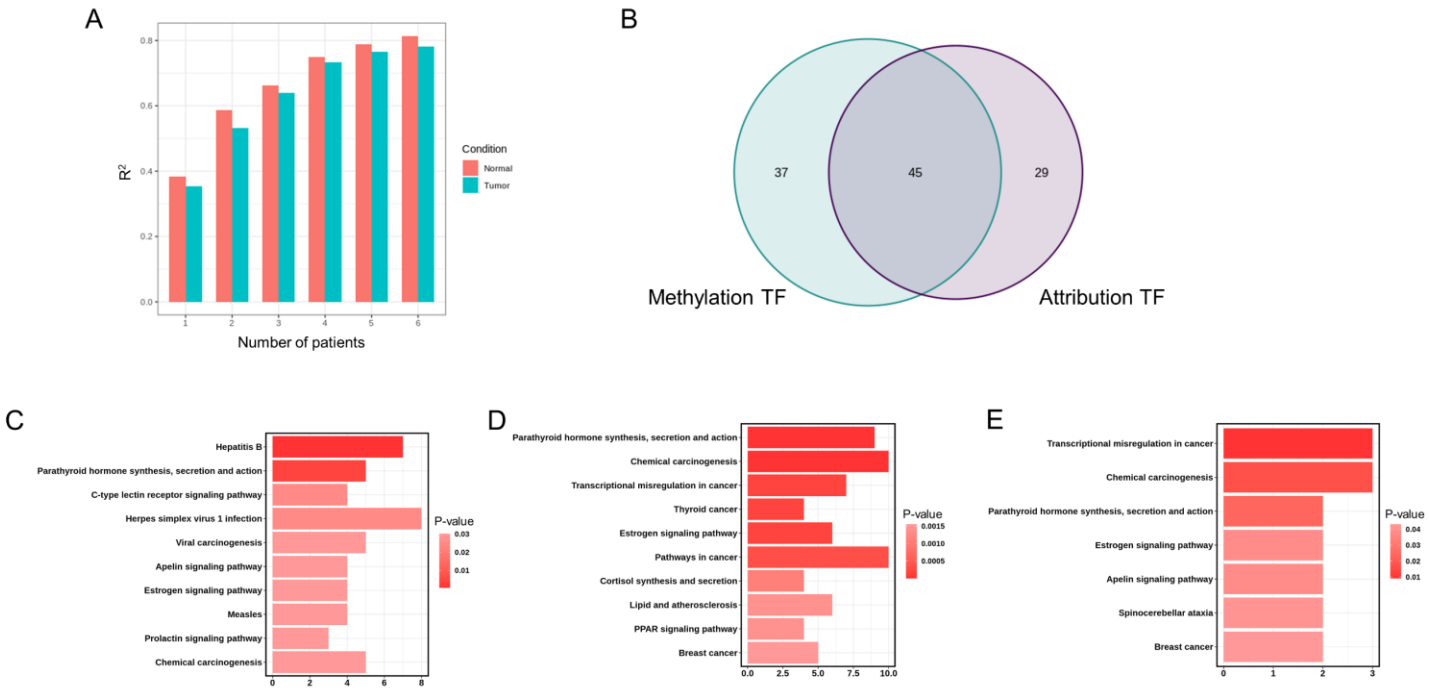

Figure 37. The analysis esophageal cancer dataset with iSEGnet

(**A**) The R^2 values of iSEGnet with a distinct number of samples. (**B**) The overlapping between TF enriched on DMRs and DARs. (**C**) The top KEGG pathways of the TFs on DMRs. (**D**) The top KEGG pathways of the TFs on DARs (**E**) The top overlapped KEGG pathways of the TFs identified from both DMRs and DARs.

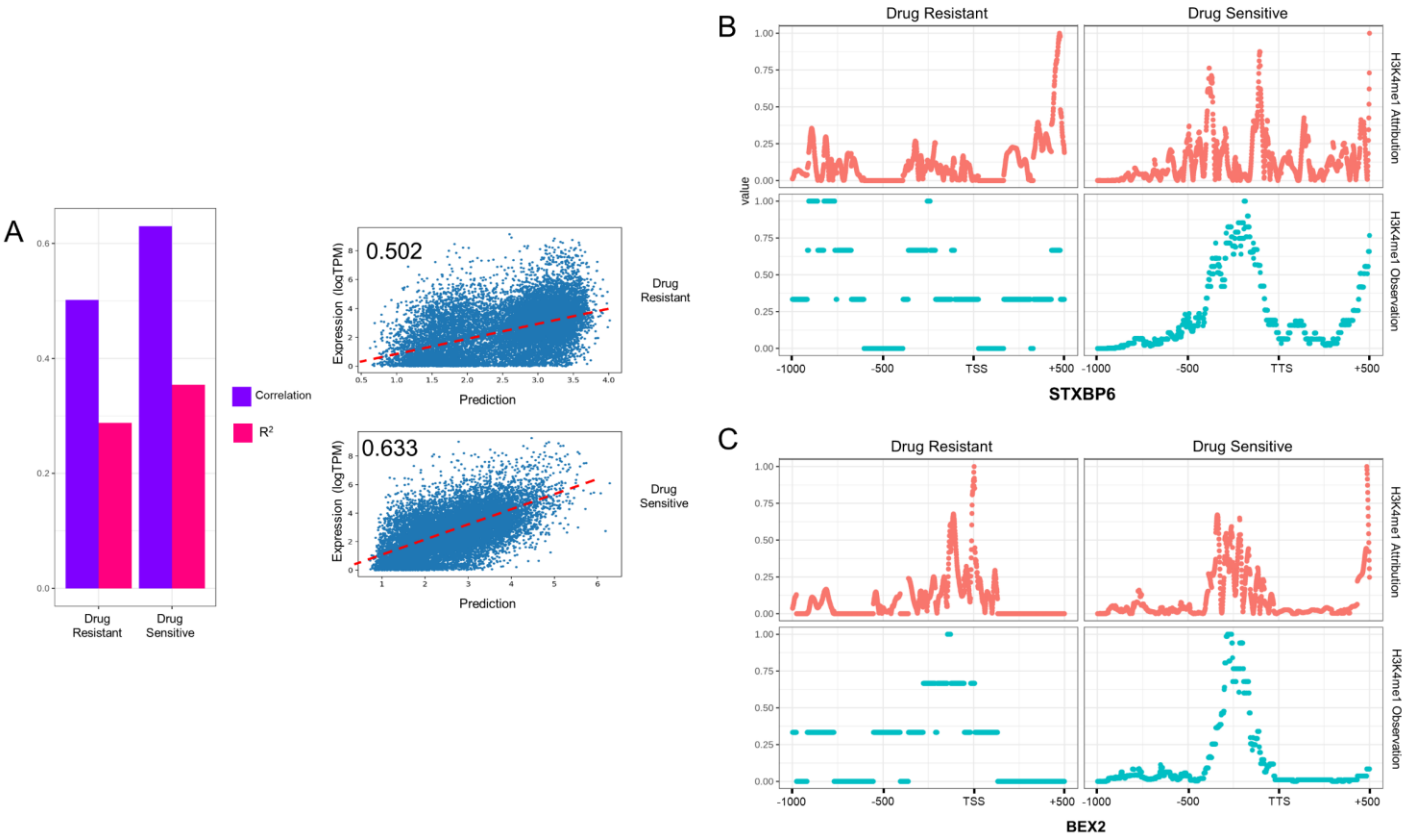

Figure 38. The analysis breast cancer cell line data with iSEGnet

(**A**) The R^2 and Pearson's correlations of iSEGnet on drug-resistant (TAMR) and drug-sensitive (MCF7) breast cancer cell lines. (**B**) (**C**) The attributions of H3K4me1 and the H3K4me1 ChIP-seq signals on the regulatory regions of STXBP6 and BEX2.

4.4 Summary

Both transcriptional and post-transcriptional modifications control gene expression. The primary elements that regulate gene transcription are DNA sequence and epigenetic modifications. It is still difficult to fully understand their intricate relationships and how they affect gene expression in research. We created iSEGnet, a deep learning framework, to predict the gene expression using data on epigenetic changes and DNA sequences on genes and their cis-regulatory areas. We use data from six different cell lines and types selected from the ENCODE to show that our has a better performance than other machine learning approaches in gene expression prediction. The study of learned models also demonstrates that particular areas around transcription start and stop sites are crucial for controlling gene expression. We located the regions in these areas that are most likely to affect gene expression for a particular epigenetic alteration using the Integrated Gradients approach. By contrasting the transcription factor binding sites discovered using ChIP-seq, we also demonstrate that these identified high attribution regions are concentrated in true active regulatory areas. Furthermore, using two cancer multi-omics datasets, we show how iSEGnet can identify putative transcription factors with regulatory roles in cancer.

Chapter 5 Conclusion

5.1 Single cell RNA-seq data analysis to identify the cell heterogeneity during lung injury and regeneration

The results demonstrate cellular heterogeneity in lung endothelial cells during LPS injury and regeneration. Two major EC subpopulations were identified in the lung endothelium, one with a predominant inflammatory signatures and the other with a predominant developmental signature. In the late stage of injury, there was an emergence of a subpopulation that had a predominant proliferation signature. This characterization was achieved by performing scRNA-seq at certain time points during lung inflammatory injury and the subsequent resolution. The existence of distinct EC subpopulations in the lung was validated by RNA-FISH and immunofluorescence. Finally, we inferred the activities of possible regulatory transcription factors in the EC subpopulations from the scRNA-seq datasets.

scRNA-seq has been used to identify the heterogeneity in endothelial cells in different organs. For example, from the single cell data from brain vasculature, 3 major endothelial subpopulations, venous EC, arterial EC and capillary EC, were found. They have essential functions in the migration of myeloid cells [82]. From the single cell study in human lymph nodes, 6 subpopulations of lymphatic ECs were revealed, which located in different regions of lymph nodes [84]. In our study, we only used the endothelial cells from lung microvessels (diameters < 100 μm) and obtained the transcriptomic profiles in these cells. Therefore, we didn't have arterial or venous EC subpopulations in our data. Microvascular endothelial cells are the major cell type in lung endothelium.

In future studies, spatial single cell RNA-seq could be used at different time points

during lung injury and regeneration. Spatial single cell RNA-seq is emerging as an approach to identify the relationship between cells and their locations. Using spatial single cell RNA-seq, one could identify the location of subpopulations in the lung microvessels. Are the cells from same subpopulation grouping together in the lung or are they mixed with different subpopulations? Is one particular subpopulation located in a certain area or are the positions of these cells randomly distributed?

5.2 The Bayesian Inference Transcription Factor Activities Model

We built a method to infer TF activities by integrating existing ChIP-seq data on transcription factor binding and scRNA-seq data. The Bayesian Inference Transcription Factor Activity Model (BITFAM) combines transcription factor ChIP-seq data with binding information and single cell RNA-seq data. We demonstrate that the inferred transcription factor activities from BITFAM correspond to the known functions of the cell types. This inferred transcription factor activity profile can also be used to identify cell heterogeneity as well as other downstream analysis. The inferred regulatory strengths between transcription factors and target genes also match the biological functions of the transcription factor.

We used selected transcription factors to validate the accuracy of BITFAM on inferring the transcription factor activities. These selected transcription factors had known biological functions in particular cell types. In future applications of BITFAM, one could utilize BITFAM to learn the activities of transcription factors which are not yet established. This will uncover the biological functions of these transcription factors in different cell types. BITFAM could also be used to generate biological hypotheses about the regulatory

mechanisms in certain cells. These could be experimentally tested by targeted knockdown or deletion in cells or animal models.

Importantly, the mRNA levels of transcription factors in single cell RNA-seq data did not match their known regulatory functions in the specific cells. The possible reasons for this are 1) the dropout events in single cell RNA sequencing led to missing detection of transcription factor mRNA, 2) the post-translation modification on the transcription factors influenced their activities and don't correlate to mRNA levels. This means that relying on mRNA levels of TFs is inadequate to identify cell type specific TFs in scRNA-seq data. Instead, BITFAM can infer the transcription factors activities in cells even when there was minimal expression of the TFs because BITFAM uses the aggregate of the TF targets and not the levels of TFs themselves. The levels of the TF targets are a much better reflection of the actual TF activities than the mRNA of the TFs themselves because TF activity is often regulated post-translationally by phosphorylation which would impact the TF target mRNA levels but not the TF mRNA level.

There are numerous ChIP-seq datasets available for each individual transcription factor (depends on how many labs are studying that transcription factor). This leads to varying levels of quality and quantity of the necessary prior knowledge in a Bayesian inference model. It is also unclear if ChIP-seq data collected from one cell type is as meaningful for investigation with a scRNA-seq derived from another cell or tissue type since current ChIP-seq data are produced from bulk cells or tissues. The limitations of using bulk ChIP-seq data generated in various cell types to a specific single cell RNA-seq data which generated from a new cell type might be partially solved by scATAC-seq data. scATAC-seq could establish the context-specific single cell chromatin accessibility and

provide information about transcription factor binding regions and its target genes. One could improve BITFAM by adding these scATAC-seq data into the prior knowledge of the Bayesian hierarchical model. In the future version of BITFAM, we can integrate scATAC-seq data into the model, therefore provide more accurate transcription factors target genes by intersecting ChIP-seq targets with the open chromatin regions. By doing so, the quality of the inference could be improved

5.3 Assessing Importance of DNA Sequence and Epigenetics Profiles for Gene expression using a Deep Neural Network

We developed iSEGnet, a deep learning framework that predicts mRNA abundance using DNA sequences and epigenetic modifications on gene regulatory regions. The input regions for iSEGnet that generate the best prediction performance are the combination of 1000bps upstream and 500bps downstream around TSSs and 500bps upstream and 500bps downstream around TTSs. We used data from six cell lines/types in the ENCODE project to validate that iSEGnet has better performance than other machine learning models, such as random forest and SVM. We discovered the epigenetic modifications and the regulatory areas that are crucial for gene expression of certain genes using Integrated Gradients. By analyzing transcription factor binding motifs enrichment and the overlap with the transcription factor ChIP-seq peak regions, we demonstrated that these regions might have regulatory functions. In order to show that iSEGnet could be used to pinpoint particular regulatory areas related to genes that had differential expression across various conditions, such as tumor and normal tissues, we lastly applied it to two cancer multi-omics datasets. In order to identify significant

transcription factors and regulatory areas that could affect gene expression under various circumstances, iSEGnet integrates multi-omics of few replicates.

The common transcription regulation patterns across genes under a certain situation or in a particular cell type are learned by iSEGnet using one copy of omics profiles because all expressed genes are considered as learning samples. Although regulatory areas and epigenetic modifications differ across genes, iSEGnet may identify patterns from a variety of hidden layers and kernels in deep CNNs that are beneficial for forecasting gene expression. However, there is a chance of losing information on the connections between various regulatory area parts under the existing CNN design. Recurrent neural networks might be used to overcome this restriction (RNN). RNN is often used to represent sequential data in applications like time-series data analysis and natural language processing. A promising architecture for predicting biological contexts from sequential data, such as DNA methylation and chromatin accessibility, has been shown using the combination of RNN and CNN.

	baseline	LPS 6 hours	LPS 24 hours	LPS 2 days	LPS 3 days	LPS 7 days
Cell number	8,191	6,527	5,158	4,608	5,318	6,171
Median gene number per cell	1,568	1,147	1,252	1,770	1,897.5	1,872
Total gene number	17,035	16,409	16,069	16,501	16,664	16,966
UMI* number per cell	3,003	2,273	2,544.5	3,788	4,289	3,694
Total UMI* number	29,013,732	16,689,633	15,099,705	19,572,952	27,346,395	25,398,294
Cell number of inflamEC	3460	2759	2278	1531	-	3083
Median gene number of inflamEC	1,502	1,090	1,229.5	1,758	-	1,884
Total gene number of inflamEC	15,280	14,730	14,649	14,638	-	15,499
Cell number of devEC	2206	2601	1561	1424	-	1683
Median gene number of devEC	1,567.5	1,154	1,242	1,723.5	-	1,903
Total gene number of devEC	15,142	14,861	14,059	14,650	-	15,234
Cell number of devEC	-	-	-	-	635	-
Median gene number of proEC	-	-	-	-	2,951	-
Total gene number of proEC	-	-	-	-	14,265	-

Table 1. Information of scRNA-seq of lung endothelial cells at different time points.

	Tabula Muris lung	*Tabula Muris* heart	*Tabula Muris* brain	Blood cell development	CRISPRi
Number of cells	5449	4365	3401	2730	5174
Number of most variable expressed genes	4552	5337	5617	1830	4844
Number of transcription factors	106	87	105	106	103

Table 2. The description of the datasets used in BITFAM.

CITED LITERATURE

1. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights* **14**, 1177932219899051 (2020).
2. Lee, J., Hyeon, D.Y. & Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med* **52**, 1428-1442 (2020).
3. Wu, C. et al. A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. *High Throughput* **8** (2019).
4. Roth, S.C. What is genomic medicine? *J Med Libr Assoc* **107**, 442-448 (2019).
5. Ng, P.C. & Kirkness, E.F. Whole genome sequencing. *Methods Mol Biol* **628**, 215-226 (2010).
6. Sjoblom, T. et al. The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-274 (2006).
7. Estivill, X. & Armengol, L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet* **3**, 1787-1799 (2007).
8. Waddington, C.H. The epigenotype. 1942. *Int J Epidemiol* **41**, 10-13 (2012).
9. Stricker, S.H., Koferle, A. & Beck, S. From profiles to function in epigenomics. *Nat Rev Genet* **18**, 51-66 (2017).
10. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics* **33**, 245-254 (2003).
11. Kornberg, R.D. & Thomas, J.O. Chromatin structure; oligomers of the histones. *Science* **184**, 865-868 (1974).
12. Kaplan, N. et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362-366 (2009).
13. Klemm, S.L., Shipony, Z. & Greenleaf, W.J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**, 207-220 (2019).
14. Song, L. & Crawford, G.E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* **2010**, pdb.prot5384 (2010).
15. Mieczkowski, J. et al. MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nature Communications* **7**, 11485 (2016).
16. Yan, F., Powell, D.R., Curtis, D.J. & Wong, N.C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* **21**, 22 (2020).
17. Cusanovich, D.A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914 (2015).
18. Bannister, A.J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res* **21**, 381-395 (2011).
19. Allfrey, V.G., Faulkner, R. & Mirsky, A.E. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS*. *Proceedings of the National Academy of Sciences* **51**, 786-794 (1964).
20. Jambhekar, A., Dhall, A. & Shi, Y. Roles and regulation of histone methylation in animal development. *Nat Rev Mol Cell Biol* **20**, 625-641 (2019).
21. Park, P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669-680 (2009).
22. Furey, T.S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet* **13**, 840-852 (2012).
23. Kharchenko, P.V., Tolstorukov, M.Y. & Park, P.J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26**, 1351-1359 (2008).

24. Davis, C.A. et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**, D794-D801 (2018).
25. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
26. Valouev, A. et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5**, 829-834 (2008).
27. Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y. & Kolpakov, F. GTRD: a database on gene transcription regulation-2019 update. *Nucleic Acids Res* **47**, D100-D105 (2019).
28. Robertson, K.D. DNA methylation and human disease. *Nat Rev Genet* **6**, 597-610 (2005).
29. Zemach, A., McDaniel, I.E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916-919 (2010).
30. Greenberg, M.V.C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* **20**, 590-607 (2019).
31. Feng, S. et al. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* **107**, 8689-8694 (2010).
32. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322 (2009).
33. Zhou, L. et al. Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. *Sci Rep* **9**, 10383 (2019).
34. Li, Y. & Tollefsbol, T.O. DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol Biol* **791**, 11-21 (2011).
35. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**, 1509-1517 (2008).
36. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).
37. Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349 (2008).
38. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat Rev Genet* **20**, 631-656 (2019).
39. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol* **18**, 83 (2017).
40. Yan, J., Risacher, S.L., Shen, L. & Saykin, A.J. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief Bioinform* **19**, 1370-1381 (2018).
41. Bersanelli, M. et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* **17 Suppl 2**, 15 (2016).
42. Chapman-Rothe, N. et al. Chromatin H3K27me3/H3K4me3 histone marks define gene sets in high-grade serous ovarian cancer that distinguish malignant, tumour-sustaining and chemo-resistant ovarian tumour cells. *Oncogene* **32**, 4586-4592 (2013).
43. Gopi, L.K. & Kidder, B.L. Integrative pan cancer analysis reveals epigenomic variation in cancer type and cell specific chromatin domains. *Nat Commun* **12**, 1419 (2021).
44. Ampuja, M. et al. Integrated RNA-seq and DNase-seq analyses identify phenotype-specific BMP4 signaling in breast cancer. *BMC Genomics* **18**, 68 (2017).
45. Wei, X. et al. Identification of open chromosomal regions and key genes in prostate cancer via integrated analysis of DNaseseq and RNAseq data. *Mol Med Rep* **18**, 2245-2252 (2018).
46. Long, M.D. et al. Dynamic patterns of DNA methylation in the normal prostate epithelial differentiation program are targets of aberrant methylation in prostate cancer. *Sci Rep* **11**, 11405 (2021).

47. Zheng, W. et al. Regnase-1 suppresses TCF-1+ precursor exhausted T-cell formation to limit CAR-T-cell responses against ALL. *Blood* **138**, 122-135 (2021).
48. Cao, W. et al. Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma. *Nat Commun* **11**, 3675 (2020).
49. Nica, A.C. & Dermitzakis, E.T. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci* **368**, 20120362 (2013).
50. Zhao, T., Hu, Y., Zang, T. & Wang, Y. Integrate GWAS, eQTL, and mQTL Data to Identify Alzheimer's Disease-Related Genes. *Front Genet* **10**, 1021 (2019).
51. Patel, D. et al. Cell-type-specific expression quantitative trait loci associated with Alzheimer disease in blood and brain tissue. *Transl Psychiatry* **11**, 250 (2021).
52. Taavitsainen, S. et al. Single-cell ATAC and RNA sequencing reveal pre-existing and persistent cells associated with prostate cancer relapse. *Nat Commun* **12**, 5307 (2021).
53. Jia, G. et al. Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nat Commun* **9**, 4877 (2018).
54. Li, S. et al. Epigenetic Landscapes of Single-Cell Chromatin Accessibility and Transcriptomic Immune Profiles of T Cells in COVID-19 Patients. *Front Immunol* **12**, 625881 (2021).
55. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865-868 (2017).
56. Pucher, B.M., Zeleznik, O.A. & Thallinger, G.G. Comparison and evaluation of integrative methods for the analysis of multilevel omics data: a study based on simulated and experimental cancer data. *Brief Bioinform* **20**, 671-681 (2019).
57. Zhang, S. et al. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res* **40**, 9379-9391 (2012).
58. Shen, R., Olshen, A.B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906-2912 (2009).
59. Ray, P., Zheng, L., Lucas, J. & Carin, L. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics* **30**, 1370-1376 (2014).
60. Lock, E.F., Hoadley, K.A., Marron, J.S. & Nobel, A.B. Joint and Individual Variation Explained (Jive) for Integrated Analysis of Multiple Data Types. *Ann Appl Stat* **7**, 523-542 (2013).
61. Lanckriet, G.R., De Bie, T., Cristianini, N., Jordan, M.I. & Noble, W.S. A statistical framework for genomic data fusion. *Bioinformatics* **20**, 2626-2635 (2004).
62. Kim, S., Jhong, J.H., Lee, J. & Koo, J.Y. Meta-analytic support vector machine for integrating multiple omics data. *BioData Min* **10**, 2 (2017).
63. Stetson, L.C., Pearl, T., Chen, Y. & Barnholtz-Sloan, J.S. Computational identification of multi-omic correlates of anticancer therapeutic response. *BMC Genomics* **15 Suppl 7**, S2 (2014).
64. Ma, B. et al. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in Biology and Medicine* **121**, 103761 (2020).
65. Chaudhary, K., Poirion, O.B., Lu, L. & Garmire, L.X. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin Cancer Res* **24**, 1248-1259 (2018).
66. Sharifi-Noghabi, H., Zolotareva, O., Collins, C.C. & Ester, M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* **35**, i501-i509 (2019).
67. Xu, J. et al. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. *BMC Bioinformatics* **20**, 527 (2019).
68. Chung, R.H. & Kang, C.Y. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. *Gigascience* **8** (2019).

69. Holzinger, E.R., Dudek, S.M., Frase, A.T., Pendergrass, S.A. & Ritchie, M.D. ATHENA: the analysis tool for heritable and environmental network associations. *Bioinformatics* **30**, 698-705 (2014).
70. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
71. Welch, J.D. et al. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887 e1817 (2019).
72. Argelaguet, R. et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* **14**, e8124 (2018).
73. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
74. Consortium, I.T.P.-C.A.o.W.G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020).
75. Schlomm, T. [Results of the CGC/TCGA Pan-Cancer Analysis of the Whole Genomes (PCAWG) Consortium]. *Urologe A* **59**, 1552-1553 (2020).
76. Liao, J.K. Linking endothelial dysfunction with endothelial cell activation. *J Clin Invest* **123**, 540-541 (2013).
77. Jambusaria, A. et al. Endothelial heterogeneity across distinct vascular beds during homeostasis and inflammation. *Elife* **9** (2020).
78. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. & Teichmann, S.A. The technology and biology of single-cell RNA sequencing. *Mol Cell* **58**, 610-620 (2015).
79. Jordao, M.J.C. et al. Single-cell profiling identifies myeloid cell subsets with distinct fates during neuroinflammation. *Science* **363** (2019).
80. Farbehi, N. et al. Single-cell expression profiling reveals dynamic flux of cardiac stromal, vascular and immune cells in health and injury. *Elife* **8** (2019).
81. Schyns, J. et al. Non-classical tissue monocytes and two functionally distinct populations of interstitial macrophages populate the mouse lung. *Nat Commun* **10**, 3964 (2019).
82. Vanlandewijck, M. et al. A molecular atlas of cell types and zonation in the brain vasculature. *Nature* **554**, 475-480 (2018).
83. Lukowski, S.W. et al. Single-Cell Transcriptional Profiling of Aortic Endothelium Identifies a Hierarchy from Endovascular Progenitors to Differentiated Cells. *Cell Rep* **27**, 2748-2758 e2743 (2019).
84. Takeda, A. et al. Single-Cell Survey of Human Lymphatics Unveils Marked Endothelial Cell Heterogeneity and Mechanisms of Homing for Neutrophils. *Immunity* **51**, 561-572 e565 (2019).
85. Vila Ellis, L. et al. Epithelial Vegfa Specifies a Distinct Endothelial Population in the Mouse Lung. *Dev Cell* **52**, 617-630 e616 (2020).
86. Matthay, M.A., Ware, L.B. & Zimmerman, G.A. The acute respiratory distress syndrome. *J Clin Invest* **122**, 2731-2740 (2012).
87. Liu, M. et al. Sox17 is required for endothelial regeneration following inflammation-induced vascular injury. *Nat Commun* **10**, 2126 (2019).
88. Riemondy, K.A. et al. Single cell RNA sequencing identifies TGFbeta as a key regenerative cue following LPS-induced lung injury. *JCI Insight* **5** (2019).
89. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083-1086 (2017).
90. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).
91. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).
92. Maizels, R.M. & Withers, D.R. MHC-II: a mutual support system for ILCs and T cells? *Immunity* **41**, 174-176 (2014).

93. Oliphant, C.J. et al. MHCII-mediated dialog between group 2 innate lymphoid cells and CD4(+) T cells potentiates type 2 immunity and promotes parasitic helminth expulsion. *Immunity* **41**, 283-295 (2014).
94. Park, C., Kim, T.M. & Malik, A.B. Transcriptional regulation of endothelial cell and vascular development. *Circ Res* **112**, 1380-1400 (2013).
95. Moreno-Layseca, P., Icha, J., Hamidi, H. & Ivaska, J. Integrin trafficking in cells and tissues. *Nat Cell Biol* **21**, 122-132 (2019).
96. Nishikawa, K., Ayukawa, K., Hara, Y., Wada, K. & Aoki, S. Endothelin/endothelin-B receptor signals regulate ventricle-directed interkinetic nuclear migration of cerebral cortical neural progenitors. *Neurochem Int* **58**, 261-272 (2011).
97. Gillich, A. et al. Capillary cell-type specialization in the alveolus. *Nature* **586**, 785-789 (2020).
98. Travaglini, K.J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619-625 (2020).
99. Chen, G.Y., Tang, J., Zheng, P. & Liu, Y. CD24 and Siglec-10 selectively repress tissue damage-induced immune responses. *Science* **323**, 1722-1725 (2009).
100. Sugita, S. et al. Acquisition of T regulatory function in cathepsin L-inhibited T cells by eye-derived CTLA-2alpha during inflammatory conditions. *J Immunol* **183**, 5013-5022 (2009).
101. Werner, S.L. et al. Encoding NF-kappaB temporal control in response to TNF: distinct roles for the negative regulators IkappaBalpha and A20. *Genes Dev* **22**, 2093-2101 (2008).
102. Gao, S., Dai, Y. & Rehman, J. A Bayesian inference transcription factor activity model for the analysis of single-cell transcriptomes. *Genome Res* (2021).
103. Tabula Muris, C. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372 (2018).
104. Zheng, G.X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
105. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* **14**, 618-630 (2013).
106. Regev, A. et al. The Human Cell Atlas. *Elife* **6** (2017).
107. Svensson, V., Vento-Tormo, R. & Teichmann, S.A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* **13**, 599-604 (2018).
108. Zeisel, A. et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142 (2015).
109. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput Biol* **14**, e1006245 (2018).
110. Wu, Y., Tamayo, P. & Zhang, K. Visualizing and Interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding. *Cell Syst* **7**, 656-666 e654 (2018).
111. Amir el, A.D. et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* **31**, 545-552 (2013).
112. Wang, D. & Gu, J. VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder. *Genomics Proteomics Bioinformatics* **16**, 320-331 (2018).
113. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).
114. Zurauskiene, J. & Yau, C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* **17**, 140 (2016).
115. Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S. & Kluger, Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods* **16**, 243-245 (2019).
116. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053-1058 (2018).

117. Duren, Z. et al. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc Natl Acad Sci U S A* **115**, 7723-7728 (2018).
118. Moon, K.R. et al. Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* **37**, 1482-1492 (2019).
119. Jung, M. et al. Unified single-cell analysis of testis gene regulation and pathology in five mouse strains. *Elife* **8** (2019).
120. Kiselev, V.Y., Andrews, T.S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* **20**, 273-282 (2019).
121. Lopes, H.F. & West, M. BAYESIAN MODEL ASSESSMENT IN FACTOR ANALYSIS. *Statistica Sinica* **14**, 41-67 (2004).
122. MacKay, D.J.C. in Maximum Entropy and Bayesian Methods: Santa Barbara, California, U.S.A., 1993. (ed. G.R. Heidbreder) 221-234 (Springer Netherlands, Dordrecht; 1996).
123. Wainwright, M.J. & Jordan, M.I. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning* **1**, 1-305 (2008).
124. Ghahramani, Z. & Matthew, J.B. Variational Inference for Bayesian Mixtures of Factor Analysers. 449--455 (2000).
125. Team, R.C. R: A Language and Environment for Statistical Computing. (2020).
126. Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D.M. Automatic differentiation variational inference. *J. Mach. Learn. Res.* **18**, 430–474 (2017).
127. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-496 (2004).
128. Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, 10008 (2008).
129. Angerer, P. et al. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241-1243 (2016).
130. Hubert, L. & Arabie, P. Comparing partitions. *Journal of Classification* **2**, 193-218 (1985).
131. Huang da, W., Sherman, B.T. & Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009).
132. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018 (2011).
133. Kulakovskiy, I.V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* **46**, D252-D259 (2018).
134. Fornes, O. et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **48**, D87-D92 (2020).
135. Bai, J. & Li, K. Statistical analysis of factor models of high dimension. *The Annals of Statistics* **40**, 436-465 (2012).
136. Buettner, F., Pratanwanich, N., McCarthy, D.J., Marioni, J.C. & Stegle, O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol* **18**, 212 (2017).
137. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**, 1724-1735 (2007).
138. Hand, D. Latent Variable Models and Factor Analysis: A Unified Approach, Third Edition by David J. Bartholomew, Martin Knott, Irini Moustaki, Vol. 81. (2013).
139. Paul, F. et al. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **164**, 325 (2016).
140. Genga, R.M.J. et al. Single-Cell RNA-Sequencing-Based CRISPRi Screening Resolves Molecular Drivers of Early Human Endoderm Development. *Cell Rep* **27**, 708-718 e710 (2019).

141. De Val, S. & Black, B.L. Transcriptional control of endothelial cell development. *Dev Cell* **16**, 180-195 (2009).
142. Lazrak, M. et al. The bHLH TAL-1/SCL regulates endothelial cell migration and morphogenesis. *J Cell Sci* **117**, 1161-1171 (2004).
143. Nutt, S.L., Heavey, B., Rolink, A.G. & Busslinger, M. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature* **401**, 556-562 (1999).
144. Wu, X. et al. Mafb lineage tracing to distinguish macrophages from other immune lineages reveals dual identity of Langerhans cells. *J Exp Med* **213**, 2553-2565 (2016).
145. Takahashi, S. et al. Role of GATA-1 in proliferation and differentiation of definitive erythroid and megakaryocytic cells in vivo. *Blood* **92**, 434-442 (1998).
146. Kimura, A. et al. The transcription factors STAT5A/B regulate GM-CSF-mediated granulopoiesis. *Blood* **114**, 4721-4728 (2009).
147. Suh, H.C. et al. C/EBPalpha determines hematopoietic cell fate in multipotential progenitor cells by inhibiting erythroid differentiation and inducing myeloid differentiation. *Blood* **107**, 4308-4316 (2006).
148. Lee, S. et al. Carbohydrate-binding protein CLEC14A regulates VEGFR-2- and VEGFR-3-dependent signals during angiogenesis and lymphangiogenesis. *J Clin Invest* **127**, 457-471 (2017).
149. Liu, G.J. et al. Pax5 loss imposes a reversible differentiation block in B-progenitor acute lymphoblastic leukemia. *Genes Dev* **28**, 1337-1350 (2014).
150. Gheorghe, M. et al. A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Res* **47**, 7715 (2019).
151. Cheneby, J. et al. ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res* **48**, D180-D188 (2020).
152. Baker, R.G., Hayden, M.S. & Ghosh, S. NF-kappaB, inflammation, and metabolic disease. *Cell Metab* **13**, 11-22 (2011).
153. Wang, B. et al. SIMLR: A Tool for Large-Scale Genomic Analyses by Multi-Kernel Learning. *Proteomics* **18** (2018).
154. Kiselev, V.Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14**, 483-486 (2017).
155. Li, E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet* **3**, 662-673 (2002).
156. Santos-Rosa, H. et al. Methylation of Histone H3 K4 Mediates Association of the Isw1p ATPase with Chromatin. *Mol Cell* **70**, 983 (2018).
157. Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467-484 (2019).
158. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
159. Zhang, Z. et al. Deep learning in omics: a survey and guideline. *Brief Funct Genomics* **18**, 41-57 (2019).
160. Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* **15** (2018).
161. Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107**, 2926-2931 (2010).
162. Dong, X. et al. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology* **13**, R53 (2012).
163. Cheng, C. et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biology* **12**, R15 (2011).

164. Singh, R., Lanchantin, J., Robins, G. & Qi, Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* **32**, i639-i648 (2016).
165. Koumakis, L. Deep learning models in genomics; are we there yet? *Comput Struct Biotechnol J* **18**, 1466-1473 (2020).
166. Zhang, S. et al. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res* **44**, e32 (2016).
167. Alquicira-Hernandez, J., Sathe, A., Ji, H.P., Nguyen, Q. & Powell, J.E. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol* **20**, 264 (2019).
168. Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* **50**, 1171-1179 (2018).
169. Avsec, Z. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196-1203 (2021).
170. Kelley, D.R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* **28**, 739-750 (2018).
171. Agarwal, V. & Shendure, J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep* **31**, 107663 (2020).
172. Kelley, D.R. Cross-species regulatory sequence activity prediction. *PLoS Comput Biol* **16**, e1008050 (2020).
173. Sekhon, A., Singh, R. & Qi, Y. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. *Bioinformatics* **34**, i891-i900 (2018).
174. Zeng, W., Wang, Y. & Jiang, R. Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. *Bioinformatics* **36**, 496-503 (2020).
175. Gao, S., Rehman, J. & Dai, Y. Assessing comparative importance of DNA sequence and epigenetic modifications on gene expression using a deep convolutional neural network. *Comput Struct Biotechnol J* **20**, 3814-3823 (2022).
176. Achinger-Kawecka, J. et al. Epigenetic reprogramming at estrogen-receptor binding sites alters 3D chromatin landscape in endocrine-resistant breast cancer. *Nature Communications* **11**, 320 (2020).
177. Abadi, M. et al. in Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation 265–283 (USENIX Association, Savannah, GA, USA; 2016).
178. Dillon, J.V. et al. arXiv:1711.10604 (2017).
179. Kingma, D.P. & Ba, J. (2015).
180. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
181. Buitinck, L. et al. arXiv:1309.0238 (2013).
182. Sundararajan, M., Taly, A. & Yan, Q. in Proceedings of the 34th International Conference on Machine Learning - Volume 70 3319–3328 (JMLR.org, Sydney, NSW, Australia; 2017).
183. Klaise, J., Van Looveren, A., Vacanti, G. & Coca, A. Alibi Explain: algorithms for explaining machine learning models. *Journal of Machine Learning Research* **22**, 1-7 (2021).
184. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* **49**, D545-D551 (2021).
185. Chen, E.Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
186. Kuleshov, M.V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-97 (2016).
187. Xie, Z. et al. Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **1**, e90 (2021).

188. Chen, H., Liu, H. & Qing, G. Targeting oncogenic Myc as a strategy for cancer treatment. *Signal Transduction and Targeted Therapy* **3**, 5 (2018).
189. McLeay, R.C. & Bailey, T.L. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11**, 165 (2010).
190. Bailey, T.L., Johnson, J., Grant, C.E. & Noble, W.S. The MEME Suite. *Nucleic Acids Res* **43**, W39-49 (2015).
191. Castro-Mondragon, J.A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **50**, D165-D173 (2022).

APPENDIX: Copyright Permission

JCI Insight

Open access

JCI Insight is a Gold Open Access journal in which all content is freely available without charge to the user or their institution. Effective with the August 20, 2020, issue, all content is published with a Creative Commons Attribution License (CC BY 4.0).

The ASCI retains copyright to many articles published prior to August 20, 2020. While all content is freely available, some use of articles to which the ASCI holds copyright may require permission. (See “Obtaining permission to use JCI Insight content” below.) Copyright or license information is noted on each article.

JCI Insight deposits all content in PubMed and PubMed Central. Authors of articles published in JCI Insight do not need to deposit their articles separately. The Journal meets cOAlition S requirements for authors who receive funding support from Plan S signatories and is fully compliant with Howard Hughes Medical Institute (HHMI), Wellcome, and UK Research and Innovation (UKRI) open access policies; and is indexed in Securing a Hybrid Environment for Research Preservation and Access (Sherpa Romeo) and included in the Directory of Open Access Journals (DOAJ).

Obtaining permission to use JCI Insight content

Articles published on or after August 20, 2020, are published with the CC BY 4.0 license, which allows for reuse without permission as long as the original source is cited. Permission may be required for use of content published prior to August 20, 2020, where the ASCI retains copyright. Permission can be obtained via Copyright Clearance Center. Copyright or license information is noted on each article.

Genome Research

Elsevier, Computational and Structural Biology Journal

Author rights

The below table explains the rights that authors have when they publish with Elsevier, for authors who choose to publish either open access or subscription. These apply to the corresponding author and all co-authors.

Author rights in Elsevier's proprietary journals	Published open access	Published subscription
Retain patent and trademark rights	√	√
Retain the rights to use their research data freely without any restriction	√	√
Receive proper attribution and credit for their published work	√	√
Re-use their own material in new works without permission or payment (with full acknowledgement of the original article): 1. Extend an article to book length 2. Include an article in a subsequent compilation of their own work 3. Re-use portions, excerpts, and their own figures or tables in other works.	√	√
Use and share their works for scholarly purposes (with full acknowledgement of the original article): 1. In their own classroom teaching. Electronic and physical distribution of copies is permitted 2. If an author is speaking at a conference, they can present the article and distribute copies to the attendees 3. Distribute the article, including by email, to their students and to research colleagues who they know for their personal use 4. Share and publicize the article via Share Links, which offers 50 days' free access for anyone, without signup or registration 5. Include in a thesis or dissertation (provided this is not published commercially) 6. Share copies of their article privately as part of an invitation-only work group on commercial sites with which the publisher has a hosting agreement	√	√
Publicly share the preprint on any website or repository at any time.	√	√
Publicly share the accepted manuscript on non-commercial sites	√	√ using a CC BY-NC-ND license and usually only after an embargo period (see Sharing Policy for more information)
Publicly share the final published article	√ in line with the author's choice of end user license	×
Retain copyright	√	×

VITA

Shang Gao, Ph.D.

Education

Ph.D. Candidate: Bioinformatics, 09/2016 to 08/2022

University of Illinois at Chicago - Chicago, IL

Projects:

- **A Bayesian Inference Transcription Factor Activity Model for the Analysis of Single Cell Transcriptomes.**
- **Assessing Importance of DNA Sequence and Epigenetics Profiles for Gene expression using a Deep Convolutional Neural Network.**
- **Single-cell transcriptomic profiling of lung endothelial cells identifies dynamic inflammatory and regenerative subpopulations.**

Master of Engineering: Bioengineering, 09/2012 to 06/2014

Nankai University - Tianjin, China

- Thesis: Study on the Mechanism of Cucumber Male Sterility and the Mapping of Male Sterile Gene

Bachelor of Science: Biotechnology, 09/2007 to 06/2011

Nankai University - Tianjin, China

- Thesis: The Fabrication of SSPP Gene Over-expression Vector and Acquisition of Transgenic Plant

Work History

Research Assistant, 09/2017 to Current

University of Illinois at Chicago – Chicago, IL

Publication

- **Shang Gao**, Yang Dai, Jalees Rehman. *A Bayesian Inference Transcription Factor Activity Model for the Analysis of Single Cell Transcriptomes.* **Genome Res**. 2021. 31: 1296-1311.
- Lianghui Zhang, **Shang Gao**, Zachary White, Yang Dai, Asrar B. Malik, Jalees Rehman. *Single-cell transcriptomic profiling of lung endothelial cells identifies dynamic inflammatory and regenerative subpopulations.* **JCI Insight.** 2022: 2379-3708 **(Co-first author)**
- **Shang Gao**, Jalees Rehman, Yang Dai. *Assessing Importance of DNA Sequence and Epigenetics Profiles for Gene expression using a Deep Neural Network.* **Computational and Structural Biotechnology Journal** 2022. 20: 3814-3823.
- Alyne Simões, Lin Chen, Zujian Chen, Yan Zhao, **Shang Gao**, Phillip T. Marucha, Yang Dai, Luisa A. DiPietro, Xiaofeng Zhou*. Differential microRNA profile underlies the divergent healing responses in skin and oral mucosal wounds.* **Sci Rep** 9, 7160 (2019).
- Bethany Baumann, Giovanni Lugli, **Shang Gao**, Morgan Zenner, Larisa Nonn. *High levels of PIWI-interacting RNAs are present in the small RNA landscape of prostate epithelium from vitamin D clinical trial specimens*. **Prostate**. 2019 Jun;79(8):840-855.

Selected Presentations

- **Shang Gao**, Yang Dai, Jalees Rehman. *A Bayesian Inference Transcription Factor Activity Model for the Analysis of Single Cell Transcriptomes*. 2019 GEMS Symposium student talk session 2019, Oct. 4, Chicago, IL
- **Shang Gao**, Lianghui Zhang, Asrar B. Malik, Yang Dai, Jalees Rehman. *A Novel Target-Based Clustering Approach to Analyze Single Cell RNA Seq Data Identifies Distinct Endothelial Cell Subpopulations Following Acute Inflammatory Injury.* Poster presented at Keystone Symposia: Single Cell Biology 2019, Jan. 14-17, Denver, CO.
- **Shang Gao**, Lianghui Zhang, Asrar B. Malik, Yang Dai, Jalees Rehman. *Single Cell RNA Seq Identifies Distinct Lung Endothelial Cell Subpopulations Following Acute*

Inflammatory Injury. American Heart Association Scientific Sessions 2018, Nov. 10-12. Chicago, IL.