SINGLE-CELL GENE EXPRESSION VARIABILITY: FUNCTIONAL ASSESSMENT AND APPLICATIONS IN FUNCTIONAL GENOMICS

A Dissertation

by

DANIEL CAMILO OSORIO HURTADO

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	James J. Cai
Committee Members,	Alan Dabney
	Jerome Menet
	David Riley
Head of Department,	Todd O'Hara

May 2021

Major Subject: Biomedical Sciences

ABSTRACT

Gene expression variability has been associated with specific roles in cell function. However, its functional implications in multicellular organization have not been systematically tested due to technical limitations. Furthermore, their potential application in functional genomics that may increase our understanding of the regulatory mechanisms driving different cell states and the active role of the genes on them have not been evaluated. Thanks to the development of single-cell RNA-seq techniques allowing the measurement of the transcriptome profile in thousands of cells in a single experiment, now it is possible to characterize the synchronized patterns of expression of genes participating in the same biological processes or under the regulation of the same transcription factor. This allows identifying cells under the same cellular state and genes' functional relationships driving those cellular states without the need for genetic manipulations.

Here we introduce three new single-cell RNA-seq datasets (from lymphoblastoid cell lines, *Ahr*, and *Malat1* knockouts). We also introduce cell-type and tissue-specific thresholds for single-cell RNA-seq quality control and novel computational methods to increase our understanding of the biological implications of the single-cell gene expression variability in a high-level order in multicellular organisms and its applications for the identification of differentially regulated genes driving the observed cellular states, and the prediction of the cell-type-specific functional roles of the genes. Our results provide evidence supporting the *'variation is function'* hypothesis suggesting that the aggregate cellular function may depend on the single-cell gene expression variability observed among cells of the same type under the same environment. We show that the reported thresholds for single-cell RNA-seq quality control accurately discriminate between healthy and low-quality cells in different tissues and cell-types. We also demonstrate that our novel computational methods based on gene expression variability and unsupervised machine learning algorithms allow unraveling the regulatory mechanisms underlying cell behaviors and the accurate prediction of the perturbations caused by the deletion of a gene in a gene regulatory network revealing the gene's function in a cell type-specific manner.

DEDICATION

To Andrés, who was my partner in crime and motivation during this adventure.

ACKNOWLEDGMENTS

I want to thank Dr. Cai for his support and mentorship during my training, as well as to the past and current members of the lab for their friendship and thoughtful discussions. I also want to thank Dr. Andrew Hillhouse and Dr. Chris Blazier for the help with the preparation of the single-cell samples and raw data processing that made possible Chapters 2 and 4. Dr. Jianhua Huang and his students Yan Zhong and Guanxun Li for the helpful discussion on data analysis that contributed to the analysis presented in Chapter 2. I also want to thank all the members of Dr. Chapkin's lab from the Texas A&M nutrition department for their reflective questions that motivated Chapter 3, and Dr. Leif Andersson, Dr. Michael Criscitiello, Dr. Ivan Ivanov, Dr. Vijayanagaram Venkatraj, Dr. Simon Mitchell, and Dr. Barbara Gastel for their for insightful comments and writing suggestions that helped to improve Chapter 4. This work would not be possible without the contributions of all the other co-authors: Dr. James J. Cai, Dr. Xue Yu, Dr. Peng Yu, Dr. Erchin Serpedin, Dr. Jianhua Huang, Dr. Laurie A. Davidson, Dr. Robert S. Chapkin, Dr. Andrew Hillhouse, Dr. Jingshu Chen, Dr. Yanan Tian, Yan Zhong, Guanxun Li, and Qian Xu.

Additionally, I would like to thank the academic advisors from the Office of Research and Graduate Studies of the College of Veterinary Medicine Biomedical Sciences: Mrs. Smith, Mrs. Cooper, Mrs. Ojala, Dr. Kessler, and Dr. Gustafson-Seabury, for all their support and guidance since the first day of this process. Finally, but not less important, I want to thank my family and friends for all their love and care during this time.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professor James J. Cai of the Department of Veterinary Integrative Biosciences, the Interdisciplinary Program of Genetics, and the Department of Electrical and Computer Engineering who acted as the chair of the committee. Additionally, Professors Alan Dabney of the Department of Statistics, Jerome Menet of the Department of Biology, and David Riley of the Department of Animal Sciences served as members of the committee.

Chapter 2 contributions are as follows: Daniel Osorio, Dr. Xue Yu, Dr. Peng Yu, Dr. Erchin Serpedin and Dr. James J. Cai conceived and designed the project; Daniel Osorio and Dr. Xue Yu cultured the cells; Daniel Osorio and Dr. James J. Cai performed bioinformatics analysis, Daniel Osorio, Dr. Xue Yu, Dr. Peng Yu, Dr. Erchin Serpedin and Dr. James J. Cai analyzed the data; Daniel Osorio and Dr. James J. Cai wrote the manuscript. All authors reviewed the manuscript.

Chapter 3 contributions are as follows: Daniel Osorio conceived and designed the project, performed data collection and analyzed the data, Daniel Osorio and Dr. James J. Cai wrote the manuscript. All authors reviewed the manuscript.

Chapter 4 contributions are as follows: Dr. James J. Cai conceptualized the project; Daniel Osorio, Yan Zhong and Dr. Jianhua Huang developed the methodology; Daniel Osorio and Dr. James J. Cai implemented the software; Daniel Osorio and Guanxun Li validated the results; Dr. Xue Yu performed the experimental analysis; Dr. James J. Cai drafted the original version of the manuscript; Daniel Osorio, Dr. Erchin Serpedin and Dr. Jianhua Huang reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Chapter 5 contributions are as follows: Dr. James J. Cai and Daniel Osorio designed the workflow and conceptualized the study. Daniel Osorio implemented the software. Daniel Osorio, Yan Zhong, and Guanxun Li performed data analysis. Dr. James J. Cai and Dr. Jianhua Huang.

supervised the data analysis. All authors contributed to the writing of the manuscript.

Chapter 6 contributions are as follows: Dr. James J. Cai conceived and designed the workflow and data analysis and implemented the MATLAB version of the software. Daniel Osorio designed the workflow and implemented the R version of the software. Yan Zhong, Guanxun Li and Qian Xu contributed to the software design and data analysis. Dr. James J. Cai and Dr. Jianhua Huang. supervised the data analysis. Dr. Laurie A. Davidson and Dr. Robert S. Chapkin. performed the *Ahr* KO experiment; Dr. Jingshu Chen and Dr. Yanan Tian performed the *Malat1* KO experiment; Dr. Andrew Hillhouse performed the single-cell RNA-seq experiments. All authors contributed to the writing of the manuscript.

The experimental procedures described in sections 6.2.7, and 6.2.8 were performed by Dr. Chapkin's Lab and Dr. Yanan Tian's Lab respectively. All other work conducted for the dissertation was completed by the student independently.

Funding Sources

This study was supported by Texas A&M University T3 grant for Dr. James J. Cai, Dr. Erchin Serpedin, and Dr. Peng Yu. Dr. James J. Cai was also supported by NIH grant R21AI126219, the Texas A&M University X-Grant (2019), and the DoD GW200026 grant. Daniel Osorio and Yan Zhong were supported by the 2020 Award of Texas A&M Institute of Data Science (TAMIDS) Data Resource Development Program. Dr. Robert S. Chapkin was supported by the Allen Endowed Chair in Nutrition & Chronic Disease Prevention, NIH R01-ES025713, R01-CA202697, and R35-CA197707 grants.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Gene expression (GE) is a cellular phenotype under genetic and environmental regulation [1]. At the molecular level, GE starts with the association of RNA polymerase and a transcription factor (TF), which leads to the binding of the polymerase in the promoter region (in a double-stranded state) of specific genes regulated by the TF. The RNA polymerase then separates the DNA into single strands, creating an open chromatin state, where the template strand can be read in the 3' to 5' direction. Nucleotides are then quickly paired with their complementary bases until they reach the terminator signal. After that, the RNA polymerase leaves the DNA, releases the mRNA, and the two DNA strands come back together and reform the double helix [2]. This event occurs in each cell independently for each gene, in a synchronized manner essential to maintain the cellular homeostasis and respond to specific stimuli.

The recent advancement of technologies to measure the gene expression profile at the single-cell level (single-cell RNA-seq) in thousands of cells in the same experiment has increased our understanding of the biological processes happening in the cells [3, 4]. This new resolution of the expression profiles has been widely adopted to refine the categories of established cell types and systematically and reproductively examine complex tissues' cellular composition [5]. The power of single-cell RNA-seq is also often used to classify new cellular states among the same cell type [6]. Under the same conditions, cells of the same type and state can still exhibit substantial cell-to-cell differences in their gene expression profiles [7]. These changes are known as single-cell gene expression variability.

At the bulk level, GE variability is measured among individuals of a population, and it is known to be responsible for producing phenotypic variability [8]. In medicine, phenotypic variability has been associated with disease susceptibility due to the generation of extreme phenotypes and reduced response to drugs that target genes with highly variable expression profiles among individuals [9, 10]. At the single-cell level, however, their functional roles in cell homeostasis are still under research. Gene expression variability can be classified into intrinsic and extrinsic variability depending on their origin. Biochemical processes like transcription and translation produce 'intrinsic' variability, and fluctuations in the quantities or states of other cellular components and extracellular signals often contribute indirectly to gene expression changes and the generation of 'extrinsic' variability [11].

Intrinsic GE variability has been hypothesized to be crucial for population-level function in a hypothesis called *variation-is-function* [12]. The *variation-is-function* hypothesis states that the intrinsic cell-to-cell variability indicates a diversity of hidden functional capacities, which may facilitate cells' collective behavior. This collective behavior is essential for the function and normal development of cells and tissues. The loss of this collective cellular behavior may result in disease. Thus, investigation of the intrinsic cell-to-cell variability may contribute to understanding pathological processes associated with disease development. Identifying the highly variable genes in different biological scenarios may also help identify non-suitable targets for treating diseases [10]. Testing this hypothesis requires measuring each cell's individual gene expression profile within a highly homogeneous population of cells. Micro-environmental perturbations and stochastic factors at the cellular level, such as local cell density, cell size, shape, and rate of proliferation, cell cycle have to be controlled [13, 14, 15]. The characterization of the relationship of the gene expression variability on cell function requires understanding which genes show greater or less cell-to-cell variability in their expression [16]. Under the *variation-is-function* hypothesis, different tissues or cell types must have different sets of highly variable genes [12]. These highly variable genes should be enriched with functions that reflect the biological processes associated with the respective tissues or the cell types.

GE variability measured across cells also allow us to infer relevant interactions between genes without the need for genetic manipulations [17]. By using reverse engineering, single-cell GE variability enables identifying the regulatory mechanisms associated with its own generation [18, 19]. Those regulatory mechanisms are usually represented in a gene regulatory network (GRN) at different resolution levels from a single individual. A GRN is a graph that describes the relationship between transcription factors, associated proteins, and their target genes [20]. The analysis of

GRNs promotes the interpretation of cell states, cell functions, and regulatory mechanisms that underlie cell behaviors. However, constructing single-cell GRNs (scGRNs) using data from single-cell RNA-seq and then effectively comparing constructed scGRNs presents significant analytical challenges. Single-cell RNA-seq data is highly sparse, usually containing more than 70% zero values [21]. The sparsity level makes mapping GRNs from single-cell data a challenging task that requires a robust construction method to avoid misleading results and inappropriate conclusions [22]. The comparison between two scGRNs is also a difficult task. Comparing each edge of the graph between networks would be ill-powered when the networks represent thousands of genes' regulatory relationships. This comparison's main goal is to detect alterations in gene regulation between cellular states that may not be recovered by just testing the differential expression and relating those alterations to cellular function changes [23, 24].

Since transcription factors regulate eukaryotic cells' gene expression, generating patterns across the expression profiles of all the target genes regulated by the same transcription factors [25]. These patterns have allowed mapping a significant proportion of interactions between biochemical entities in the cell. Moreover, synchronized expression patterns (co-expression) are characteristic of genes involved in the same biological process or metabolic pathway. Those relationships represented in the gene regulatory network are known to allow predicting the perturbations such as the caused by the deletion or malfunction of a gene [26, 27]. Given that at the transcriptome level, the topmost usually perturbed (differentially regulated) genes after a gene knockout (KO) are associated with the biological process or metabolic pathway in which the knocked-out gene is involved [28]. This predictable behavior of the cells allows characterizing the biological processes in which a gene is involved in different cell states. Compared with the traditional way to decode the cellular role of a gene consisting of inactivating one of the organism's genes and then characterize the displayed phenotype compared with the one from a wild-type organism with a similar genetic background, the approach based on the analysis of the network topology requiring only the wild-type RNA-seq data is more suitable to identify the functional role of the genes under different scenarios in a time and cost-effective manner.

Here I present my research using the gene expression variability observed in public and newly generated single-cell RNA-seq experiments. I used this property of the datasets to test the variation-is-function hypothesis, to construct and compare single-cell gene regulatory networks among different cellular states and perturbations to detect alterations in gene regulation and relate those alterations into changes in cellular function, as well as to predict the cell-type-specific functional roles of the genes in different cell states.

This research project was based on the fact that gene expression variability is ubiquitous in all kingdoms' organisms and has been associated with specific roles in cell function [7, 12]. However, its functional implications in cell function and multicellular organization have not been systematically tested due to technical limitations. Furthermore, their potential application in functional genomics that may increase our understanding of the regulatory mechanisms driving different cell states and the active role of the genes on them have not been evaluated. The foundations were that the development of new high-throughput technologies such as single-cell RNA-seq allowing us to measure all genes' expression also provides us with an extraordinary opportunity to characterize the biological processes happening in every single cell of an organism [4]. The expression profiles of the genes across cells also contain information about the synchronized patterns of expression of genes participating in the same biological processes or under the regulation of the same transcription factor allowing the identification of functional relationships between genes without genetic manipulations [17]. Our rationale was to use available public datasets and develop new computational methods that make use of the variability of the gene expression, to increase our understanding of the biological implications of such variability in a high-level order in multicellular organisms, identify differentially regulated genes driving the observed phenotypes through the construction and comparison of gene regulatory networks, and to predict the cell-type-specific functional roles of genes using the information provided by the constructed single-cell gene regulatory networks' topology.

In the next five chapters, I describe the single-cell RNA-seq data generation and characterization of two lymphoblastoid cell lines (LCLs) with different ascendance (Chapter 2). The comparison of the mitochondrial content across all the datasets included in the PanglaoDB [29] database to generate thresholds for quality control purposes (Chapter 3). The evaluation of the *variation-is-function* hypothesis [30] taking advantage of the distributional properties of the single-cell RNA-seq data and the advance of the multivariate methods to select homogeneous populations of cells (Chapter 4). The development of scTenifoldNet (Chapter 5), a workflow based on machine learning methods to construct and compare single-cell gene regulatory networks using single-cell RNA-seq data collected from different conditions, as well as the development of scTenifoldKnk (Chapter 6), another workflow to perform in-silico knockout experiments using single-cell RNA-seq data from wild-type control samples used as input.

2. SINGLE-CELL RNA SEQUENCING OF A EUROPEAN AND AN AFRICAN LYMPHOBLASTOID CELL LINE *

2.1 Introduction

Immortalized cell lines are continuously growing cells derived from biological samples. Lymphoblastoid cell lines (LCLs) are one of the important members among many immortalized cell lines [31]. LCLs are usually established by infecting human peripheral blood lymphocytes in vitro with Epstein-Barr virus (EBV). The viral infection selectively immortalizes resting B cells, giving rise to an actively proliferating B cell population [32]. LCLs exhibit a low somatic mutation rate in continuous culture, making them the preferred choice of storage for individuals' genetic material [33]. As one of the most reliable, inexpensive, and convenient sources of cells, LCLs have been used by several large-scale genomic DNA sequencing efforts such as the International HapMap and the $1,000$ Genomes projects [34, 35], in which a large collection of LCLs were derived from individuals of different genetic backgrounds, to document the extensive genetic variation in human populations.

LCLs are also an in vitro model system for a variety of molecular and functional assays, contributing to studies in immunology, cellular biology, genetics, and other research areas [36, 37, 38, 39, 40, 41, 42]. It is also believed that gene expression in LCLs encompasses a wide range of metabolic pathways specific to individuals where the cells originated [43]. LCLs have been used in population-scale RNA sequencing projects [44, 45, 46], as well as epigenomic projects [47]. For many LCLs used as reference strains, both genomic and transcriptomic information is available, making it possible to detect the correlation between genotype and expression level of genes and infer the potential causative function of genetic variants [48]. Furthermore, compar-

This chapter has been reprinted from: **Osorio, Daniel, Xue Yu, Peng Yu, Erchin Serpedin, and James J. Cai. 'Single-cell RNA sequencing of a European and an African lymphoblastoid cell line'.** ***Scientific data*** **6, no. 1 (2019): 1-8.** Available at: https://www.nature.com/articles/s41597-019-0116-4. Licensed under a Creative Commons Attribution 4.0 International License https://creativecommons.org/licenses/by/4.0/, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

isons of gene expression profiles of LCLs between populations such as between Centre d'Etude du Polymorphisme Humain – Utah (CEPH/CEU) and Yoruba in Ibadan, Nigeria (YRI), have revealed the genetic basis underlying the differences in transcriptional activity between the two populations [46, 49].

With the advent of single-cell RNA sequencing (single-cell RNA-seq) technology [50, 4], our approach for understanding the origin, global distribution, and functional consequences of gene expression variation is ready to be extended. For example, data generated from single-cell RNA-seq provide an unprecedented resolution of the gene expression profiles at single cell level, which allows the identification of previously unknown subpopulations of cells and functional heterogeneity in a cell population [51, 52, 53].

In this study, we used single-cell RNA-seq to assess the gene expression across thousands of cells from two LCLs: GM12878 and GM18502. Cells were prepared using a Chromium Controller (10× Genomics, Pleasanton, CA) as described previously [4] and sequenced using an Illumina Novaseq. 6000 sequencer. We present this dataset on the single-cell gene expression profile for more than $7,000$ cells from GM12878 and more than $5,000$ from GM18502. GM12878 is a popular sample that has been widely used in genomic studies. For example, it is one of three 'Tier 1' cell lines of the Encyclopedia of DNA Elements (ENCODE) project [47, 54]. GM18502, derived from the donor of African ancestry, serves as a representative sample from the divergent population. The two cell lines are part of the International HapMap project, and genotypic information is available for both of them [34]. We also processed and sequenced an additional sample of $1:1$ mixture of GM12878 and GM18502 using the same single-cell RNA-seq procedure. Our dataset presented here provides a suitable reference for those researchers interested in performing between-populations comparisons in gene expression at the single-cell level, as well as for those developing new statistical methods and algorithms for single-cell RNA-seq data analysis.

2.2 Materials and Methods

2.2.1 Cell culture

GM12878 and GM18502 cell lines were purchased from the Coriell Institute for Medical Research. Cells were cultured in the Roswell Park Memorial Institute (RPMI) Medium 1640 supplemented with 2mM L-glutamine and 20% of non-inactivated fetal bovine serum in T25 tissue culture flasks. Flasks with 20mL medium were incubated on the upright position at 37°C under 5% of carbon dioxide. Cell cultures were split every three days for maintenance. Note that authentication test and mycoplasm contamination screening on these freshly purchased cell lines were not undertaken in this study.

2.2.2 Growth curve

Four culture flasks for each cell line were started with approximately 200, 000 viable cells/mL to measure the growth rate of each cell line. Cells were prepared and cultured as described above. Viable cell number was estimated on a daily basis for four days. Briefly, 100uL suspended cells from each flask were taken every day, to visualize the viable cells, the samples were stained using 10uL of Trypan Blue (0.4%), and live cells were counted manually using a Neubauer counting chamber.

2.2.3 Single cell preparation

Single-cell sample preparation was conducted according to Sample Preparation Demonstrated Protocol provided by 10× Genomics as follows: 1mL of cell suspensions from each cell line (day 4, stable phase) was pelleted in Eppendorf tubes by centrifugation (400g, 5min). The supernatant was discarded, and the cells pellet was then resuspended in 1x PBS with 0.04% BSA, followed by two washing procedures by centrifugation (150g, 3min). After the second wash, cells were resuspended in ≈ 500uL 1x PBS with 0.04% BSA followed by gently pipetting mix 10-15 times. Cells were counted using an Invitrogen Countess automated cell counter (Thermo Fisher Scientific, Carlsbad, CA) and the viability of cells was assessed by Trypan Blue staining (0.4%).

2.2.4 Generation of single cell GEMs (Gel bead in EMulsion) and sequencing libraries

Libraries were prepared using the 10× Genomics Chromium Controller in conjunction with the single-cell 3' v2 kit. Briefly, the cell suspensions were diluted in nuclease-free water according to manufacturer instructions to achieve a targeted cell count of $5,000$ for each cell line. The cDNA synthesis, barcoding, and library preparation were then carried out according to the manufacturer's instructions. Libraries were sequenced in the North Texas Genome Centre facilities on a Novaseq. 6000 sequencer (Illumina, San Diego).

2.2.5 Mapping of reads to transcripts and cells

Sample demultiplexing, barcode processing, and unique molecular identifiers (UMI) counting were performed by using the 10× Genomics pipeline CellRanger v.2.1.0 with default parameters. Specifically, for each library, raw reads were demultiplexed using the pipeline command `'cellranger mkfastq'` in conjunction with `'bcl2fastq'` (v2.17.1.14, Illumina) to produce two fastq files: the read 1 file contains 26-bp reads, each consists of a cell barcode and a unique molecule identifier (UMI), and the read 2 file contains 96-bp reads including cDNA sequences. Reads then were aligned to the human reference genome (GRCh38), filtered, and counted using `'cellranger count'` to generate the gene-barcode matrix. Summary metrics of barcoding and sequencing from raw data are given in Table 2.1.

2.2.6 Quality control

Expression matrices were processed using Seurat (v2.3.4) R package [55]. Briefly, for each library, the expression matrix was loaded using the `'Read10X'` function, and the default log-normalization was performed using the `'NormalizeData'` function, followed by a cantering and scaling of the normalized values by using the `'ScaleData'` function. Quality control (QC) measures, including UMI count, the number of genes detected per cell, and the percentage of mitochondrial transcripts were calculated. Cells with a proportion of mitochondrial reads lower than 10% and a library size smaller than $2.5\times$ standard deviation (SD) from the average library size were considered good quality cells.

	GM12878	GM18502	Mixture
Estimated Number of Cells	7,247	5,530	5,828
Mean Reads per Cell	65,466	91,493	83,326
Median Genes per Cell	2,954	3,960	3,621
Number of Reads	474,436,605	505,958,821	485,628,282
Valid Barcodes	97.20%	97.30%	97.20%
Sequencing Saturation	50.30%	53.50%	53.30%
Q30 Bases in Barcode	94.90%	94.80%	94.80%
Q30 Bases in RNA Read	90.20%	89.60%	89.90%
Q30 Bases in Sample Index	91.50%	93.40%	92.20%
Q30 Bases in UMI	94.80%	93.40%	94.70%
Reads Mapped to Genome	93.90%	93.70%	93.70%
Reads Mapped Confidently to Genome	92.00%	92.00%	92.00%
Reads Mapped Confidently to Intergenic Regions	2.60%	2.70%	2.70%
Reads Mapped Confidently to Intronic Regions	12.90%	13.10%	12.80%
Reads Mapped Confidently to Exonic Regions	76.50%	76.20%	76.50%
Reads Mapped Confidently to Transcriptome	72.60%	71.90%	72.50%
Reads Mapped Antisense to Gene	0.90%	0.90%	0.90%
Fraction Reads in Cells	90.70%	91.70%	89.80%
Total Genes Detected	21,329	20,701	21,151
Median UMI Counts per Cell	18,214	25,973	22,608

Table 2.1: Summary metrics for 10× Genomics single-cell RNA-seq barcoding and sequencing of three LCL samples (GM12878, GM18502, and the $1:1$ mixture).

2.2.7 Cell cycle phase and population assignment

Cell cycle phase assignment was made using the `'CellCycleScoring'` function in the Seurat R package [55], which uses the phase-specific marker genes, given by the `'cc.genes'` dataset [56]. Cell population assignment, i.e., assigning cells in the mixture sample back to the cell line (GM12878 or GM18502) they belong to, was made using the Brunet algorithm [57] for non-negative matrix factorization, in the NMF (v.0.21) R package [58]. A set of marker genes ($n = 252$) with absolute log-fold change > 2.5 identified by comparing the pure cell lines was used as inputs and the resulting probabilities after $2,000$ iterations were used to assign each cell in the mixture to either GM12878 or GM18502.

2.2.8 Dimensionality reduction

Expression matrices from GM12878, GM18502, and the mixture sample were merged and log-normalized using the function `'MergeSeurat'`. The resultant matrix was then centered and scaled. Highly variable genes were identified using function `'FindVariableGenes'` in the Seurat R package [55]. Identified highly variable genes were used as input to produce the t-Distributed Stochastic Neighbour Embedding (t-SNE) projection using the `'RunTSNE'` function with standard settings (`perplexity = 30`, `theta = 0.5`, `maximum_iteration = 1000`, `learning_rate = 250`, and `momentum_reduction = 0.5`, by using the first 5 components from the principal component analysis). The Uniform Manifold Approximation and Projection (UMAP) was produced with the same set of highly variable genes as input using the function `'RunUMAP'` with standard settings (`min_dist = 0.3`, `metric = correlation`, `n_neighbors = 30`).

2.2.9 single-cell RNA-seq versus bulk RNA-seq

For both GM12878 and GM18502, transcriptome has been previously sequenced using bulk RNA-seq. The availability of these existing data allowed us to examine the correlation between gene expression levels measured using single-cell RNA-seq and bulk RNA-seq in the same LCLs. Thus, we downloaded the raw fastq files of bulk RNA-seq experiments from the Gene Expression Omnibus (GEO) database using accessions GSM484896 [59] (for GM12878) and GSM2392689 [60] (for GM18502) and quantified gene expression for both samples using Salmon [61] (v.0.12.0) against the human transcriptome (GRCh38). In addition, we also compared gene expression measured using single-cell RNA-seq in GM12878 and GM18502 with the average gene expression measured in multiple samples from CEU and YRI populations. To do so, we downloaded the bulk RNA-seq data of 91 CEU and 89 YRI LCLs from the website of the Geuvadis RNA-seq project of 1,000 Genomes. The expression of each gene was measured as the mean of transcripts per million (TPM) values across all individuals of CEU or YRI population. To visualize the relationship of the single-cell gene-expression profiles of the two cell lines with their respective population, a principal component analysis (PCA) was performed. The input data for PCA was batch-effect corrected using the `'removeBatchEffect'` function in the limma (v.3.4.0) R package [62] and quantile normalized using the `'normalize.quantiles'` function in the preprocessCore (v.1.46.0) R package.

2.3 Results

Here we present the single-cell RNA-seq gene expression profile for $7,045$ and $5,189$ cells for GM12878 and GM18502, respectively. For GM12878, the median UMI counts per cell is $18,214$ and the median number of genes detected (at least 1 UMI) per cell is $3,167$; for GM18502, $25,973$ and $3,891$. Figure 2.1 is a heatmap of log-transformed expression data of top 200 highly expressed genes in the two LCLs. Cells are grouped by their cell cycle phases (G1, S, and G2/M) and sorted within each group by their library size. Among the top expressed genes, there are several immunoglobulin genes such as *IGLC2*, *IGHA1*, *IGKC*, *IGLC3*, and *IGHM*. These genes are not only expressed highly on average but also expressed highly variably across cells—i.e., highly expressed in one set of cells but no expression in another set of cells. We consider that this highly variable expression pattern can be attributed to immunoglobulin gene rearrangement. During the formation of the naïve-B cells, gene rearrangement process occurs to reshuffle different subunits of the variable (V), diversity (D) and joining (J) segments of immunoglobulin genes, resulting in the generation of a wide range of organism-specific antigen receptors that allow the immune system to recognize foreign molecules and initiate differential immune responses [63, 64]. LCLs are produced through the rapid proliferation of few EBV-driven B cells from the blood cell population [65]. Thus, our single-cell RNA-seq data of GM12878 and GM18502 offer a 'snapshot' of highly diverse immunoglobulin rearrangement profiles in a much larger population of polyclonal B cells found in the two donors.

We also performed single-cell RNA-seq with a $1:1$ mixture sample of the two LCLs and obtained data for additional $5,820$ cells with a median UMI counts per cell of $22,608$ and a median number of genes detected per cell of $3,625$. This mixture sample can be considered as a technical replicate for both GM12878 and GM18502. The use of the mixture sample facilitates direct com-

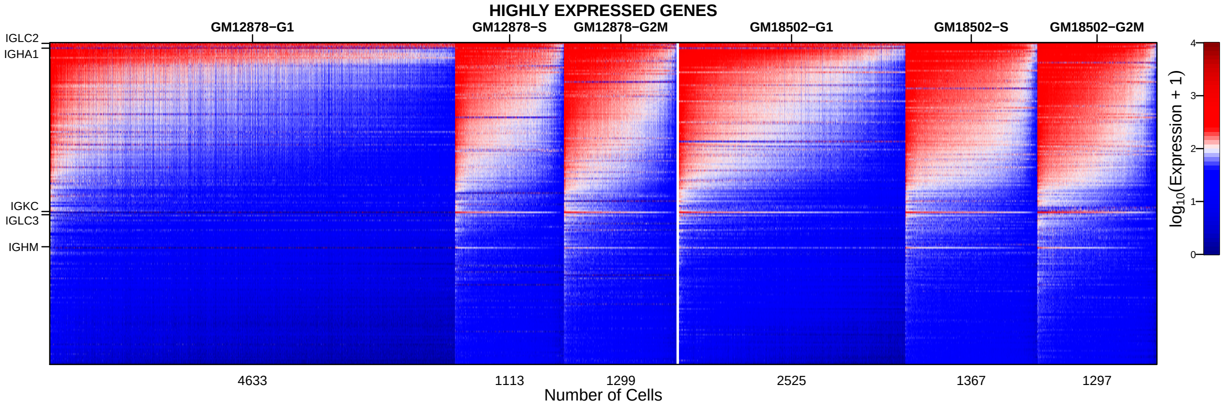


Figure 2.1: Heatmap of single-cell gene expression levels of the top 200 genes highly expressed in GM12878 and GM18502. Values are log-transformed UMI counts. For coloring purposes, values are truncated at a range between 0 and 4. Genes are arranged by the expression level. Cells are grouped according to cell cycle phases and sorted by their library size within each group. Immunoglobulin genes are labeled.

parison of gene expression between GM12878 and GM18502 because cells from two cell lines in the mixture were processed simultaneously in the same reaction, maximally eliminating the batch effect. We found that cells in the mixture were able to be assigned back to their original cell lines almost unambiguously using a non-negative matrix factorization algorithm (see Section 2.2.7). Furthermore, the average gene expression measured in cells in the mixture, after discriminating cells in the mixture and assigning them to their respective one of original cell lines, was virtually indistinguishable from that measured in the original 'pure' cells (Fig. 2.2).

The percentage of mitochondrial transcripts, an indicator of apoptotic cells, was computed for all cells sequenced in all the three samples. We found that no more than 0.4% of cells, that is, 26 cells from GM12878, 6 from GM18502, and 23 cells from the mixture sample, surpass the commonly used threshold of 10% mitochondrial transcripts [66]. This suggests that the majority of cells processed and sequenced were viable. Furthermore, as the 10× Genomics Chromium technology relies on droplets to partitioning cells and barcoding, it is normal some of them contain multiple cells in the cell droplet, making the estimation of the frequency of multiplets a critical aspect of quality control [67]. There are several ways to identify multiplets [68, 69, 70]. Here

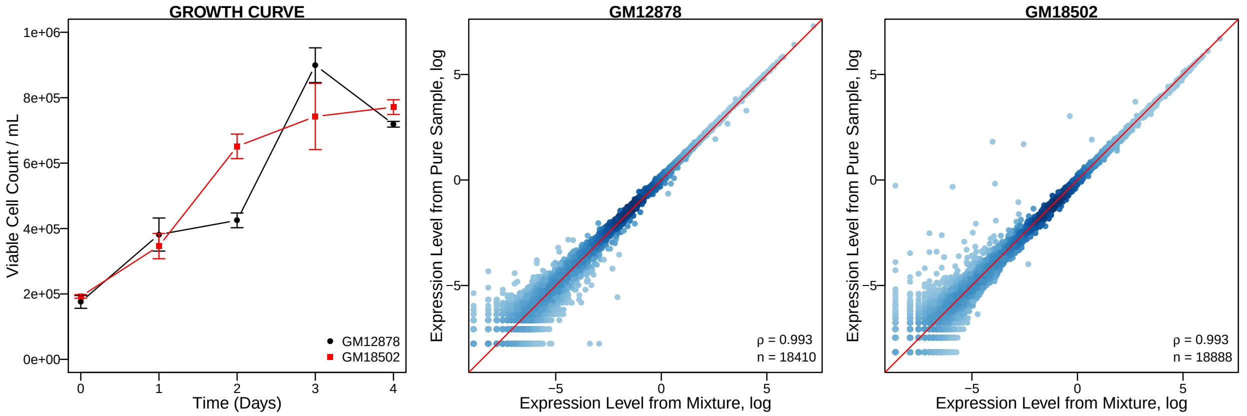


Figure 2.2: Cell growth curves and the gene expression correlations between samples. (a) Growth curve of the GM12878 and GM18502 cultured in the same RPMI 1640 medium. (b) Spearman's correlation between the gene expressions profiles UMI average of the cells assigned to the CEU population from the mixture and those from the pure GM12878 cell line. Values were log-transformed, and each dot represents a gene. (c) Spearman's correlation between the gene expression (average UMI) of cells assigned to the YRI population from the mixture and those from the pure GM18502 cell line. Values are log-transformed, and each dot represents a gene.

we adopted the threshold of 2.5× SD from the average library size for each cell. Based on this threshold, only 171 cells were considered to be multiplets for GM12878, 66 for GM18502, and 87 for the mixture (Fig. 2.3). These results support the quality of the dataset.

In either t-SNE or UMAP projection, no separation was observed between cells from the two pure cell lines, GM12878 and GM18502, and cells from the corresponding replicates of the two pure cell lines in the mixture (Fig. 2.4). This result suggests that cells in the mixture have the global expression profiles indistinguishable from those of cells of their original samples. Population signal of each sample allows a sample to be separated from others in the first two t-SNE or UMAP dimensional spaces. Furthermore, for each cell line, cells of different cell cycle phases are not entirely separated – a continuous path between the different clusters of cells exist. This allows researchers interested in cell cycle development to perform pseudo-time analysis [71]. Also, cells in the same cell cycle phase tend to be spread out and form a spectrum of cells in intermediate stages, indicating that cell proliferation is a continuous process and researchers interested in this process can use this dataset to refine reference cell sub-populations by their characterized

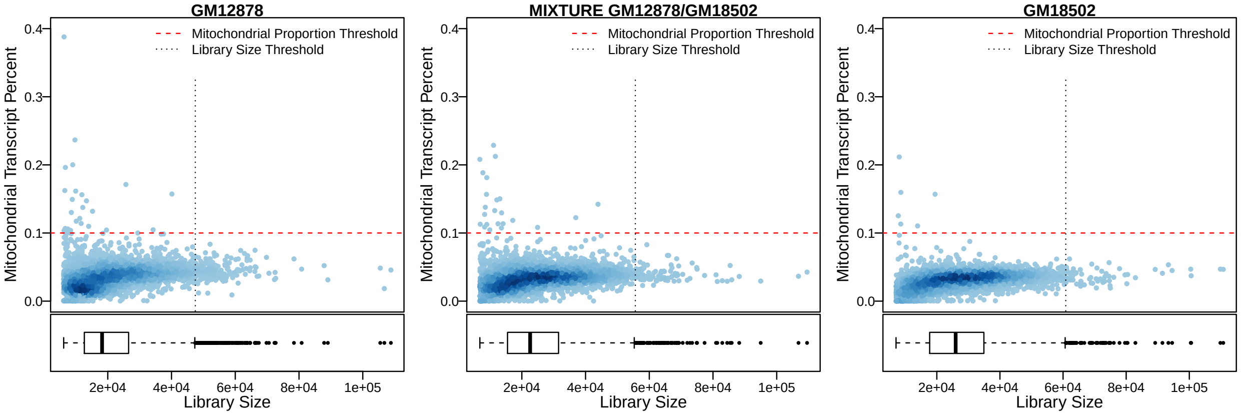


Figure 2.3: Distribution of the single-cell gene expression profiles under the defined quality control thresholds. There are $6,848$ cells for the GM12878, $5,117$ for the GM18502 and $5,710$ for the mixture sample within the range of thresholds. These cells are considered to be of high quality.

expression profiles.

For both GM12878 and GM18502, we conducted correlation analyses to validate our single-cell RNA-seq expression data using bulk RNA-seq expression information as a reference. We first compared gene expression measured using single-cell RNA-seq and bulk RNA-seq in the same LCL, GM12878 or GM18502. We also compared gene expression measured using single-cell RNA-seq in GM12878 (and GM18502) with the average gene expression in corresponding population CEU (and YRI). We found that in all cases the correlations are highly significant and strong with Spearman correlation coefficients (SCCs) of 0.78, 0.58, 0.76, and 0.77, respectively (Fig. 2.5). Thus, when single-cell RNA-seq data are pooled across cells, genes' expression levels are largely recapitulated as they were measured using bulk RNA-seq. These results further support the quality of our single-cell RNA-seq dataset. We note that the SCC (0.58) between GM18502 single-cell RNA-seq and GM18502 bulk RNA-seq is lower than that (0.78) between GM12878 single-cell RNA-seq and GM12878 bulk RNA-seq. This may be due to differences in cell population state at the time when GM18502 cells were harvested for single-cell RNA-seq and bulk RNA-seq.

As long-lasting supplies of cells containing genotypic and phenotypic information matching that of B-cell origins, LCLs have contributed significantly to biomedical research. We present a high-quality dataset of single-cell RNA-seq from homogenous cell populations of two LCLs, in-

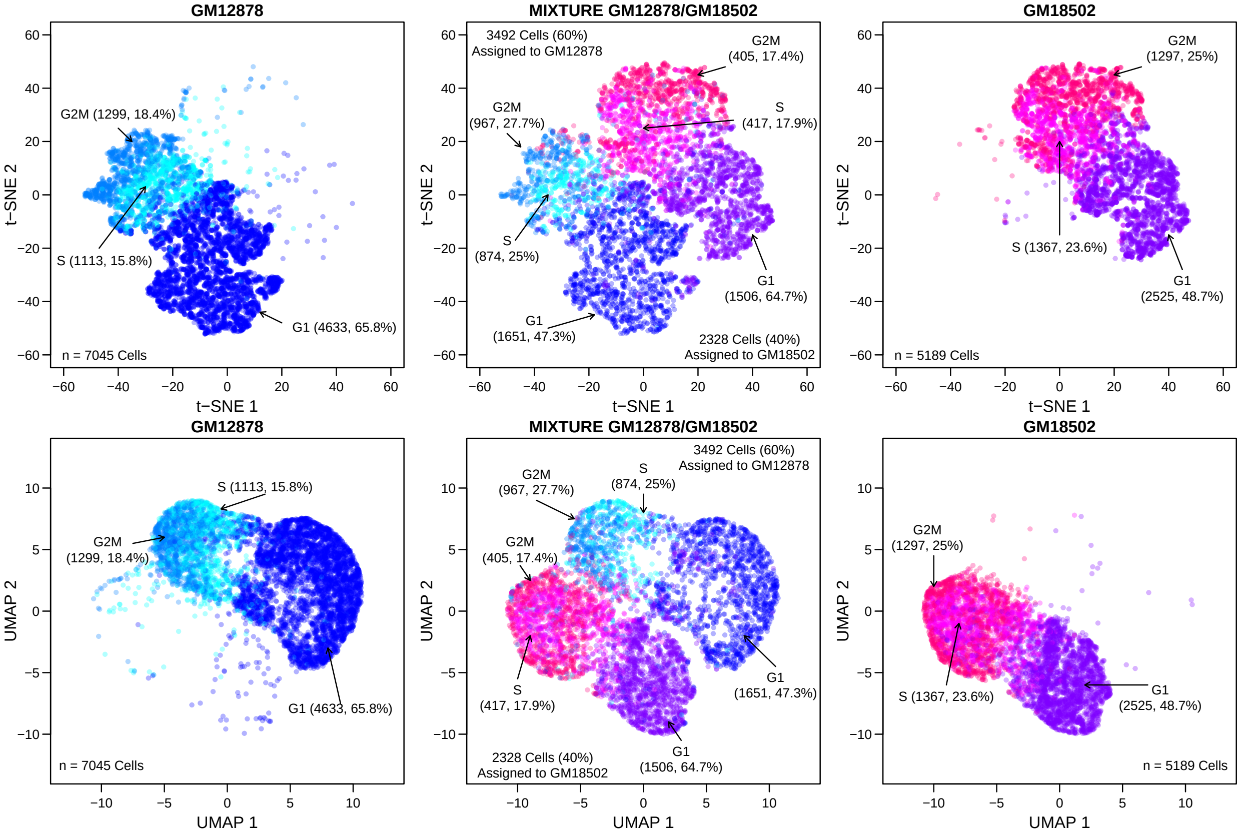


Figure 2.4: Plots of t-SNE and UMAP projections generated from the pooled single-cell RNA-seq data of GM12878, GM18502, and the mixture samples. Separate panels are used to show cells labeled and colored differentially according to their cell line name and cell cycle state.

cluding GM12878—one of the most popular reference cell lines. Our dataset provides information that can be used to quantify cell-to-cell variability in gene expression and study cellular states and associated gene expression changes. It also informs the analysis and comparison of gene expression at the single-cell level between European and African LCLs. The data from the mixture sample are a suitable resource for estimating the technical variability of single-cell RNA-seq and can also be used to calibrate statistical methods for data normalization and batch effect correction.

2.3.1 Data Records

The sequencing data from this study have been submitted as the BioProject reference (PRJNA508890), with descriptions of the Biosamples (SUB4895416, SUB4895422, SUB4895423). Raw data of three samples have been deposited at the National Center for Biotechnology Infor-

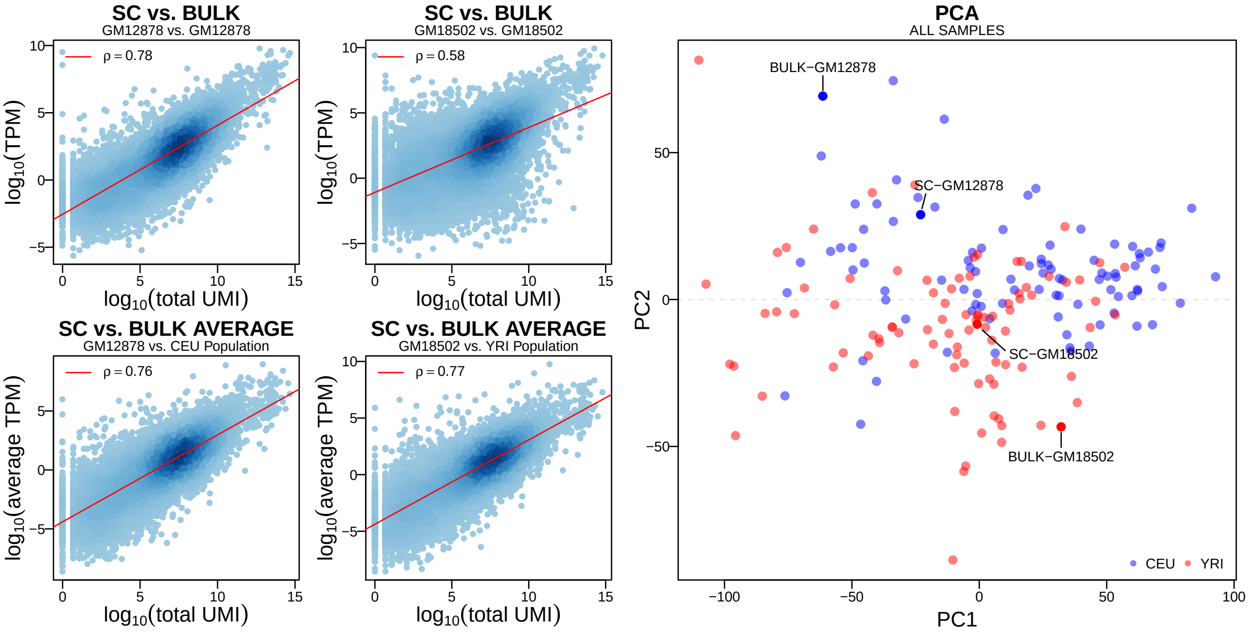


Figure 2.5: Gene expression correlations between single-cell sample, bulk-cell sample, and population average of bulk-cell samples. (a) Spearman's correlation between the gene expressions profiles at the single-cell level and the bulk expression level (TPM) for GM12878 and GM18502. (b) Spearman's correlation between the gene expressions profiles at the single-cell level for the GM12878 and GM18502 compared to the average bulk level expression (average TPM) for the available samples of CEU and YRI. Values are log-transformed, and each dot represents a gene. (c) PCA plot shows the similarity between the same samples' gene expression profiles obtained using bulk RNA-seq and single-cell RNA-seq.

mation (NCBI) Sequence Reads Archive (SRA) with accession ID: SRP172838. For each sample, data include unprocessed single-cell RNA-seq reads in two raw fastq files (*R1.fastq.gz for cell barcodes and UMIs, and *R2.fastq.gz for RNA reads), as well as an expression matrix file in matrix market exchange format (*.mtx) with columns corresponding to cells and row to genes. UMI matrices of this study have been deposited with the Gene Expression Omnibus at GEO: GSE126321. The identifiers for the columns and rows are included in separated files (barcodes.tsv and genes.tsv). These processed files correspond to the output produced by the cell ranger pipeline. In addition, a supplementary table with the barcodes, population, UMI count, gene count, and mitochondrial transcript levels is included.

2.3.2 Code Availability

All the required code to replicate the feature characterization of GM12878 or GM18502 and the mixture, as well as all figures included in this document, are available in a public repository on GitHub at https://github.com/cailab-tamu/sciData-LCL.

3. SYSTEMATIC DETERMINATION OF THE MITOCHONDRIAL PROPORTION IN HUMAN AND MICE TISSUES FOR SINGLE-CELL RNA-SEQUENCING DATA QUALITY CONTROL *

3.1 Introduction

Single-cell RNA-seq (single-cell RNA-seq) experiments have improved the resolution of our knowledge about cellular composition and cellular behavior in complex tissues [72]. A critical step during single-cell RNA-seq data processing is to perform quality control (QC) over the cells sequenced transcriptomes [73]. The QC process usually involves applying user-defined thresholds for different metrics computed for each individual cell to filter out doublets and ‘low-quality’ cells [74]. Commonly used QC metrics include the total transcript counts (also known as the library size), the number of expressed genes and the mitochondrial proportion (mtDNA%, i.e. the ratio of reads mapped to mitochondrial DNA-encoded genes to the total number of reads mapped). Defining the proper thresholds of QC metrics is a complex task that requires a vast knowledge of the cellular diversity in the tissue under study. Thresholds may be uniquely set for each sample, as they are dependent on the cells or tissue being processed [75].

This study focuses on the systematic determination of a threshold for mtDNA% – the fraction of mitochondrial counts per cell – in single-cell RNA-seq QC. Mitochondrial content is known to interact with the nuclear genome, drive alternative splicing and regulate nuclear gene expression and is also associated with cancer, degenerative diseases and aging [76, 77]. High numbers of mitochondrial transcripts are indicators of cell stress, and therefore mtDNA% is a measurement associated with apoptotic, stressed and low-quality cells [78, 21, 79]. However, mtDNA% threshold depends highly on the tissue type and the questions being investigated [80]. The mtDNA% threshold is of economic and biological importance. A wrongly defined, very stringent mtDNA%

This chapter has been reprinted from: **Osorio Daniel, Cai James J., ‘Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control’, *Bioinformatics*, 2020, btaa751.** Available at: https://doi.org/10.1093/bioinformatics/btaa751. Licensed by Oxford University Press with license # 5000230599558.

threshold may cause bias in the recovered cellular composition of the tissue under study. This bias may force the researchers to increase the sample size to capture enough cells (which may not have the normal biological behavior of the cell type) under the threshold, and thus increase the cost of the experiment. Inversely, a relaxed threshold of mtDNA% may allow apoptotic, low-quality cells to remain in the analysis, resulting in the identification of wrong biological patterns.

To reduce the bias caused by the use of arbitrary mtDNA% thresholds, Ma *et al.* [81] proposed an unsupervised method to optimize the threshold for each given input data. This computationally expensive data-driven procedure, which defines the threshold as a function of the distribution of the data, due to the lack of reference values, is not able to identify bias induced during the library preparation. Without such standard references, the values of mtDNA% thresholds fluctuate with different input datasets. For example, a largely failed experiment may generate a dataset, in which most cells have an inflated mtDNA%. Accordingly, the optimized threshold based on these inflated values may be unreasonably high. Therefore, having a uniform and standardized threshold for single-cell RNA-seq data analysis is essential. It improves the reproducibility of experiments and simplifies the automatization of bioinformatic pipelines [82].

Through analysis of bulk RNA-seq data produced by the Illumina Body Tissue Atlas, Mercer *et al.* [83] reported the mtDNA% for 16 human tissues. They found that the mtDNA% ranges from 5% or less in tissues with low energy requirements up to $\approx 30\%$ in the heart due to the high energy demand of cardiomyocytes. Based on that study, early publications of single-cell RNA-seq datasets used the 5% threshold reported for tissues with low energy demands (e.g. adrenal, ovary, thyroid, prostate, testes, lung, lymph and white blood cells) as default for data QC [84]. Furthermore, the 5% threshold has been adapted as the default parameter by Seurat – one of the most popular software packages for single-cell RNA-seq data analysis [55]. These have made a 5% practical standard for single-cell RNA-seq data analyses. Nevertheless, due to the lack of reference values for mtDNA% in different species, technologies, tissues and cell types, the optimal value for a standardized threshold is still an open question in the field.

PanglaoDB is a single-cell RNA-seq database providing uniformly processed, annotated count matrices for thousands of cells from hundreds of single-cell RNA-seq experiments. The data source of PanglaoDB is the sequence read archive (SRA) database of the National Center for Biotechnology Information (NCBI). With the datasets from PanglaoDB, it is possible to systematically evaluate the optimal threshold of mtDNA% for different experimental settings that may vary across platforms, technologies, species, tissues or cell types [29, 85]. Here, we present a systematic analysis of the mtDNA% in more than 5 million cells reported in over one thousand datasets in PanglaoDB [29]. We compared the mtDNA% reported for different technologies, species, tissues and cell types. By analyzing the data provided by hundreds of experiments together, we reach the consensus reference values for more than 40 human tissues and more than 120 mouse tissues. Furthermore, we evaluated the validity of using the 5% threshold in different humans and mice tissues and showed that omitting the mtDNA% as a QC filter led to erroneous biological interpretations of the data.

3.2 Materials and Methods

Datasets in the PanglaoDB database, available at the time of analysis (in January 2020), were downloaded and processed using R 3.6.2 [86] through an 'in-house' script using the XML [87] and xml2 [88] packages. The library size (total number of counts), the total number of detected genes and the total number of counts that match with the mitochondrial genes (mitochondrial counts) were estimated for all cells in each of downloaded datasets. The SRA/SRS identifiers, species, protocol, tissue, cell type and barcode of each experiment were obtained and associated with each cell.

Only the cells with more than $1,000$ counts and with the total number of counts greater than two times the average library size in the same sample were retained for analysis. In addition, a polynomic regression of degree 2 (to account for saturation) was applied to establish the 95% confidence intervals of the predicted total number of genes as a function of the library size per cell. Cells with an observed total number of genes below or above expectation limits were removed from the analysis. The same procedure was applied a second time to establish the 95% confidence intervals of the predicted mitochondrial counts as a function of library size. An ordinary least squares (OLS) regression model was used to fit the data and cells with exceptionally high or low mitochondrial counts were removed from the analysis.

Subsequently, the mtDNA% value was computed for each cell as the ratio between the mitochondrial counts and the library size of the cell. The mtDNA% values were then compared between cells from different settings: species, technologies, tissues and cell types. To compare the mtDNA% between humans and mice cells, we used the Welch two-sample t-test and used the Wilcoxon sum-rank test to cross-validate the results. To evaluate the reliability of the 5% threshold, a comparison to evaluate whether the mean was < 0.05 threshold value was performed using the t-test for each tissue and cell type independently using the data generated by the 10× Genomics Chromium system, after filtering out groups with less than $1,000$ cells.

Example datasets (SRS3703557, SRS3545826 and SRS2397417) were downloaded from the PanglaoDB along with cell clustering results and cell type information. Count matrices were processed using the 'Seurat' R package to generate low-dimensional representations. Differential expression analysis was performed using 'MAST' [89] to compare the transcriptome profile of clusters exhibiting a median mtDNA% higher than 5% against that of other clusters with the same cell type but with a median mtDNA% lower than 5%. Using the sorted list of fold-changes reported by MAST, we performed Gene Set Enrichment Analysis (GSEA) to test the enrichment of the 'Apoptosis' pathway from the KEGG database [90] in the clusters with increased mtDNA%. To run the GSEA analysis efficiently, we used the multilevel function included in the 'fgsea' R package [91].

3.3 Results

We downloaded a total of $5,530,106$ cells reported in $1,349$ datasets from the PanglaoDB database. From those, we removed $278,607$ cells with a total number of counts smaller than 1000 or above two times the average library size in the sample where it was sequenced. Also, $80,225$ cells with no mitochondrial counts were removed. The remaining $5,171,274$ cells were used to establish the 95% confidence intervals of the predicted total number of genes as a function of the library size per cell. We found that the relationship between the number of genes and the library size is monotonically positive ($\rho = 0.89$; P $< 2.2 \times 10^{-16}$), which is consistent with that previously reported [85]. We also found that the expected total number of genes reaches saturation at a point close to the 1×10^5 library size counts. In this step, we removed $157,960$ cells because they have a total number of quantified genes above ($n = 5,509$) or below ($n = 152,451$) the 95% confidence interval limit defined from the prediction.

Next, we accounted for outliers in the mitochondrial counts in relative to the library size. This procedure has been shown to be critical to differentiate apoptotic cells of preapoptotic and healthy cells in a supervised experiment [92]. To do so, we used the OLS regression and computed the confidence interval of prediction between the mitochondrial counts and the library size with data from all $5,013,314$ cells. We found that the relationship is noisy but positive and linear ($r = 0.65$, P $< 2.2 \times 10 - 16$. Following this procedure, we identified $333,712$ cells with mitochondrial counts above ($n = 178,671$) or below ($n = 155,041$) computed confidence interval limits, which were also removed. After this step, $4,679,602$ cells were retained for the study.

With the cleaned dataset, we estimated that the mtDNA% per cell is distributed between the minimum of 0.17% and the maximum of 14.64%, considerably lower than the upper limit previously reported (up to 30% in heart) using the bulk RNA-seq generated by the Illumina Body Tissue Atlas [83]. Next, we performed a comparison to evaluate whether there is a difference in the average mtDNA% cross different species. The PanglaoDB database contains human and mouse datasets; therefore, our comparison was between human and mouse. We performed the Welch two-sample t-test and use the Wilcoxon-sum rank test to validate the results. Both tests converged to the same conclusion, that is, the average mtDNA% in human cells is significative higher than that in mice cells (P$<$ 2.2×10^{-16}, in both cases) as is displayed in Figure 3.1 A.

Then, we compared the mitochondrial content between human and mouse data, stratified by the type of single-cell RNA-seq technologies, by which the data are obtained. These technologies include drop-seq, C1 Fluidigm and 10× Genomics. Our results confirm our previous finding. In all cases wherever data allowed, no matter which technology is used, the same pattern was recovered. That is, human cells have significantly larger mtDNA% than mice cells (Fig. 3.1 B.).

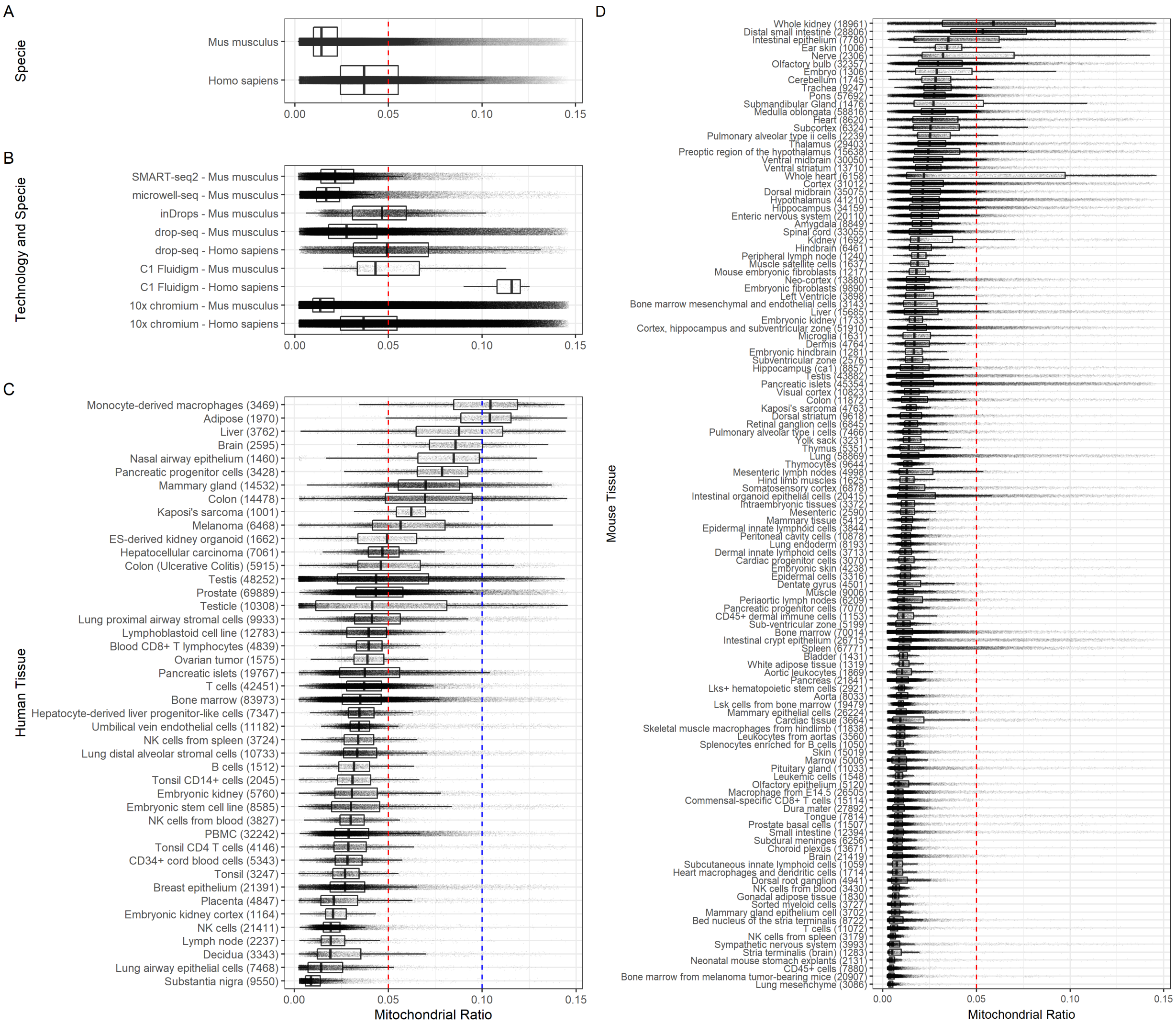


Figure 3.1: Boxplots showing the differences in mtDNA% across species, technologies and tissues. Each dot represents a cell; the red line is the early established 5% threshold, and the blue line is the 10% threshold for human cells proposed here. In parenthesis (C and D), the number of cells in the stated tissue. (A) The difference in mtDNA% between human and mice cells. (B) The differences in mtDNA% between human and mice cells by the technology used to generate the data. (C) Boxplots of mtDNA% across 44 human tissues. (D) Boxplots of mtDNA% across 121 mouse tissues

Most importantly, for all cases where mitochondrial content in humans was evaluated, the 75th percentile was located above the threshold, suggesting that the early defined 5% is not appropriate for human cells. Note that 91.3% ($n = 4,271,613$) of cells analyzed here were processed using the 10× Genomics chromium system. Next, we decided to perform the comparison of the mitochondrial content between tissues and cell types using only the data generated using the 10× Genomics technology.

For humans, we identified 44 tissues, for which more than $1,000$ cells are available in the database. From those 44 tissues, 13 (29.5%) showed an average mtDNA$\%$ significantly higher than 5%. The 13 human tissues are nasal airway epithelium, monocyte-derived macrophages, testicle, colon (ulcerative colitis), liver, colon, melanoma, mammary gland, ES-derived kidney organoid, pancreatic progenitor cells, adipose, Kaposi's sarcoma and brain. However, as displayed in Figure 3.1 C., 18 of the 44 human tissues (40%) have a portion of the interquartile range over the 5% threshold. Only two of them, monocyte-derived macrophages and adipose, have an average mtDNA$\%$ higher than 10%. This result supports our observation that the early defined 5% is not appropriate for human tissues. We conclude that the new standardized threshold for human tissues should be 10% instead. At the cell-type level, we found similar patterns. From 37 different cell types with more than $1,000$ cells derived from human samples, 13 of them (35.1%) have an average mtDNA$\%$ greater than 5%, but none of them have an average mtDNA$\%$ greater than 10%. The 13 cell types are hepatocytes, epithelial cells, neutrophils, cholangiocytes, smooth muscle cells, keratinocytes, Langerhans cells, spermatocytes, ductal cells, beta cells, luminal epithelial cells, macrophages and embryonic stem cells.

Furthermore, only 4 of them (epithelial cells, Langerhans cells, spermatocytes and macrophages) have a portion of the interquartile range above the 10% threshold. For mice, when the mtDNA$\%$ was compared across the cell types with at least $1,000$ cells reported in the database, 7 of 74 cell types showed an average mtDNA$\%$ greater than 5%. The 7 cell types are proximal tubule cells, distal tubule cells, hepatocyte, cardiomyocytes, Leydig cells, intercalated cells and choroid plexus cell. In contrast to the identified 44 human tissues, there are many more mouse tissues (121) with more than $1,000$ cells reported in the database. Among them, only 3 (2.5%) showed an average mtDNA% significantly higher than 5% (whole kidney, whole heart and distal small intestine). Furthermore, only 6 mouse tissues (whole kidney, intestinal epithelium, whole heart, nerve, distal small intestine and submandibular gland) have a portion of the interquartile range over the 5% threshold (Fig. 3.1 D.). These findings indicate that the 5% threshold early proposed in the field is an appropriate standardized threshold for mouse tissues.

To evaluate the effect of mtDNA% in the analysis of single-cell RNA-seq data, we downloaded three datasets from the PanglaoDB database using accessions: SRS3703557, SRS3545826 and SRS2397417. The first dataset contains $9,238$ cells from the mouse heart, the second $7,448$ cells from mouse lung and the third $9,057$ cells from the human umbilical vein.

First, we evaluated the effect of genes encoded in the mitochondrial genome (for short, mtGenes) on the cell clustering results. We compared the low-dimensional representations generated by t-SNE using the PCA result as a prior with and without the mtGenes. We found that, in all three examples, the mitochondrial content does not affect significantly the structure of the data in low dimensional representation, allowing to recover clearly in both cases (with and without mtGenes) the clusters reported by the PanglaoDB database (Fig. 3.2). These results confirm findings previously reported in an evaluation study of different computational pipelines for single-cell RNA-seq data preprocessing [93]. We also found that even without considering the mtGenes for the generation of the low-dimensional representation, low-quality cells with high mtDNA% tend to cluster together (Fig. 3.2 B.; Clusters 19, 0, 9, 7 and 11 in SRS3703557; cluster 5 in SRS3545826 and cluster 11 in SRS2397417).

Next, we evaluated the significance of the threshold to identify low-quality cells. For each example, after identifying the clusters containing cells with high mtDNA% of a cell type, we used MAST to compare their expression profiles against other clusters with most of the cells below the threshold. The fold-change values reported by MAST were used as the input of GSEA analysis to test for the significance of the Apoptosis pathway in the KEGG database.

For the first example, we focused on cardiomyocytes, a cell type associated with high mtDNA%

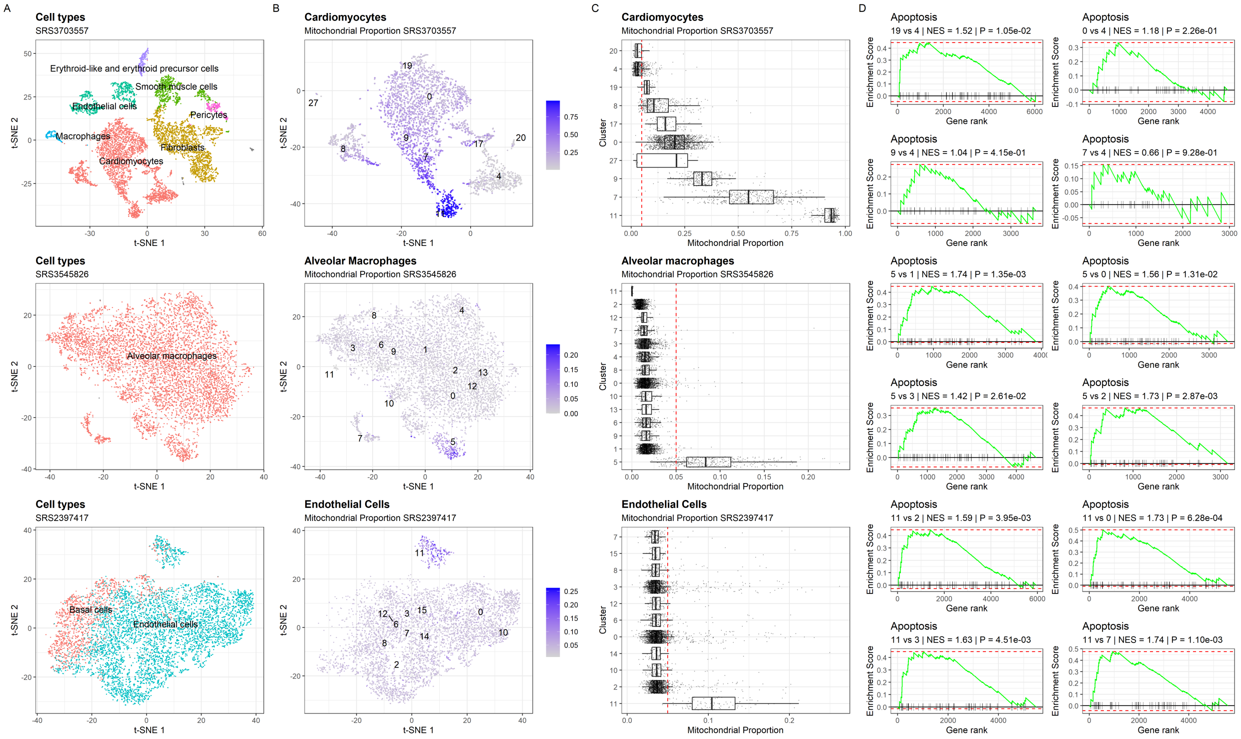


Figure 3.2: Case examples showing the effect of omitting the mtDNA% QC filter in the analysis of single-cell RNA-seq data. (A) t-SNE representation of all the cell populations included in the dataset generated by excluding the mitochondrial genes from the list of highly variable genes before principal component analysis (PCA). Each dot represents a cell and they are colored by cell type. (B) t-SNE representation of cell type used as an example colored in the function of the mtDNA% in each cell. Clusters reported by the PanglaoDB are labeled. (C) Boxplot showing the distribution of the mtDNA% across clusters. The red line is the early established 5% threshold. (D) GSEA analysis of the Apoptosis pathway between clusters with a high proportion of low-quality cells and others containing high-quality cells

with nine clusters reported by the PanglaoDB database (Fig. 3.2 C.) for this dataset. We compared cells with a higher mtDNA% level in clusters 19, 0, 9, 7 and 11, which form a larger cluster, against cells with a lower mtDNA% level in cluster 4. We found that the number of genes with detectable expression decreases with the increase of mtDNA% in cells (as shown by the x-axes of the first two rows of subplots in Fig. 3.2 D.). This anticorrelation is expected as the increased mtDNA% is likely to be associated with cell breakout events. When a breakout occurs (due to the differences in copy numbers given by a single nucleus and several mitochondria by cell), generate an increase in the abundance of mitochondrial content, and reducing the reads that will be mapped to nuclear genes, resulting in fewer genes' expression detected. Indeed, we found even a small increase of the mitochondrial content (comparing cluster 19 and cluster 0 versus cluster 4) led to a huge decrease in the number of expressed genes ($> 6,000$ genes versus $\approx 5,000$ genes, see the first row of subplots in Fig. 3.2 D.). The number of genes included in the GSEA analysis, in turn, influences the value of Normalized Enrichment Score (NES), which is used to assess the significance of the apoptosis pathway. Despite this influence, a positive NES value was recovered in all the cases for the tested cardiomyocytes clusters, as well as in the other two examples of mouse alveolar macrophages (cluster 5 against others) and human endothelial cells (cluster 11 against others), suggesting a consistently higher expression (positive $\log_2$ fold-change) of apoptotic pathway genes among cells with mtDNA% above the threshold (Fig. 3.2 D.).

In summary, we reported a new set of mtDNA% reference values across human and mice tissues and cell types for single-cell RNA-seq QC. Based on our analytical results, we suggest a standardized mtDNA% threshold of 10% for single-cell RNA-seq QC of human samples. For mouse samples, we found that the early defined threshold of 5% accurately discriminates between healthy and low-quality cells, bringing to evidence that under a well-performed single-cell RNA-seq QC, clusters containing cells with high mtDNA% exhibiting signatures of apoptosis, like those shown in the example datasets, should be excluded from being used to make biological interpretations. Thus, we suggest that all published mouse studies, in which single-cell RNA-seq QC was based on the mtDNA% value greater than 5%, should be re-evaluated because the use of any mtDNA%

higher than 5% is likely to be an overshoot over the threshold, resulting in apoptotic cells being utilized in the subsequent analyses.

4. SINGLE-CELL EXPRESSION VARIABILITY IMPLIES CELL FUNCTION *

4.1 Introduction

Cells are fundamental units of cellular function. Cells in multi-cellular organisms can be organized into groups, or cell types, based on shared features that are quantifiable. A multicellular organism is usually composed of cells of many different types – each is a distinct functional entity differing from the other. Within the same cell type, cells are nearly identical and are considered to carry the same cell function or biological processes associated with the cell type that ensures the homeostatic state of the organism where the cell is present.

The recent development of single-cell RNA sequencing (scRNA-seq) technologies has brought the increasingly high-resolution measurements of gene expression in single cells [94]. This power has been widely adopted to refine the categories of known cell types and analyze complex tissues systematically and reproducibly [5]. The power of scRNA-seq has also been harnessed to identify novel cellular states among the same type of cells [6].

Cells of the same type and at the same state may still show marked intrinsic cell-to-cell variability in gene expression or single cell expression variability (scEV), even under the same environmental conditions [95, 96, 7]. The importance of this intrinsic variability is increasingly appreciated [97, 98]. Changes in the magnitude of scEV have been associated with development [99, 100, 101, 102], aging [103, 104], and pathological processes [105, 106].

Dueck and colleagues [12] put forward the so-called ‘variation is function’ hypothesis, saying that scEV per se might be crucial for population-level function. They used the term ‘single cell variation or variability’ to refer to diversity within an ensemble that has been previously defined as being generally homogeneous, rather than diversity of cell types that are clearly distinct and already

This chapter has been reprinted from: **Osorio, Daniel; Yu, Xue; Zhong, Yan; Li, Guanxun; Serpedin, Erchin; Huang, Jianhua Z.; Cai, James J. 2020. ‘Single-Cell Expression Variability Implies Cell Function’ *Cells* 9, no. 1: 14.** Available at: https://www.mdpi.com/2073-4409/9/1/14. Licensed under a Creative Commons Attribution 4.0 International License https://creativecommons.org/licenses/by/4.0/, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

recognized. The main focus of their question is to ask how the individual cells with different gene expression levels may interact to causally generate higher-level function. If the hypothesis turns out to be true, it means that the intrinsic cell-to-cell variability is an indicator of a diversity of hidden functional capacities, which facilitate the collective behavior of cells. This collective behavior is essential for the function and normal development of cells and tissues [107, 108]. The loss of this collective cellular behavior may result in disease. Thus, investigation of the intrinsic cell-to-cell variability may contribute to the understanding of pathological processes associated with disease development.

It is worth noting that the level of intrinsic cell-to-cell variability needs to be measured within a highly homogeneous population of cells. This is because many micro-environmental perturbations and stochastic factors at the cellular level are known to change the scEV. These factors may include local cell density, cell size, shape and rate of proliferation, cell cycle [13, 14, 109, 110, 15]. To work on the cell-to-cell variability, these confounding factors have to be controlled.

Exponential scaling of scRNA-seq has made it feasible to study scEV across thousands of cells [111] and quantify scEV based on measures of statistical dispersion such as the coefficient of variation (CV) [112, 113]. The sheer number of cells sequenced in a 'typical' droplet-based scRNA-seq experiment allows us to filter out for a sizable number of highly homogeneous cells, based on the similarity between their global transcriptional profiles. With these selected core of highly similar cells, we are able to test the 'variation is function' hypothesis. Furthermore, using established statistical methods, we are able to control for many sources of technical variation that may confound the measurement of scEV to obtain an unbiased estimate. For instance, single-molecule capture efficiency, 3' end bias due to single-cell RNA library preparation protocol, and low expression of genes are examples of known sources of technical variation [53], which should be controlled for using statistical means.

The characterization of the impact of scEV on cell function requires the understanding of which genes show greater or less cell-to-cell variability in their expression. These feature genes may carry valuable information that can facilitate the elucidation of underlying regulatory networks

[16]. Once these genes are identified, a follow-up question is whether they are tissue- or cell type-specific – i.e., whether the same genes will be identified for different tissues or cell types. Our working hypothesis is in line with the 'variation is function' hypothesis, that is, different tissues or cell types have different sets of highly variable genes (HVGs), and these HVGs should be enriched with functions that reflect the biological processes associated with the respective tissues or the cell types. To test this, we analyzed three scRNA-seq data sets generated for three different cell types. Each data set contains thousands of cells. For each cell type, we selected a highly homogenous population of cells, with the help of a newly developed dimensionality reduction method, called potential of heat-diffusion for affinity-based trajectory embedding (PHATE) [114]. We estimated scEV among the selected cells for each of these cell types and further systematically characterized functions of identified HVGs. We show that HVGs are highly specific to cell types, i.e., different cell types have different sets of HVGs; functions of HVGs precisely mirror the biological processes of the corresponding cell types.

4.2 Materials and Methods

4.2.1 LCL cell culture and scRNA-seq experiment

The lymphoblastoid cell line (LCL) GM12878 was purchased from the Coriell Institute for Medical Research. They were cultured in the RPMI-1640 medium supplied with 2mM L-glutamine and 20% of non-inactivated fetal bovine serum, incubated at 37°C under 5% CO_2 atmosphere. For maintenance, cells were subcultured every three days by adding fresh medium. For single-cell sequencing, each cell line was subcultured with $200,000$ viable cells/mL. Cells were harvested for single-cell sample preparation and sequencing on day four (stationary phase) following the sample preparation demonstrated protocol and Single Cell 3' Reagent Kits v2 user guide provided by 10× Genomics. Briefly, cells were mixed well in each flask, and 1mL of cell suspensions from each cell line were taken out. The cells were washed three times by centrifuging, suspending, and resuspending in 1× PBS with 0.04% BSA. Viable cells were then counted using an automated cell counter (Thermo Fisher Scientific, Carlsbad, CA, USA). Cells (≈ 5000 per cell line) were then pelleted and resuspended in the nuclease-free water based on the cell suspension volume calculator table, followed by GEM (gel bead-in-emulsions) generation and barcoding, the post-GEM-RT cleanup, cDNA amplification, and library construction and sequencing. The experiments were conducted at the Texas A&M Institute for Genome Sciences and Society. The sequencing was conducted in the North Texas Genome Center facilities using a Novaseq 6000 sequencer (Illumina, San Diego, CA, USA). Raw reads for each cell were analyzed using Cell Ranger (v2.0.0, 10× Genomics, Pleasanton, CA, USA) and the outputs were aligned to the human reference genome (GRCh38) to obtain the counts [115].

4.2.2 Non-LCL scRNA-seq data sets

The scRNA-seq data for lung airway epithelial cells (LAECs) was downloaded from the GEO database using accession number GSE115982. The original data was generated in the study of [116] for CCR10^{-} and CCR10^{+} LAECs. We used the data generated from the CCR10^{-} cells with the sample identifier GSM3204305. The scRNA-seq data for primary dermal fibroblasts (DFs) was generated in the study of [117]. We downloaded the data for unstimulated DFs from the ArrayExpress database using accession number E-MTAB-5988. To support our findings, we processed additional samples and compared the results obtained for the same cell type. We selected the transcriptomic profile of fibroblasts ($7,052$ and $6,503$ cells) from samples extracted from two different lung regions (GEO accessions: GSM2894834 and GSM2894835); and for the iPSC sample, we used another iPSC culture ($9,146$ cells) generated in the study of [118], and downloaded from the ArrayExpress database using accession number E-MTAB-6268. All of these data sets were produced using 10× Genomics scRNA-seq solutions.

4.2.3 Selection of highly homogeneous populations of cells

We used a supervised data analysis method to select highly homogeneous cells based on the scRNA-seq expression profile of each cell. The procedure is summarized in a flowchart. The main steps are as follows. We used Seurat (v2.2.0) [119] to assign each cell into a cell cycle phase and excluded cells that were not considered to be in G1-phase. We removed genes encoded in the mitochondrial genome from the analysis. We then selected and retained cells with a library size between 50 and 95 percentiles. We used PHATE [114] to generate a embedding plot of all remaining cells and inspected the distributions of cells in the three-dimensional plot and manually picked one 'core' cell. Finally, an additional 999 cells that were closest to the core cell, according to the Euclidean distances between cells, were selected to form the final $1,000$-cell population. This selection procedure was applied to each of the three cell types independently, as well as for the additional samples used to support our findings.

4.2.4 Identification of HVGs

Identification of highly variable genes (HVGs) was based on the assumption that high expression variability of these genes across cells relative to their mean expression is caused by biological effects rather than merely technical noise. We used the method proposed in [120], which is implemented in function `'sc_hvg'` of the scGEAToolbox package [121]. This method starts by adjusting the library size and assumes that the observed mean expression (μ_i) and the observed CV^2 ($\hat{w}_i$) of gene i among cells have the following relationship:

$$E\left(\hat{w}_i\right) \approx \frac{a_1}{\hat{\mu_i} + a_0} \tag{4.1}$$

and

$$\frac{\hat{w}_i}{\frac{a_1}{\hat{\mu_i}} + a_0} \sim \frac{X^2_{m-1}}{m-1} \tag{4.2}$$

where m is the number of cells. The values of a_0 and a_1 are estimated by generalized linear regression (GLM). The residual term $\frac{\hat{w}_i}{\frac{a_1}{\hat{\mu_i}} + a_0}$ for each gene is used to test if the observed CV^2 is significantly larger than the expected CV^2 via a chi-squared test. Multiple testing p-value adjustments were performed by controlling FDR [122].

4.2.5 Function enrichment analyses

To identify the overrepresented biological functions of HVGs in different cell types, we performed the GO enrichment analysis using Enrichr [123, 124] and GOrilla [125]. Enrichr was conducted for HVGs (FDR < 0.01) against the rest of the expressed genes with respect to pathways collected in the Reactome pathway knowledgebase [126]. GOrilla was performed with the list of genes sorted in descending order of their residual variability.

4.2.6 Analyses of co-expression network and regulatory regions of HVGs

MAGIC [24] was used to impute the expression matrix. The co-expression networks were constructed using 1-correlation as a distance measure, using SBEToolbox [127]. The motif analysis of the regulatory regions associated with the HVGs was performed using the GREAT [128]. Genomic coordinates for the HVG genes from the Human Reference Genome (hg19) were downloaded from the Ensembl Biomart [129] and converted into bed format using an in-house script. Identified motifs were searched against the JASPAR database [130] to match the binding sites of corresponding TFs.

4.2.7 Data availability

The data sets used in this study and computer code are available.

1. LCLs GM12878 scRNA-seq data in the GEO database with accession number GSE126321: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126321

2. LAEC scRNA-seq data in the GEO database with accession number GSE115982: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115982

3. DF scRNA-seq data in ArrayExpress database with accession number E-MTAB-5988: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5988

4. Human iPSC scRNA-seq data in ArrayExpress database with accession number E-MTAB-6687: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6687

5. Fibroblasts scRNA-seq data in GEO database with accession number GSM2894834: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2894834

6. Fibroblasts scRNA-seq data in GEO database with accession number GSM2894835: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2894835

7. Human iPSC scRNA-seq data in ArrayExpress database with accession number E-MTAB-6268: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6268/

8. Computer codes used to analyze data: https://github.com/cailab-tamu/HVG

9. The following are available online at https://www.mdpi.com/2073-4409/9/1/14/s1, Figure S1: Flowchart of selection of a highly homogeneous population of cells; Figure S2: UMAP embedding plots for the $1,000$ selected cells of each of three cell types. Figure S3: Scatter plots between mean and residual CV^2. Figure S4: Comparison between two methods: HVG [120] and VEG [131], for correcting mean-variance dependency. Figure S5: Venn diagram showing overlap between HVGs identified in three cell types. Figure S6: Correlation between scEV and population-level expression variability across genes of functional sets. Figure S7: PHATE 3-D embedding plot for cells colored according to *IGLC2* expression level in cells. Table S1: HVGs of lymphoblastoid cell line (LCL) cells (FDR < 0.01). Table S2: HVGs of lung airway epithelial cells (LAECs) (FDR < 0.01). Table S3: HVGs of dermal fibroblasts (DFs) (FDR < 0.01). Table S4: Significantly enriched motifs found in HVGs. Table S5: HVGs of induced pluripotent stem cell (iPSCs) (FDR < 0.01).

4.3 Results

4.3.1 Single-cell RNA sequencing and selection of highly homogenous cells

In this study, we experimented with the transcriptomic profiles of three different human cell types, namely, lymphoblastoid cell line (LCL), lung airway epithelial cell (LAEC), and dermal fibroblast (DF). We estimated single-cell expression variability (scEV) for each of these cell types, individually.

To obtain the scRNA-seq data for LCL, we cultured GM12878, an LCL strain widely used in genomic research, prepared cells using a 10× Genomics Chromium Controller, and sequenced a total of $7,045$ cells [115]. This data has been deposited in the NCBI Gene Expression Omnibus (GEO) database (accession number GSE126321). For the other two cell types, LAEC and DF, we obtained the scRNA-seq data for $3,863$ and $2,553$ cells from the studies of [116] and [117], respectively. In addition, we also processed two more samples for fibroblasts and one for iPSC cells (see Section 4.2.7 for data availability) to cross-validate our findings. All scRNA-seq data sets of the three cell types and the additional samples used were produced using 10× Genomics droplet-based solution and made use of unique molecular identifiers (UMIs) [132].

For each cell type, we employed a data analysis procedure, a filter pipeline on scRNA-seq data, to select highly similar populations of cells (see Section 4.2 for materials and methods). These selected cells are a representative population of each the cell type. Briefly, we first excluded mitochondrial DNA-encoded genes from the analysis. We then excluded cells in the S- or G2/M phases and only retained G1-phase cells. We also excluded cells with library size < 55 percentile or > 99 percentile. Finally, we used PHATE to produce the low-dimensionality representation of the remaining cells to inspect between-cell structure driven by heterogeneity in gene expression. PHATE is a visualization method that captures both local and global nonlinear structure in data by an information-geometry distance between data points [114]. As seen from the PHATE projection (Figure 4.1 A.), several 'arms' of cells show the structure of the cell-to-cell relationship. Based on the observation, we manually picked one 'core' cell at the root of the arms of cells in the middle of the cell cloud (red circle in Figure 4.1 A.). The core cell and 999 nearest cells around it were then selected using the k-nearest neighbors algorithm to form the final population of $1,000$ cells, which was used for subsequent data analyses.

To examine the homogeneity of selected cells, we used t-distributed stochastic neighbor embedding (t-SNE) [133] to position all $1,000$ selected cells in the two-dimensional t-SNE space. Compared to PHATE, t-SNE is a more commonly used nonlinear visualization algorithm for revealing structures in high-dimensional data, emphasizing local neighborhood structure within the data. When running t-SNE, we experimented with a series of perplexity values to produce multiple plots for the same population of selected cells. t-SNE is known to be sensitive to hyperparameters

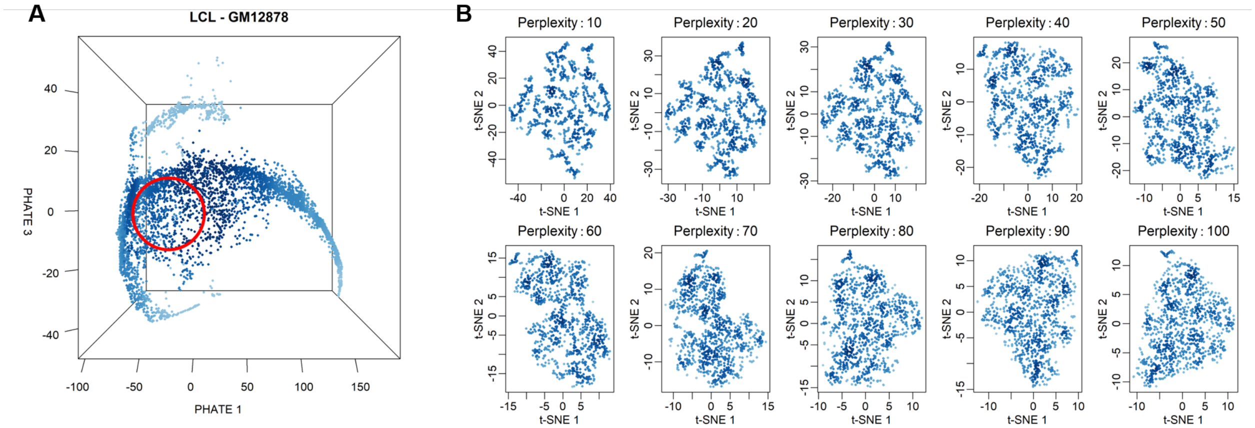


Figure 4.1: Selection of a highly homogenous cell population for variability analysis. (A) Three-dimensional PHATE embedding plot for G1-phase cells of GM12878. Each point represents a single cell in the three-dimensional space. The red circle indicates the approximate positions of 1000 selected cells. (B) Embedding plots generated for the 1000 selected cells with a t-SNE algorithm with a series of perplexity values.

[134]. In general, when different parameter values are given, t-SNE tends to produce different cell clustering plots. However, for our selected cells, no structure is observed in any of these t-SNE embedding plots (Figure 4.1 B.). The same results were obtained for the other two cell types as well as using the uniform manifold approximation and projection (UMAP) as an alternative embedding algorithm [134] (Supplementary Figure S2). Thus, we confirm that cells selected with our filter pipeline are highly homogenous populations of representative cells for each cell type.

4.3.2 Identification of highly variable genes

Highly variable genes (HVGs) are expressed variably across homogeneous cells of the same type. For each cell type, we used the method of [120] to identify HVGs from scRNA-seq data of the homogeneous population of selected cells. In this method, the relationship between the squared coefficient of variation (CV^2) of genes and their average expression (μ) is considered. The relationship between log-transformed CV^2 and log-transformed μ is fitted with a generalized linear model (GLM) by using a gamma distribution, and the expected CV^2 for a given μ is calculated with the fitted curve. The log-transformed ratio between observed CV^2 and expected CV^2 [= log(observed CV^2) - log(expected CV^2], called ‘residual variability’, is used as the measurement of scEV. Since

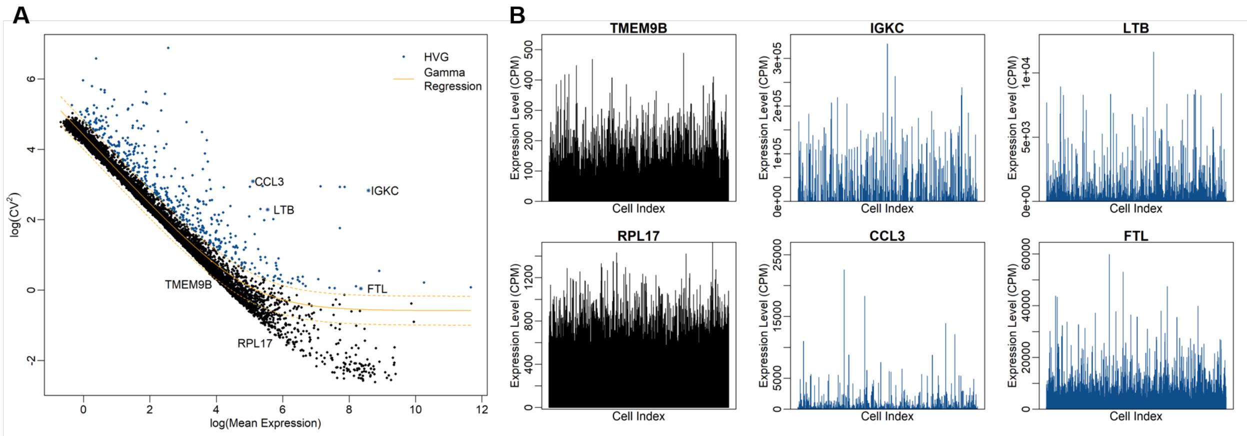


Figure 4.2: Identification of highly variable genes (HVGs). (A) The relationship between CV2 and mean expression of genes in LCL GM12878. The orange line shows the trend for the gamma GLM fit curve between CV2 and mean expression and used to identify HVGs. For each gene, the residual variability is calculated as the difference between observed CV2 and expected CV2 from the fitted curve. (B) Expression profiles of selected HVGs and lowly variable genes across cells. Cells are unsorted and remain a random order. Each vertical line is a cell, and the height of line indicates the level of gene expression in counts per million (CPM) in a cell.

the expected CV^2 captures the variability originated from technical noise, the residual variability is considered to be an unbiased measure of biological variability. Indeed, after the correction, the μ-CV^2 correlation disappeared (Supplementary Figure S3). We repeated the procedure for correcting -CV dependency using another method [131] and obtained the qualitatively similar results in terms of identified HVGs and enriched functions (Supplementary Figure S4). Here, we only report the results obtained using the method of [120].

After the μ-CV^2 dependency correction, we identified 465, 466, and 364 HVGs at a false-discovery rate (FDR) of 0.01 for LCL, LAEC, and DF, respectively (Supplementary Tables S1–3). To visualize the expression variability of genes, we plot CV^2 against μ, both on the logarithmic scale, for LCL (Figure 4.2 A.). Each dot represents a gene; all genes together give a characteristic cloud showing the μ and CV^2 of gene expression. Genes above the GLM fitting curve, e.g., *IGKC*, *CCL3*, *LTB*, and *FTL*, are more variable than expectation, whereas genes below the curve, e.g., *TMEM9B* and *RPL17*, are less variable (Figure 4.2 B.).

4.3.3 Cell-type origin determines the function of highly variable genes

To assess the biological functions of HVGs in different cell types, we performed enrichment analyses. We found that enriched gene ontology (GO) terms are largely distinct and reflect respective cell functions of each of the three cell types (Table 4.1). For example, LCL HVGs (e.g., *CCL22* and *IFI27*) are more likely to be involved in cytokine- or interferon-signaling pathways, and also, more generally, the innate immune system; LAEC HVGs (e.g., *COL1A1*, *MMP1*, and *IL17C*) collagen formation and extracellular matrix organization; DF HVGs (e.g., *KRT14*, *ACAN*, and *FLG*) keratinization and regulation of cell proliferation. DF HVGs also include *SFRP2*, *DPP4*, and *LSP1*, which are marker genes defining major fibroblast subpopulations in human skin [135]. Taken together, these results show that different cell types have different sets of HVGs with substantial scEV, associated with cell-type-specific functions.

If two cell types have shared function, then we expect to see the overlap in their HVG-associated functions. This, indeed, is the case. There are some overlaps between enriched functions between the three cell types we examined here. For example, the cytokine signaling pathway is enriched for both LCL and LAEC, and extracellular structure organization is enriched for both LAEC and DF. Meanwhile, across all three cell types, there are 14 shared HVGs genes (*CDC20*, *CLEC2B*, *CLIC3*, *CTSC*, *CYP1B1*, *DUSP2*, *HES1*, *MT1E*, *NPW*, *SOX4*, *STMN1*, *TK1*, *TRIB3*, and *UCHL1*; Supplementary Figure S5), with diverse cellular and molecular functions.

4.3.4 Functions associated to HVGs are conserved across tissues and subpopulations of cells

To determine how stable the functions associated with the HVGs identified in a cell type are, we used additional samples and random selection of the core cell. We analyzed two other fibroblast samples obtained from different body tissues (see Section 4.2.7 for data availability), as well as the initially included dermal sample. For each sample, after quality control, we randomly selected a core cell and their 999 more similar cells; then, we identified the set of HVGs (FDR < 0.01 and fold-change > 1.5) present in each subpopulation as described before (see Section 4.2 for materials and methods). Under these thresholds, we identified 221, 226, and 228 HVGs for dermal,

Cell Type	Highly Variable Genes, Top 50	Enriched GO Terms, Top 10	Enriched Reactome Pathways, Top 10
Lymphoblastoid Cell Line (LCL)	*ANKRD37 ATF3 BIN1 BMP4 CAMP CCL22 CCL3 CCL3L3 CCL4 CCL4L2 CCR7 CD69 CD7 CD83 CDKN1A CTSC CYP1B1 DHRS9 DUSP2 FSCN1 HIST1H1C IER3 IFI27 IGHG1 IGHG3 IGHM IGKC ITM2A KCNMA1 LINC00176 LINC01588 LMNA LTA LTB MAL MIER2 MIR155HG MYC NFKBIA PMCH PRSS2 RGS1 RGS16 RGS2 RP11-291B21.2 S100A4 SFN TNFAIP2 TUBB4B WFDC2*	Signal transduction (GO:0007165) Response to stimulus (GO:0050896) Immune response (GO:0006955) Response to biotic stimulus (GO:0009607) Immune system process (GO:0002376) Response to external biotic stimulus (GO:0043207) Response to external stimulus (GO:0009605) Cytokine-mediated signaling pathway (GO:0019221) Defense response (GO:0006952) Response to chemical (GO:0042221)	Immune system (R-HSA-168256) Chemokine receptors bind chemokines (R-HSA-380108) Interferon alpha/beta signaling (R-HSA-909733) Cytokine signaling in immune system (R-HSA-1280215) Interferon signaling (R-HSA-913531) Peptide ligand-binding receptors (R-HSA-375276) G alpha (i) signaling events (R-HSA-418594) Innate Immune System (R-HSA-168249) Interferon gamma signaling (R-HSA-877300) Cell cycle (R-HSA-1640170)
Lung airway epithelial cell (LAEC)	*AMTN ANKRD1 AREG CCL2 CCL5 CCL7 COL1A1 COL1A2 COL3A1 COL6A1 COL6A3 CRCT1 CTGF CXCL5 CXCL6 FBXO32 GREM1 HAS2 IFNL1 IFNL2 IFNL3 IGFBP5 IGFL1 IL17C IL23A KRT14 KRT6B KRT81 LY6D MEG3 MMP1 MSMB OVOS2 PI3 POSTN PPBP RP11-338I21.1 S100A7 S100A8 S100A9 SERPINB2 SERPINB3 SERPINB4 SLC15A2 SPARC SUGCT SULF1 TEX26-AS1 TNFAIP6 TSLP*	Regulation of multicellular organismal process (GO:0051239) Regulation of signaling receptor activity (GO:0010469) Response to stimulus (GO:0050896) Regulation of cell proliferation (GO:0042127) Developmental process (GO:0032502) Extracellular matrix organization (GO:0030198) Response to chemical (GO:0042221) Response to organic substance (GO:0010033) Regulation of developmental process (GO:0050793) Regulation of response to stimulus (GO:0048583)	Extracellular matrix organization (R-HSA-1474244) Assembly of collagen fibrils and other multimeric structures (R-HSA-2022090) Cytokine signaling in immune system (R-HSA-1280215) Collagen formation (R-HSA-1474290) Signaling by interleukins (R-HSA-449147) Chemokine receptors bind chemokines (R-HSA-380108) Peptide ligand-binding receptors (R-HSA-375276) Collagen biosynthesis and modifying enzymes (R-HSA-1650814) Integrin cell surface interactions (R-HSA-216083) Class A/1 (rhodopsin-like receptors) (R-HSA-373076)
Dermal fibroblast (DF)	*ACAN ACTA2 ACTC1 CEMIP CLU COMP CTSC CXCL1 DCN DKK1 FLG G0S2 GAL HIST1H4C IGFBP5 IGFBP7 IL1RL1 KCNMA1 KRT14 KRT17 KRT19 KRT34 KRT81 KRTAP1-5 KRTAP2-3 LCE1F LUM MGP MMP1 MMP3 MT1X NMB OLFM2 PCP4 PENK PGF PI16 POSTN PPP1R14A PTTG1 PTX3 RARRES2 RGCC SCG5 SERPINE2 SFRP2 SFRP4 STMN2 TFPI2 TNFRSF11B*	Regulation of signaling receptor activity (GO:0010469) Developmental process (GO:0032502) Keratinization (GO:0031424) Anatomical structure development (GO:0048856) Regulation of cell proliferation (GO:0042127) Regulation of multicellular organismal process (GO:0051239) Extracellular matrix organization (GO:0030198) Extracellular structure organization (GO:0043062) Response to oxygen-containing compound (GO:1901700) Multicellular organismal process (GO:0032501)	Extracellular matrix organization (R-HSA-1474244) Regulation of insulin-like growth factor (IGF) transport and uptake by insulin-like growth factor binding proteins (IGFBPs) (R-HSA-381426) ECM proteoglycans (R-HSA-3000178) Hemostasis (R-HSA-109582) Platelet degranulation (R-HSA-114608) Dissolution of fibrin clot (R-HSA-75205) Response to elevated platelet cytosolic Ca2+ (R-HSA-76005) Negative regulation of TCF-dependent signaling by WNT ligand antagonists (R-HSA-3772470) GPCR ligand binding (R-HSA-500792) Peptide ligand-binding receptors (R-HSA-375276)

Table 4.1: Representative highly variable genes (HVGs) identified in the three cell types: LCL, LAEC, and DF, and the results of functional enrichment analyses. Genes are sorted by residual variability. The top 50 genes with the highest residual variability values are selected as representative HVGs.

Shared Highly Variable Genes	**Enriched GO Terms, Top 5**	**Enriched Reactome Pathways**
ID3 MARCKSL1 DPT ID2 CYP1B1 IGFBP2 IGFBP5 APOD IGFBP7 SFRP2 PLK2 SOX4 PI16 CTGF SGK1 CITED2 IGFBP3 SERPINE1 TIMP1 OSR2 HAS2 MYC B4GALT1 PTGDS CRYAB BAMBI MFAP5 CSRP2 LUM PGF THBS1 FGF7 MT2A MT1X WISP2 ADAMTS1	Collagen-containing extracellular matrix (GO:0062023) Extracellular matrix (GO:0031012) Epithelial cell proliferation (GO:0050673) Negative regulation of cell migration (GO:0030336) Extracellular matrix organization (GO:0030198)	Extracellular matrix organization (R-HSA-1474244) Regulation of insulin-like growth factor (IGF) transport and uptake by insulin-like growth factor binding proteins (IGFBPs) (R-HSA-381426)

Table 4.2: Shared highly variable genes (HVGs) identified in the three fibroblast samples: dermal, lung distal, and lung proximal, and the results of functional enrichment analysis.

lung distal, and lung proximal fibroblasts, respectively. Among the identified HVGs from the three samples, we found 36 genes that statistically enrich (FDR < 0.05) for specific biological processes historically associated with fibroblasts (Table 4.2). The small overlap found, as well as the functional enrichment for the extracellular matrix, were previously described in [136, 137], where it is shown that fibroblasts are a remarkably plastic cell type differing between human tissues where they develop unique morphologies and physiologic functions but still have a commonly associated role, the extracellular matrix organization, and maintenance.

4.3.5 HVGs as part of the regulatory network with high cell-type specificity

Next, we set out to test whether HVGs are co-expressed and thus tend to form co-expression networks [138]. We first imputed the expression matrix and then constructed the co-expressed network using the top 50 HVGs for each cell type. For LCLs, the network contains two main modules centered on *NFKBIA* and *IGHG1*, respectively (Figure 4.3 A.). *NFKBIA* encodes the *NF-κB* inhibitor that interacts with *REL* dimers to inhibit *NF-κB*/*Rel* complexes [139, 140]. For LAECs, two modules are centered on *IL23A*/*TNFAIP6* and *COL1A1* (Figure 4.3 B.); for DF, *KRTAP2-3* and *IGFBP7* (Figure 4.3 C.). Thus, functions of 'hub' genes in HVG co-expression networks are closely relevant to the function of corresponding cell type. These results are another line of evidence that scEV implies cell function. The transcription of multiple HVGs may be involved in the same underlying regulatory activities, giving rise to the co-expression network, as we observed. Thus, we wondered whether scEV in several different HVGs is driven by activities of one or few common TFs. To address this question, we searched for upstream regulators of the HVGs defined

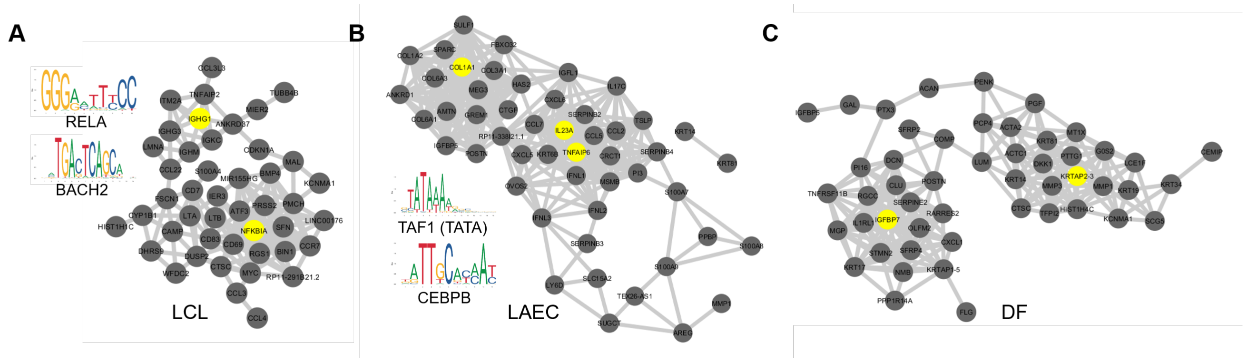


Figure 4.3: Co-expression networks of top HVGs. (A) Co-expression network between most-variable HVGs of LCL and two enriched binding motifs identified in these HVGs. (B) and (C) are for LAEC and DF, respectively. Genes labeled in yellow are the ones acting as a "hub" with high betweenness centrality and closely relevant to the cell-type function.

by our analysis (see Section 4.2 for materials and methods). We identified significant enriched TF binding motifs upstream of HVGs, four for LCL, and five for LAEC (Supplementary Table S4). No significantly enriched motif was identified for DF. The known motifs of LCL HVGs include that of the *NF-κB* subunit gene, *RELA*, and that of *BACH2* (Figure 4.3 A.). The known motifs of LAEC HVGs include the TATA box and that of *CEBPB* (Figure 4.3 B.).

To further explore the involvement of HVGs in the cell type-specific regulatory network, we focused on LCL HVGs in a well-studied gene regulatory network that orchestrates B cell fate dynamics [141, 142, 143]. This known regulatory network involves eight genes, including three LCL HVGs – *PRDM1* (or *Blimp-1*), *AICDA* (or *AID*), *IRF4*, two key regulatory genes with binding motifs enriched in targeting LCL HVGs (see above) – *RELA* and *BACH2*, and three other key regulators – *BCL6*, *PAX5*, and *REL* (*cRel*) (Figure 4.4 A.).

We examined the inter-relationship between across-cell expressions of three LCL HVGs (Figure 4.4 B.). The scatter plot shows that the directionality of the correlation between *AICDA* and *IRF4* depends on the expression level of *PRDM1*. Among cells with relatively low expression of *PRDM1*, expressions of *AICDA* and *IRF4* are negatively correlated. Whereas, among cells in which *PRDM1* is highly expressed, expressions of *AICDA* and *IRF4* are positively correlated. This nonlinear relationship between expressions of HVGs suggests they are embedded in a tightly

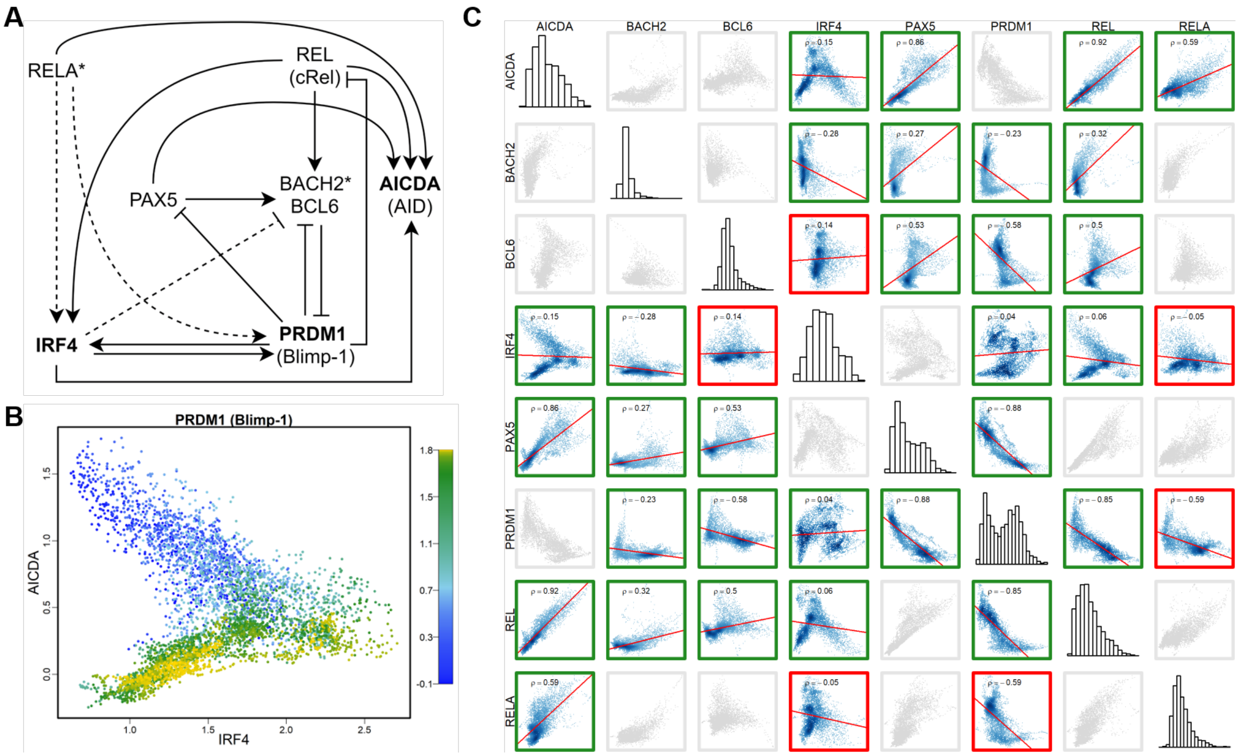


Figure 4.4: Gene regulatory network and correlation matrix of LCL HVGs. (A) An *NF-κB* regulatory network model for activated B cell (ABC) - antibody secreting cell (ASC) differentiation, modified from [143]. Bold font indicates HVGs; asterisk indicates the upstream TFs targeting HVGs; solid line dashed line indicates the regulatory relationship supported by the correlation between two corresponding genes, and the dashed line indicates regulatory relationship not supported by the expression correlation between genes. (B) Scatter plot of cells, showing the correlation between expression levels of three HVGs: *IRF4*, *AICDA* (*AID*), and *PRDM1* (*Blimp-1*). The color bar indicates the expression level of PRDM1 (*Blimp-1*). (C) Spearman correlation matrix between expression levels of eight genes involved in the model. Green boxes indicate that the sign of the correlation between two genes is consistent with the effect (induction/repression) of the relationship between the two in the regulatory model. Red boxes indicate inconsistency, while gray boxes indicate no direct relationship according to the model.

regulated expression network. Thus, we examined the all-by-all Spearman correlation between expressions of all eight genes in this regulatory network using the imputed data of the homogenous LCLs (Figure 4.4 C.). By comparing the sign of correlation coefficient of each pair of genes with the regulatory effect of the gene pair in the model network, we found that the correlation matrix can be used to correctly recover 15 out of 18 direct regulatory relationships. The result suggests that, even in this highly homogenous population of LCLs, cells retain gene regulatory network activities that orchestrate cell fate dynamics as in their original B cells.

4.3.6 Single-cell expression variability in LCLs is positively correlated with between - individual expression variability

Next, we examined the relationship between scEV and inter-individual expression variability. We distinguish between the two different types of variabilities at different organizational levels. Specifically, the former is cell-to-cell variability in a population of cells, and the latter is inter-individual variability at the human population level. We again focused on LCLs, for which population-scale gene expression data are available from the Geuvadis RNA-seq project of $1,000$ Genomes samples. The bulk RNA-seq data was downloaded as a normalized expression matrix of FPKM values. We retained data for all LCLs of European ancestry (CEU) [44]. With the residual variability estimated from scRNA-seq of GM12878 and that estimated from the CEU population, we tested the correlation between the two estimates across genes. When the test was conducted with all genes ($n = 8,424$), we obtained a significant but weak positive correlation (SCC, $\rho = 0.19$, P $= 1.2 \times 10^{-9}$). We wondered whether this positive correlation was driven by subsets of genes. To identify these gene sets, we conducted the correlation tests for the GO-defined gene sets one by one. Across all gene sets tested, the average SCC for gene sets defined by GO biological process (BP) and molecular function (MF) terms are on average $\rho = 0.28$ and $\rho = 0.23$, respectively. Strikingly, we found a small number of gene sets that produced SCC much higher than averages. The functions of these gene sets include B-cell activation involved in immune response (GO:0002322), cytokine receptor activity (GO:0004896), cellular response to drug (GO:0035690), and regulation of tyrosine phosphorylation of stat protein (GO: 0042509; Figure

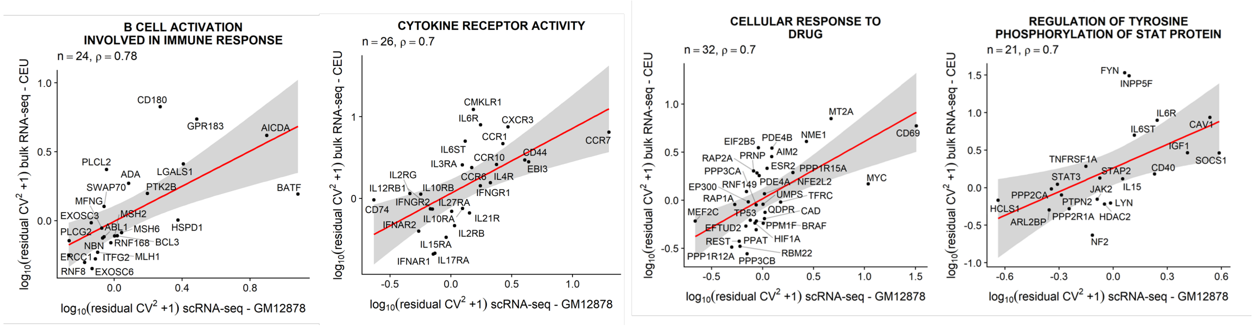


Figure 4.5: Correlation between scEV (i.e., residual variability estimated from LCL GM12878) and the population-level expression variability (measured in LCLs derived from unrelated individuals of European ancestry, CEU) between genes of selected gene sets. More examples can be found in Supplementary Figure S6.

4.5), as well as leukocyte chemotaxis (GO: 0030595) and phospholipase activity (GO:0004620; for more examples, see Supplementary Figure S6). Thus, for these gene sets, scEV may contribute to the establishment of between-individual expression variability.

4.3.7 No enriched functions associated with HVGs identified in human induced pluripotent stem cells (iPSCs)

Finally, we argued, if scEV is the indicator of cell type-specific function, then scEV in undifferentiated cells should not be associated with any cellular functions. To test this, we obtained the scRNA-seq data from the study of [144] (see Section 4.2.7 for data availability). The data was generated from human iPSCs [145]. Same as other cell types examined in this study, these iPSCs were also prepared using the 10× Genomics Chromium controller. The released data contains five samples. We used the first batch (Sample 4) of the data to perform the HVG detection and function enrichment tests, using the same procedure applied to other cell types. When plotting the relationship between log-transformed CV^2 and log-transformed average expression (μ), we found almost no genes showing large CV^2 deviated from the regression curve (Figure 4.6 A.) –a pattern differs substantially from those of the other three cell types (Figure 4.6 B.). This pattern suggests that, for the majority of genes in iPSCs, scEV can be explained by technical noise or sampling stochasticity. In other words, iPSCs lack biological variability in their single-cell expression. Nevertheless,

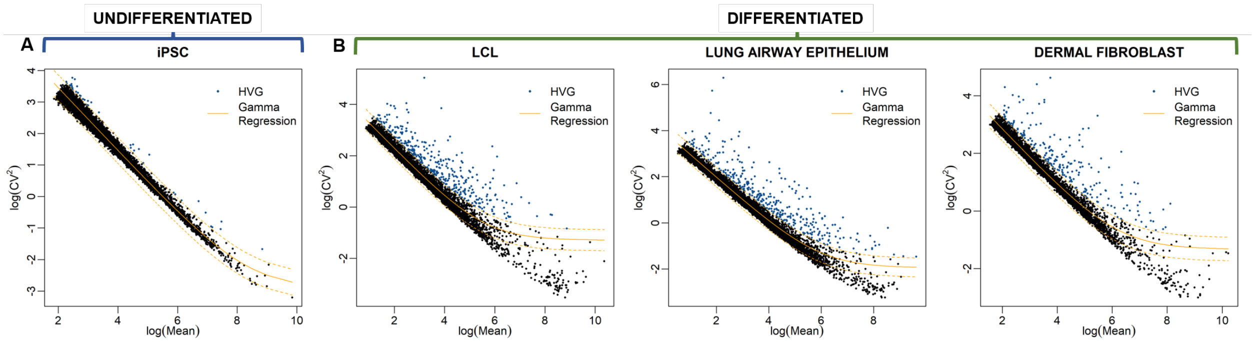


Figure 4.6: Comparison of the magnitude of single-cell expression variability (scEV) among genes between (A) undifferentiated induced pluripotent stem cells (iPSCs) and (B) three differentiated cell types: Lymphoblastoid cell line (LCL), lung airway epithelial cell (LAEC), and dermal fibroblast (DF). For each cell type, the relationship between coefficient of variation squared (CV2) and mean expression of genes is shown.

we still identified 79 iPSC HVGs (Supplementary Table S5) but could not associate any significant (FDR < 0.05) enriched function with them. To further validate our findings, we performed the same analysis using another iPSC sample (see Section 2.7 for data availability) recovering the same pattern, a low number of HVGs (4) that are not significative enriched for a specific function (*ID3*, *LEFTY1*, *MALAT1*, *TAGLN*). These negative results are consistent with our prediction given by the ‘variation is function’ hypothesis: undifferentiated iPSCs are not expected to be associated with any cell-type-specific function.

4.4 Discussion

Single-cell expression variability (or scEV) is sometimes called gene expression noise, emphasizing the stochastic nature of transcriptional activities in cells [146, 147]. Interrogating scEV data has provided insights into gene regulatory architecture [148, 149]; manipulating the magnitude of scEV, through using noise enhancers or scEV-modulating chemicals, has been an approach to achieve drug synergies [150]. Understanding the origin and functional implications of scEV has long been appreciated [95, 96, 7, 151].

In this study, we focused on scEV in human cells. More specifically, we characterized different genes’ expression variability levels within a highly homogeneous population of genetically identical (or nearly isogenic) cells under the same environmental condition. We quantified scEV in highly homogeneous populations of a sizable number of viable cells. Working with cells of the same type, for example, LCL, we started by preprocessing data from thousands of cells. We found that, even though we had firstly preprocessed the data and retained only cells with similar library size and in the same cell cycle phase, it was not enough. There were still marked substructures, shown as branches of cells, in the embedding cloud of cells (Figure 4.1 A.), as revealed by the new embedding algorithm [114]. Retrospectively, we applied the trajectory analysis and found out that one of the longest branches contained cells with elevated expression of immunoglobulin genes (Supplementary Figure S7).

Similarly, marked substructures were observed in the embedding plots of the other two cell types, LAEC and DF. Genes that were differentially expressed and drove the formation of branches of LAECs and DFs were different from those in LCL cells. Thus, there is no single or a small set of marker genes that can be used to capture cellular heterogeneity across different cell types, making the definition of populations of homogenous cells a tedious task. Our work represents the first study focused on comparing scEV in highly homogeneous cell populations across genes in different cell types.

We showed that scEV estimated from homogeneous populations of cells for different cell types carries information on cell type-specific function. Information on molecular functions of cells and biological processes of a given cell type can be extracted from a set of highly variable genes (HVGs), bearing significant biological meaning (see also [30]). HVGs detected in different cell types do not overlap and can reveal the subtle differences in cellar functions between cell types. These conclusions are reached based on our investigation of three cell types and their corresponding HVGs.

First, LCLs are usually established by in vitro infection of human peripheral blood lymphocytes by the Epstein–Barr virus. The viral infection selectively immortalizes resting B cells, giving rise to an actively proliferating B cell population [32]. B cells genetically diversity by rearranging the immunoglobulin locus to produce diverse antibody repertories that allow the immune system to recognize foreign molecules and initiate differential immune responses [13, 64, 63]. LCLs are produced through the rapid proliferation of few EBV-driven B cells from the blood cell population [65]. Thus, scRNA-seq data sets of LCLs offer a 'snapshot' of highly diverse immunoglobulin rearrangement profiles in a much larger population of polyclonal B cells established in donors of these cell lines. Therefore, it is not unexpected to see quite a few immunoglobulin genes in the top list of HVGs identified in LCLs. In addition to these immunoglobulin genes, a number of other immune genes, especially C-C motif chemokine ligands (CCLs) and C-C motif chemokine receptors (CCRs), are in the list of HVGs of LCL. These genes play important roles in allowing the coordination of the activity of individual cells through intercellular communication, essential for the immune system maintains robustness [152]. The HVG co-expression network analysis revealed the key role of the *NF-κB* pathway in facilitating communications between immune cells [107, 13]. More strikingly, we were able to reconstruct nearly the entire *NF-κB* regulatory network, underlying a differentiation of activated B cells and antibody-secreting cells, by using the correlation and anti-correlation relationships between expressions of HVGs and their regulatory genes.

Second, LAEC is a key cell type playing important roles in lung tissue remodeling, and pulmonary inflammatory and immune responses [153]. The airway epithelium, playing a critical role in conducting air to and from the alveoli, is a dynamic tissue that normally undergoes slow but constant turnover. In the event of mild to moderate injury, the airway epithelium responds vigorously to re-establish an epithelial sheet with normal structure and function. HVGs identified in LAECs, which are enriched with genes involved in collagen formation, regulation of cell proliferation, and extracellular matrix organization, accurately elucidate this aspect of functions of the airway epithelium. LAECs are also central to the defense of the lung against pathogens and particulates that are inhaled from the environment. This aspect of functions is also reflected in the enriched functionality of LAEC HVGs.

Third, DFs are responsible for generating connective tissue and play a critical role in normal wound healing [136]. DFs are also commonly used in immunological studies [117, 154, 155].

HVGs identified in DFs again accurately reflect these primary aspects of DF functions, including extracellular matrix organization, keratinization, and regulation of signaling receptor activity. DF HVGs do have several categories of enriched functions overlap with those of LAEC, which is not unexpected, given that DF and LAEC have functional overlaps [156].

Our results provide evidence supporting the 'variation is function' hypothesis, proposed by [12], suggesting that the aggregate cellular function may depend on scEV. Dueck and colleagues also laid down several scenarios, including bet hedging, response distribution, fate plasticity, and so on, in which the establishment of the relationship between scEV and cell function could be attained. Our analytical framework using scRNA-seq data may be utilized in appropriate systems to test the plausibility of these different scenarios. If scEV is an accountable and credible surrogate of cell function, as we have shown in this study, then quantifying and characterizing scEV may become a first-line approach for understanding the function of cell types and tissues. Indeed, when we applied this framework to scRNA-seq data from human iPSCs, we observed no enriched gene functions and no regulatory pathways/networks associated with HVGs in iPSCs. This anti-example, showing no variation no function, further validates the "variation is function" hypothesis.

Furthermore, we have shown that, across certain sets of genes, scEV is positively correlated with population-level expression variability. This correlation provides a new possibility to design single-cell assays with one sample to approximate the population variability of certain genes' expression. This new method may be used to study disease-causing expression dysregulation because it has been a number of cases that increased population-level expression variability has been linked with diseases [157, 158, 159, 160, 161].

Pelkmans [98] pointed out in a visionary perspective article that: 'Embracing this cell-to-cell variability as a fact in our scientific understanding requires a paradigm shift, but it will be necessary'. Indeed, scRNA-seq technologies have brought revolution to gene expression analysis. The technical development gives us a new approach beyond the capacity of traditional methods that rely on experimental measurements of population-average behavior of cells to conceive regulatory network models and signal processing pathways. More importantly, for traditional methods, by averaging information across many cells, differences among cells, which may be important in explaining mechanisms, can be lost. Given the large degree of cell-to-cell expression variability even between genetically homogeneous cells, conclusions reached as for such with traditional average-based methods may be of low-resolution, incomplete, and sometimes misleading [6, 107, 16, 162].

5. SCTENIFOLDNET: A MACHINE LEARNING WORKFLOW FOR CONSTRUCTING AND COMPARING TRANSCRIPTOME-WIDE GENE REGULATORY NETWORKS FROM SINGLE-CELL DATA *

5.1 Introduction

A gene regulatory network (GRN) is a graph depicting the intricate interactions between transcription factors (TFs), associated proteins, and their target genes, reflecting the physiological condition of the cells in question. The analysis of GRNs promotes the interpretation of cell states, cell functions, and regulatory mechanisms that underlie the dynamics of cell behaviors. Multiple methods have been developed to build GRNs from data of gene expression [163, 164, 165, 166]. It is important to compare GRNs constructed using datasets from different samples because the comparison may reveal regulatory mechanisms leading to transcriptomic changes. In particular, the comparison results may help us understand what is the most significant shift in regulatory mechanisms between samples, as well as how genetic and environmental signals are integrated to regulate a cell population's physiological responses and how cell behavior is affected by various perturbations. All of these are key questions in the study of the functional participation of given GRNs. Despite the critical importance of comparative GRN analysis, relatively few methods have been established to compare GRNs [167].

Single-cell RNA-sequencing (scRNA-seq) technology has been revolutionizing the biomedical sciences in recent years. New research provides an unparalleled degree of precision in analyzing transcriptional regulation, cell history, and cell interactions with rich knowledge. It transforms previous entirely tissue-based assays into transcriptomic single-cell measurements and greatly en-

This chapter has been reprinted from: **Osorio, Daniel, Yan Zhong, Guanxun Li, Jianhua Z. Huang, and James J. Cai. 'scTenifoldNet: a machine learning workflow for constructing and comparing transcriptome-wide gene regulatory networks from single-cell data.' *Patterns* 1, no. 9 (2020): 100139.** Available at: https://www.sciencedirect.com/science/article/pii/S2666389920301872. Licensed under a Creative Commons Attribution 4.0 International License https://creativecommons.org/licenses/by/4.0/, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

hances our understanding of cell development, homeostasis, and disease. Current scRNA-seq systems (e.g., 10× Genomics) can profile transcriptomes for thousands of cells per experiment. The sheer number of measured cells can be leveraged to construct GRNs. Advanced computational methods can facilitate such an effort to reach unprecedented resolution and accuracy, revealing the network state of given cells [168, 169, 170]. Furthermore, comparative analyses among GRNs of different samples will be extremely powerful in revealing fundamental changes in regulatory networks and unraveling the transcriptional programs that govern the behaviors of cells. Since our ability to generate scRNA-seq data has outpaced our ability to extract information from it, there is a clear need to develop effective computational algorithms and novel statistical methods for analyzing and exploiting information embedded within GRNs [171].

Constructing single-cell GRNs (scGRNs) using data from scRNA-seq and then effectively comparing constructed scGRNs presents significant analytical challenges [171, 22]. A meaningful comparison of scGRNs first requires a robust construction of a GRN from scRNA-seq data. Comparing scGRNs built via an unstable solution would cause misleading results and inappropriate conclusions. The vast number of different cellular states in a sample and the technical and biological noise, as well as the sparsity of scRNA-seq data, complicate the process of scGRN construction. Often, the expression of a gene is governed by stochastic processes and also influenced by transcriptional activities of many other genes. Thus, it is difficult to tease out subtle signals and infer true connections between genes. Furthermore, a direct comparison between two scGRNs is difficult; e.g., comparing each edge of the graph between scGRNs would be ill powered when scGRNs involve thousands of genes. Taken together, the key challenge in conducting comparative scGRN analysis is to extract meaningful information from noisy and sparse scRNA-seq data, since the information is deeply embedded in the differences between highly complex scGRNs of two samples.

In this paper, we introduce a workflow for constructing and comparing scGRNs using data from scRNA-seq of different samples. The workflow, which we call scTenifoldNet, is built upon several machine learning algorithms, including principal-component (PC) regression, low-rank tensor approximation, and manifold alignment. Through several examples, we show that scTenifoldNet is a sensitive tool to detect specific changes in gene expression signatures and the regulatory network rewiring events. The input of scTenifoldNet is a pair of expression matrices from scRNA-seq of two different samples. For instance, one sample may come from a healthy donor and the other from a diseased donor. In scTenifoldNet, the two input expression matrices are simultaneously processed through a multistep procedure. The final output is a list of ranked genes, sorted according to the differential regulation level of each gene. The ranked gene list can be used to perform functional enrichment analysis to detect the enriched molecular functions and involved biological processes. The constructed scGRN can also be used to identify functionally significant modules, i.e., subsets of tightly regulated genes.

scTenifoldNet includes an innovative method for comparing two scGRNs. We are not aware of any prior work using a similar design to achieve the same analytical goal. scTenifoldNet overcomes several methodological challenges, resulting in an effective and efficient scGRN comparison method. Here, we first benchmark and demonstrate the utility of scTenifoldNet across synthetic datasets and then apply scTenifoldNet to real datasets. Our real data analyses showed scTenifoldNet's power in identifying significant genes and network modules whose regulatory patterns shift greatly between samples. Some of these findings have not been reported in the respective original studies in which the datasets were generated.

5.2 Material and methods

5.2.1 The scTenifoldNet workflow

The scTenifoldNet workflow takes two scRNA-seq expression matrices as inputs. The two matrices are supposed to be obtained from two samples of the same type of cell, such as those of different treatments or from diseased and healthy subjects. The purpose of the analysis is to identify genes whose transcriptional regulation is shifted between the two samples. The whole workflow consists of five steps: cell subsampling, network construction, network denoising, manifold alignment, and module detection.

5.2.2 Cell subsampling

Instead of using all cells of each sample to construct a single GRN, we randomly subsample cells multiple times to obtain a set of subsampled cell populations. This subsampling strategy is to ensure the robustness of results against cell heterogeneity in samples. Subsampling of each sample is performed as follows: assuming the sample has M cells, m cells ($m < M$) are randomly selected to form a subsampled cell population. The process is repeated with cell replacement t times to produce a set of t subsampled cell populations.

5.2.3 Network construction

For a given expression matrix, a PC-regression network construction method [167] is adopted to construct scGRN. PC regression is a popular multiple regression method, where the original explanatory variables are first subjected to a PC analysis (PCA) and then the response variable is regressed on the few leading PCs. By regressing on PCs ($M \ll n$, where n is the total number of genes in the expression matrix), PC regression mitigates the overfitting and reduces the computation time. To build an scGRN, each time we focus on one gene (referred to as the target gene) and apply the PC-regression method, treating the expression level of the target gene as the response variable and the expression levels of other genes as the explanatory variables. The regression coefficients from PC regression are then used to measure the strength of the association of the target gene and other genes and to construct the scGRN. We repeat this process n times, each time with one gene as the target gene. At the end, the interaction strengths between all possible gene pairs are obtained and an adjacency matrix is formed. The details of applying the PC-regression method to a scRNA-seq expression data matrix are described as follows.

More specifically, suppose $X \in \mathbb{R}^{n \times p}$ is the gene expression matrix with n genes and p cells. The i^{th} row of X, denoted by $X_i \in \mathbb{R}^p$ represents the gene expression level of the i^{th} gene in the p cells. We construct a data matrix $X_{-i} \in \mathbb{R}^{(n-1) \times p}$ by deleting X_i from X. To estimate the effects of the other $n-1$ genes to the i^{th} gene, we build a PC-regression model for X_i. First, we apply PCA to X_{-i}^T and take the first M leading principal components to construct $Z^i = (Z_1^i, \cdots, Z_M^i) \in \mathbb{R}^{p \times M}$, where $Z_m^i \in \mathbb{R}^p$ is the m^{th} principal component of X_{-i}^T, $m = 1,\ 2,\ \ldots,\ M$. Mathematically, $Z^i = X_{-i}^T V^i$, where $V^i \in \mathbb{R}^{(n-1)\times M}$ is the PC loading matrix for the first M leading principal components, satisfying $\left(V^i\right)^T V^i = I_M$. Secondly, the PC-regression method regresses X_i on Z^i and solves the following optimization problem: $\hat{\beta}^i = \arg min_{\beta^i \in R^M} \parallel X_i - Z^i \beta^i \parallel_2^2$. Then, $\hat{\alpha}^i = V^i \hat{\beta}^i \in \mathbb{R}^{n-1}$ quantifies the effects of the other $n-1$ genes to the i^{th} gene. After performing PC-regression on each gene, we collect $\{\hat{\alpha}^i\}_{i=1}^n$ together and construct an $n \times n$ weighted adjacency matrix W of the gene-gene interaction network. The i^{th} row of W is $\hat{\alpha}^i$, and the diagonal entries of W are all 0. Then we retain interactions with top $\alpha\%$ ($= 5\%$ by default) absolute value in the matrix to obtain the scGRN adjacency matrix.

5.2.4 Tensor decomposition

For each of the t subsamples of cells obtained in the cell subsampling step, we construct a network using PC-regression, as described above. Each network is represented as a $n \times n$ adjacency matrix; the adjacency matrices of the t networks can be stacked to form a third-order tensor $\mathfrak{T} \in \mathbb{R}^{(n \times n \times t)}$. To remove the noise in the adjacency matrices and extract important latent factors, the CANDECOMP/PARAFAC (CP) tensor decomposition is applied. Similar to the truncated singular value decomposition (SVD) of a matrix, the CP decomposition approximates the tensor by a summation of multiple rank-one tensors [172]. More specifically, for our problem:

$$\mathfrak{T} \approx \mathfrak{T}_d = \sum_{r=1}^{d} \lambda_r a_r \circ b_r \circ c_r,$$

where $\circ$ denotes the outer product, $a_r \in \mathbb{R}^n, b_r \in \mathbb{R}^n$, and $c_r \in \mathbb{R}^t$ are unit-norm vectors, and λ_r is a scalar. In the CP decomposition, $\mathfrak{T}_d$ is the denoised tensor of $\mathfrak{T}$, which assumes that the valid information of $\mathfrak{T}$ can be described by d rank-one tensors, and the rest part $\mathfrak{T} - \mathfrak{T}_d$ is mostly noise.

We use the function `cp` in the R package 'rTensor' to do the CP decomposition. For each sample, the reconstructed tensor $\mathfrak{T}_d$ includes t denoised scGRNs. We then calculate the average of associated t denoised networks to obtain the overall stable network. We further normalize entries by dividing them by their maximum absolute value to obtain the final scGRNs for the given sample.

For later use, denote the denoised adjacency matrices for the two samples as W_d^x and W_d^y.

5.2.5 Manifold alignment

After obtaining W_d^x and W_d^y, we compare them to identify the regulatory changes and associated genes and modules. Instead of directly comparing the two $n \times n$ adjacency matrices, we apply manifold alignment to build comparable low-dimensional features and compare these features of genes between two samples, while maintaining the structural information of the two scGRNs [173]. Manifold alignment is used here to match the local and no-linear structures among the data points of W_d^x and W_d^y and project them to the same low-dimensional space. Specifically, we use W_d^x and W_d^y to denote the pairwise similarity matrices obtained by applying the PC-regression-based network construction method, and then denoising through tensor decomposition on the two initial expression matrices, X and Y. These similarity matrices serve as the input for manifold alignment to find the low-dimensional projections $F^x \in \mathbb{R}^{n \times d}$ and $F^y \in \mathbb{R}^{n \times d}$ of genes from each sample, where $d \ll n$. In terms of the underlying matrix representation, we use $F_i^x \in \mathbb{R}^d$ and $F_i^y \in \mathbb{R}^d$ to denote the i^{th} row of F^x and F^y that reflect the features of the i^{th} gene in X and Y, respectively.

We note that W_d^x and W_d^y may include negative values, which means genes are negatively correlated. When the similarity matrix contains negative edge weights, the properties of the corresponding Laplacian are not entirely well understood [174]. To deal with this problem, we add 1 to all entries in W_d^x and W_d^y, transforming the range of W_d^x and W_d^y from $[-1, 1]$ to $[0, 2]$. As a result, all original negative relationships have a transformed value in $[0, 1)$ and all original positive relationships have a transformed value in $(1, 2]$. In this case, the projected features of two genes with a positive correlation will be closer than those with a negative correlation. For convenience, we still use W_d^x and W_d^y to denote the transformed similarity matrices of two data sets.

Now we propose a specific manifold alignment method to find appropriate low-dimensional projections of each gene. Our manifold alignment should trade off the following two requirements: (1) the projections of the same i^{th} gene in two samples should be relatively close in the projected space; and (2) if i^{th} gene and j^{th} gene in sample 1 are functionally related, their projections F_i^x and F_j^x should be close in the projected space, and the same is true for sample 2. We minimize the following loss function: $Loss\left(F^x, F^y\right) = \lambda \sum_{i=1}^{n} \parallel F_i^x - F_i^y \parallel_2^2 + \sum_{i,j=1}^{n} \parallel F_i^x - F_j^x \parallel_2^2 W_{i,j}^x + \sum_{i,j=1}^{n} \parallel F_i^y - F_j^y \parallel_2^2 W_{i,j}^y$, where $W_{i,j}^x$ and $W_{i,j}^y$ denote the (i, j) entry of W_d^x and W_d^y respectively. The first term of the loss function requires the similarity between corresponding genes across two samples; the second and third terms are regularizers preserving the local similarity of genes in each of the two networks. λ is an allocation parameter to balance the effects of two requirements.

One way to minimize the loss function is by using an algorithm similar to Laplacian eigenmaps [175], which requires the adjacency matrix to be symmetry, but in our case both W_d^x and W_d^y are asymmetric. Notice that if we symmetrize W_d^x and W_d^y by $W^x = \frac{1}{2}((W_d^x)^T + W_d^x)$ and $W^y = \frac{1}{2}((W_d^y)^T + W_d^y)$, and again denote $W_{i,j}^x$ and $W_{i,j}^y$ as the (i, j) entry of W^x and W^y, then the value of the loss function won't be changed. Thus, minimizing the loss function based on the symmetrized adjacency matrices, W^x and W^y, is equivalent to using the original adjacency matrices, W_d^x and W_d^y. Based on this observation, using linear algebra, we can write the loss function into the matrix form as $Loss\left(F^x, F^y\right) = 2trace\left(F^T L F\right)$, where: $F = \begin{bmatrix} F^x \\ F^y \end{bmatrix}$, $L = \frac{1}{2}\left(D - W\right), W = \begin{bmatrix} W^x & \frac{\lambda}{2} I \\ \frac{\lambda}{2} I & W^y \end{bmatrix}$, and D is a diagonal matrix with $D_{ii} = \sum_i W_{ij}$. L is called a graph Laplacian matrix. The default selection of λ is 0.9 times the mean value of the row sums of W^x and W^y. By further adding the constraint $F^T F = I$ to remove the arbitrary scaling factor, minimizing $Loss\left(F^x, F^y\right)$ is equivalent to solving an eigenvalue problem. The solution for $F = [f_1, f_2, \cdots, f_d]$ is given by d eigenvectors corresponding to the d smallest nonzero eigenvalues of L [176].

5.2.6 Determination of p-value of deferentially regulated genes

With $F = \begin{bmatrix} F^x \\ F^y \end{bmatrix} = [f_1, f_2, \cdots, f_d]$ obtained in manifold alignment, we calculate the distance d_j between projected data points of two samples for each gene. One may declare significant genes according to the ranking of d_j's. To avoid arbitrariness in deciding the number of selected genes, we propose to use the Chi-square distribution to determine the significance of genes [120]. Specifically, d_j^2 is derived from the summation of squares of differences of projected representations of gene j for two samples, whose distribution could be approximately Chi-square. To adjust the scale of the distribution, we compute the scaled fold-change defined as $\frac{df \cdot d_j^2}{\bar{d}^2}$ for each gene j, where $\bar{d}^2$ denotes the average of d_j^2 among all the tested genes. The scaled fold-change approximately follows the Chi-square distribution with the degree of freedom df if the gene does not perform differently in the two samples. By using the upper tail $(P\,[X > x])$ of the Chi-square distribution, we assign p-values for genes and adjust them for multiple testing using the Benjamini–Hochberg (B–H) FDR correction [122]. To determine df, since the number of the significant genes will increase as df increases, we use $df = 1$ to make a conservative selection of genes with high precision.

5.2.7 Functional enrichment analyses

Functional enrichment analysis of gene sets was performed using Enrichr [123, 124], which is a web-based, integrative enrichment analysis application based on more than 100 curated gene set libraries. The test of enriched TF targets was performed using the ChIP-X enrichment analysis (ChEA) [177] based on comprehensive results from ChIP-seq studies. Finally, predefined gene sets from the REACTOME, BioPlanet and KEGG databases were tested for the enriched functions using the pre-ranked Gene Set Enrichment Analysis (GSEA) [178].

5.2.8 Simulations of scRNAseq data and benchmarking of network methods

A systematic evaluation of state-of-the-art algorithms for inferring scGRNs was performed using BEELINE [22]. We applied scTenifoldNet/PC-regression and other scGRN inference algorithms to a data set called GSD, which is derived from a curated Boolean model [179]. These methods include PIDC [180], PPCOR [181], LEAP [182], GRNBOOST2 [183], GENIE3 [164], SCINGE [184], SINCERITIES [185], GRISLI [186], SCODE [187], GRNVBEM [188], and SCNS [189]. Due to compatibility issues, SCRIBE [190] was not included in comparison. We processed the data set through the uniform pipeline provided by BEELINE: Preprocessing, generation of a Docker containers for scTenifoldNet/PC-regression and the 11 above-mentioned algorithms for scGRN construction, parameter estimation, and postprocessing and evaluation. We compared algorithms based on their average performance among three different metrics: precision-recall curve (AUPRC), receiver operating characteristic curve (AUROC), and the time of computing.

We generated our own synthetic data sets using SERGIO – a single-cell expression simulator guided by GRNs [191]. SERGIO allows for the simulation of scRNAseq data while considering the linear and nonlinear influences of regulatory interactions between genes. SERGIO takes a user-provided GRN to define the interactions and generates expression profiles of genes in steady-state using systems of stochastic differential equations derived from the chemical Langevin equation. The time-course of mRNA concentration of gene i is modeled by: $\frac{\partial X_i}{\partial t} = P_i\left(t\right) - \lambda_i x_i\left(t\right) + q_i\left(\sqrt{P_i\left(t\right)\alpha}\right)$, where x_i is the expression of gene i, P_i is its production rate, which reflects the influence of its regulators as identified by the given GRN, λ_i is the decay rate, q_i is the noise amplitude in the transcription of gene i, and α is an independent Gaussian white noise process. In order to obtain the mRNA concentrations as a function of time, the above stochastic differential equation is integrated for all genes as follows: $\left(X_i\right)_t = \left(X_i\right)_{t_0} + \int_{t_o}^{t}\left(P_i\left(t\right) - \lambda_i x_i\left(t\right)\right)\partial t + \int_{t_0}^{t} q_i\left(\sqrt{P_i\left(t\right)}\right)\partial W_\alpha$. The simulation was focused on testing and comparing the performance of PC-regression and several other methods (SCC, MI, GENIE3) using sparse data without imputation. The relationships between 100 genes were simulated as they belong to two major modules containing 40 and 60 genes, respectively. Each module is under the influence of one TF. We used the steady-state simulations to synthesize data to generate expression profiles of 100 genes, according to the parameter setting for two modules.

For each one of the tested methods, we randomly selected $n = 10$, 50, 100, 500, 1000, 2000, and 3000 cells from the simulated data for ten times and build ten scGRN. For each n, relevance measurements (accuracy and recall) were evaluated for each of the ten networks using the match of the sign of the relationships between genes to compute the following formulas: Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ and Recall = $\frac{TP}{TP+FN}$, where TP stands for true positive, TN true negative, FP false positive, and FN false negative. For the MI and GENIE3 methods that only provide positive values, the median value was used as the center point and then the values were scaled to $[-1, 1]$ by dividing them over the maximum absolute value.

5.2.9 Code availability

scTenifoldNet has been implemented in R. The source code is available at https://github.com/cailab-tamu/scTenifoldNet, which also includes the code of the benchmarking method, auxiliary functions, and example datasets (including the simulated data used to generate Fig. 5.2). The scTenifoldNet R package is available at the CRAN repository. The following is available online: Document S1 including Figures S1–S5 and Tables S1–S7 https://ars.els-cdn.com/content/image/1-s2.0-S2666389920301872-mmc1.pdf. Document S2 including the article plus Supplemental Information https://ars.els-cdn.com/content/image/1-s2.0-S2666389920301872-mmc2.pdf

5.3 Results

To enable comparative scGRN analysis in a robust and scalable manner, we base our method on a series of machine learning methods. A key challenge of our comparative analysis is to extract meaningful differences in regulatory relationships between two samples from noisy and sparse data. Specifically, we seek to contrast scGRNs constructed from different scRNAseq expression matrices. Fig. 5.1 shows the main components of scTenifoldNet architecture. The whole workflow contains five key steps: subsampling cells, constructing multilayer scGRNs, denoising, manifold alignment, and differential regulation (DR) test. In order to produce biologically meaningful results, we made dedicated design decisions for the task in each of these steps. Next, we briefly describe the numerical methods implemented in scTenifoldNet. More technical details are presented in section 5.2.

5.3.1 Numerical Methods

The numerical methods used to construct and compare scGRNs involve the following five steps:

1. Pre-processing data and subsampling cells: The input data are two scRNAseq expression data matrices, X and Y, containing expression values for n genes in m_1 and m_2 cells from two different samples, respectively. Next, m cells in X and Y are randomly sampled to form X' and Y'. This subsampling process is repeated t times to create two collections of

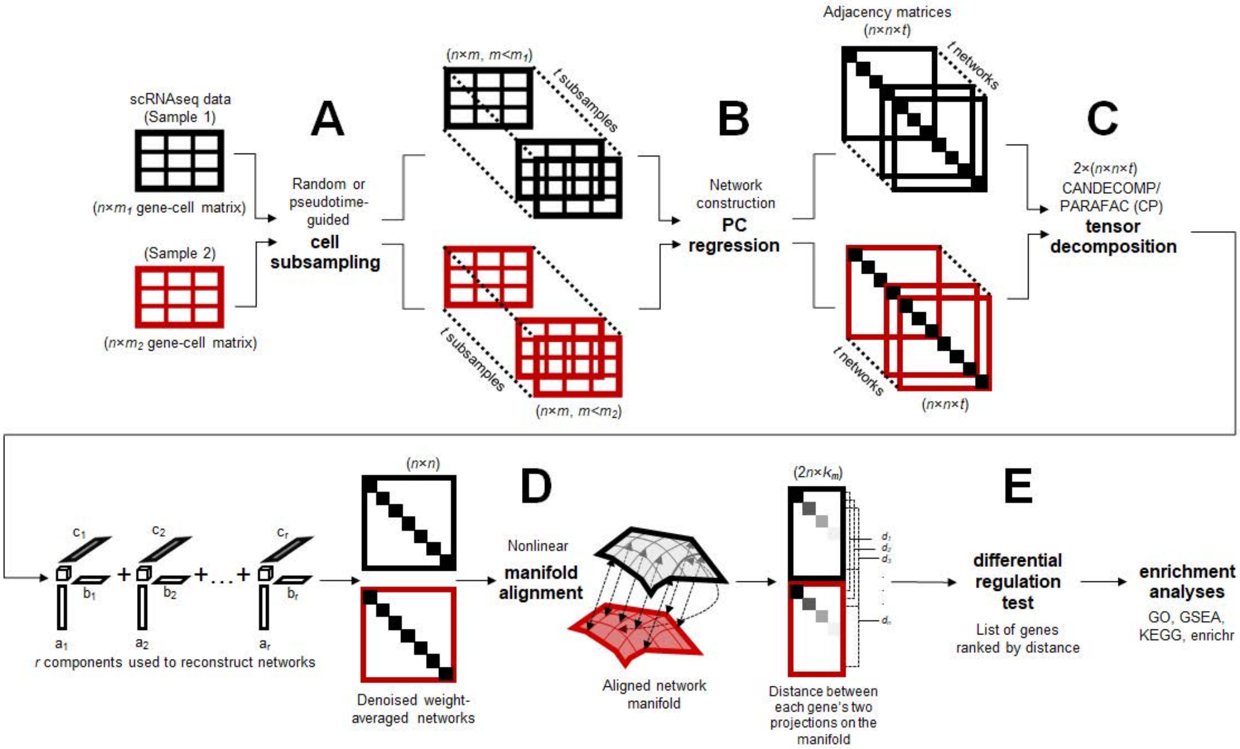


Figure 5.1: Overview of the scTenifoldNet workflow. scTenifoldNet is a machine learning framework that uses a comparative network approach with scRNA-seq data to identify regulatory changes between samples. scTenifoldNet is composed of five major steps. (A) Cell subsampling. scTenifoldNet starts with subsampling cells in the scRNA-seq expression matrices. When two samples are analyzed, each of the two samples is subsampled either randomly or following a pseudotime trajectory of cells. The subsampling is repeated multiple times to create a series of subsampled cell populations, which are subject to network construction and form a multilayer scGRN. (B) Network construction. PC regression is used for scGRN construction; each scGRN is represented as a weighted adjacency matrix. (C) Tensor denoising. Two samples produce two multilayer GRNs and form two three-order tensors, which are subsequently decomposed into multiple components. The top components of tensor decomposition are then used to reconstruct two denoised multilayer scGRNs. Then, two denoised multilayer scGRNs are collapsed by taking the average weight across layers. (D) Manifold alignment. The two single-layer average scGRNs are then aligned with respect to common genes using a nonlinear manifold alignment algorithm. Each gene is projected to a low-rank manifold space as two data points, one from each sample. (E) Differential regulation test. The distance between the two data points is the relative difference of the gene in its regulatory relationships in the two scGRNs. Ranked genes are subject to tests for their significance in differential regulation between scGRNs.

subsampled cells $\{X'_i\}$ and $\{Y'_i\}$, where $i = 1, 2, \ldots, t$.

2. Constructing initial networks: For each $X'_i \in \{X'_i\}$, $i = 1,\ 2, \ldots,\ t$, PC-regression is used to construct a GRN. The constructed GRN from X'_i is stored as a weighted graph represented with an $n \times n$ weighted adjacency matrix W^x_i. Similarly, for each $Y'_i \in \{Y'_i\}$, $i = 1, 2, \ldots, t$, we construct a GRN and represent it with an $n \times n$ weighted adjacency matrix W^y_i. Diagonal values of each adjacency matrix are set to zeros, and other values are normalized by dividing by their maximal absolute value. Each normalized adjacency matrix is then filtered by retaining only the top 5% of edges ranked using the absolute edge weight, resulting in a sparse adjacency matrix.

3. Denoising: Tensor decomposition [172] is used to denoise the adjacency matrices obtained in Step 2. The collection of t scGRNs for each sample, $\{W^x_i\}$ or $\{W^y_i\}$, is processed separately as a third-order tensor, denoted as $\mathfrak{T}^x$ or $\mathfrak{T}^y$, each containing $n \times n \times t$ elements. The CANDECOMP/PARAFAC (CP) decomposition is applied to decompose $\mathfrak{T}^x$ and $\mathfrak{T}^y$ into components. Next, $\mathfrak{T}^x$ and $\mathfrak{T}^y$ are reconstructed using top r components to obtain denoised tensors: $\mathfrak{T}^x_d$ and $\mathfrak{T}^y_d$. Denoised $\{W^x_i\}$ and $\{W^y_i\}$ in $\mathfrak{T}^x_d$ and $\mathfrak{T}^y_d$ are collapsed by taking the average of edge weights for each edge to form two denoised, averaged matrices, W^x_d and W^y_d, which are subsequently normalized as in step 2 and then symmetrized.

4. Aligning genes onto a manifold: The weighted adjacency matrices W^x_d and W^y_d are regarded as two similarity matrices for a nonlinear manifold alignment procedure. The alignment is done by solving an eigenvalue problem with a Laplacian matrix derived from the joint matrices: $W = \left[W^x_d,\ \lambda I/2; \lambda I^T/2,\ W^y_d\right]$, where λ is a tuning parameter and I is the identity matrix that reflects the binary correspondence between genes in the samples, X and Y. As the result of manifold alignment, all genes in the samples, X and Y, are projected on a shared, low dimensional manifold with a dimension $k_m \ll n$. The projections of each gene j from the samples, X and Y, are two k_m-dimensional vectors, F^x_j and F^y_j.

5. Ranking genes: For each gene j, let d_j be the Euclidean distance between the gene's two

projections F_j^x and F_j^y on the shared manifold: one is from the sample X, and the other is from the sample Y. Genes are sorted according to this distance. The greater the distance, the greater the regulatory shift.

In the following sections, we explain rationale behind each step of scTenifoldNet, as well as the selection of machine learning algorithms and implementation details.

5.3.2 Subsampling of cells

The rationale for randomly subsampling cells is close to that of ensemble learning. Ensemble learning is a technique where multi-model decisions are merged to improve overall performance. Similarly, instead of attempting to build a single scGRN, scTenifoldNet randomly samples subsets of cells from the given scRNAseq expression matrix and builds a series of 'low-precision' scGRNs with the subsampled data sets. These 'low-precision' scGRNs are then combined to obtain a 'high-precision' scGRN. As mentioned above, current scRNAseq technology can produce the transcriptome profiles of thousands of cells from each sample. It is fundamentally difficult to process high-dimensionality and large-scale scRNAseq data, especially given that there can be a substantial variation among cells. This happens even in a group of highly homogenous cells of the same type [192]. The presence of so-called outlying cells, i.e., cells showing profiles of expression deviate from those of most other cells, many influence the construction of 'high-precision' scGRNs. Therefore, subsampling offers promise as a technique for handling the noise in the input data sets. When the number of cells is small, the input data matrix may be resampled with replacement [193].

5.3.3 Constructing scGRNs using PC-regression

Although many GRN construction methods have been developed [163, 164, 166], it is unclear which one is suitable for constructing a large number of scGRNs from the subsampled data [171]. When dealing with multiple sets of input data, both the accuracy and computational efficiency of these algorithms have to be considered. We opted to use the PCNet method [167], which is based on PC-regression [194], after conducting a thorough review of the current methods. The

PC-regression method extracts the first few (e.g., $k = 3$) PCs and then uses these components as the predictors in a linear regression model fitted using ordinary least squares. The values of the transformed coefficients of genes are treated as the strength and regulatory effect between genes to generate the network. The main use of PC-regression in scTenifoldNet lies in its ability to surpass the multicollinearity problem that arises when two or more explanatory variables are linearly correlated.

5.3.4 Denoising via low-rank tensor approximation

Removing the noise from constructed scGRNs is an important step of scTenifoldNet. Here the term noise is used in a broad sense to refer to any outlier or interference that is not the quantity of interest, i.e., the true regulatory relationship between genes. For each sample, the multilayer scGRN constructed from multiple subsampled data sets is regarded as a rank-three tensor. To reduce the noise in the multilayer scGRN, we decompose the tensor and reconstruct the multilayer scGRN using leading components. The idea is similar to that of denoising using truncated singular value decomposition (SVD). After cutting a larger portion of the noise spread over the lowest singular value components, the reconstructed data matrix based on the truncated SVD would, therefore, represent the original data with reduced noise. Indeed, tensor decomposition has been used in video data analyses for denoising and information extracting purposes [195]. It has also been used to impute missing data [196]. We use the CANDECOMP/PARAFAC (CP) algorithm [197] to factorize the two multilayer scGRNs separately and regenerate all adjacency matrices using leading components. The number of components used for reconstruction can be specified and is set to 3 by default. In the real data applications, we find the tensor GRN regeneration serves for two purposes: denoising and enhancing, i.e., making main signals stronger and making less important signals weaker.

5.3.5 Manifold alignment of two scGRNs

For a gene, its position in one of the two scGRNs (i.e., denoised adjacency matrices from the two samples) is determined by its regulatory relationships with all other genes. Here we regard each gene as a data point in a high-dimensional space where components of the data point are the features, i.e., weights between the gene and all other genes in the scGRN adjacency matrix. To compare the same gene's positions in the two scGRNs, we first align the two scGRNs. To do so, we take a popular and effective approach for processing high-dimensional data, intuitively modeling the intrinsic geometry of the data as being sampled from a low-dimensional manifold—i.e., commonly referred to as the manifold assumption [198]. This assumption essentially means that local regions in the data can be mapped to low-dimensional coordinates, while the nonlinearity and high dimensionality in the data come from the curvature of the manifold. Manifold alignment produces projections between sets of data, given that the original data sets lie on a common manifold [173, 199, 200, 201]. Manifold alignment matches the local and nonlinear structures among the data points from multiple sources and projects them to the same low-dimensional space while maintaining their local manifold structure of each source. The ability to flexibly learn and accurately represent the structure in the data with manifold alignment has been demonstrated in applications in automatic machine translation, face recognition, and so on [202, 203]. Here, we use manifold alignment to match genes in the two denoised scGRNs, one from each sample, to identify cross-network linkages. Consequently, the information of genes stored in two scGRNs is aligned, meaning points close together in the low-dimensional space are more similar than points that are farther apart.

5.3.6 Ranking genes and reporting DR genes

To identify genes whose regulatory status differs between the two samples, we calculate the distance between projected data points in the manifold alignment subspace. For each gene, if the gene appears in scGRNs of both samples, there are two data points for the same genes, one from each sample. We compute the Euclidean distance between the two data points of the gene and used the distance to measure the dissimilarity in the gene's regulatory status in two scGRNs [204]. We do this for all genes shared between two samples and then rank genes by the distance. The larger the distance, the more different the gene in two samples. In this way, we obtain a list of ranked genes. These ranked genes are subject to functional annotation, such as using the pre-ranked Gene

Set Enrichment Analysis (GSEA) [178] to assess the enriched functions associated with top genes. To avoid choosing the number of selected genes arbitrarily, we compute p-values for genes using Chi-square tests, adjust p-values with a multiple testing correction, and select significant genes using 5% FDR cutoff.

5.3.7 Benchmarking the performance of scTenifoldNet using simulated data

5.3.7.1 Precision and recall of the network construction method adopted in scTenifoldNet

PC-regression is the method we adopted for scTenifoldNet to construct scGRN. It is important to ensure scTenifoldNet/PC-regression is an effective and efficient network construction method for our purpose. To this end, we conducted a systematic comparison between network construction algorithms using a published evaluation tool package called BEELINE [22]. We benchmarked scTenifoldNet/PC-regression and compared it with 11 other algorithms (see 5.2). We chose to re-use a reference data set called GSD in the BEELINE package to perform the benchmarking. GSD is the largest curated reference data set provided in the BEELINE package. We found that, when jointly comparing their precision-recall curve (AUPRC), receiver operating characteristic curve (AUROC) and computation time, scTenifoldNet/PC-regression, along with a partial information decomposition-based method [180], outperformed all other algorithms (see Supplementary Fig. S1 for details).

We also simulated scRNAseq data using a parametric method with a predefined scGRN model (see 5.2 for details). With the simulated data, we compared constructed scGRNs against the ground truth (i.e., the simulated scGRN) to estimate the accuracy of reconstruction. We tested the accuracy of scTenifoldNet/PC-regression against methods based on Spearman correlation coefficient (SCC) and mutual information (MI) [163], and GENIE3 [164]. SCC and MI methods are computationally efficient, whereas GENIE3 is not, but GENIE3 is the top-performing method for network inference in the DREAM challenges [165]. For each method, their performances in recovering gene regulatory relationships were compared against the ground-truth interactions between genes, where were generated according to pre-setting parameters. We found that scTenifoldNet/PC-regression produced more specific (better accuracy) and more sensitive (better recall) scGRNs than other methods (Fig. 5.2 A.). This is true across a wide range of settings of cell numbers in input scRNAseq expression matrices. scTenifoldNet/PC-regression is also much faster than GENIE3 (running time information is available in Supplementary Table S1).

5.3.7.2 Effect of denoising with tensor decomposition

To show the effect of tensor denoising, we simulated scRNAseq data (see 5.2) and processed the data using the first two steps of scTenifoldNet, i.e., cell subsampling followed by the construction of scGRNs using PC-regression. We subsampled 500 cells each time and generated ten scGRNs. The ten scGRNs are treated as a multilayer network or a tensor to be denoised. For each scGRN, we kept the top 20% of the links. The presence and absence of links in each scGRN were compared with those in the simulated, ground-truth scGRN to estimate the accuracy of recovery and the rate of recall. Fig. 5.2 B. contains the heatmaps of adjacency matrices of the ten scGRNs before and after denoising (small panels). We also show two collapsed scGRNs (Fig. 5.2 B., large panels), which were generated by averaging link weights across the ten scGRNs before and after denoising. These results illustrate the ability of scTenifoldNet to denoise multilayer scGRNs. For instance, tensor denoising improves the recall rate of regulatory relationships between genes by 25%. This simulation study suggests that tensor denoising could be useful for removing impacts of random dropout and other noise issues affecting the scGRN construction using scRNAseq data.

5.3.7.3 Detecting power illustrated with a simulated data set

We used simulated data to show the capability of scTenifoldNet in detecting differentially regulated (DR) genes. We first used the negative binomial distribution to generate a sparse synthetic scRNAseq data set (an expression matrix including 67% zeros in its values). This toy data set includes $2,000$ cells and 100 genes. We called it sample 1. We then duplicated the expression matrix of sample 1 to make sample 2. We modified the expression matrix of sample 2 by swapping expression values of three randomly selected genes with those of another three randomly selected genes. Thus, the differences between samples 1 and 2 are restricted in these six genes. Using scTenifoldNet with the default parameter setting, we compared the originally generated expression matrix (sample 1) against itself (sample 1 vs. sample 1) and also against the manually perturbed version (sample 1 vs. sample 2). As expected, when comparing the original matrix against itself, none of the genes was identified to be significant. However, when samples 1 and 2 were compared, the six genes whose expression values were swapped were identified as significant DR genes (Fig. 5.2 C., FDR < 0.1). These results are expected and support the sensitivity of scTenifoldNet in identifying subtly shifted gene expression programs.

5.3.8 Real data analyses

5.3.8.1 Practical considerations of real data analysis using scTenifoldNet

First of all, we address several practical questions regarding the application of scTenifoldNet to real scRNAseq data. (1) What are the input expression matrices to be compared? The input to scTenifoldNet is two matrices of gene expression values (e.g., UMI counts) as measured in two samples to be compared. In each matrix, columns represent cells, and rows represent genes. We assume that each input matrix contains a sizable number of cells. For example, a typical input matrix may contain UMI counts for $5,000$ genes and $2,000$ cells. Whether a gene is expressed among cells can be determined by examining if this gene has a nonzero UMI count in more than 5% of cells. Scaling normalization (e.g., the library size normalization) of the input UMI count matrix does not seem to affect the construction of scGRNs (Supplementary Fig. S2). In contrast, imputing the UMI count matrix using an imputation algorithm (e.g., MAGIC [24]) may have impact on the performance of scTenifoldNet (Supplementary Fig. S3). (2) How does scTenifoldNet handle cell heterogeneity? Heterogeneity in expression among cells is inevitable. scTenifoldNet is designed to tolerate a certain level of such heterogeneity as long as cells are of the same type. scTenifoldNet is not a data preparation tool. It also does not perform any clustering analysis for cells; it does not assign cells into cell types. We assume all cells in both input matrices are of the same type. Otherwise, the results would be difficult to interpret. To solve this problem, a specific tool (to prioritize cell types most responsive to biological perturbations) has been developed

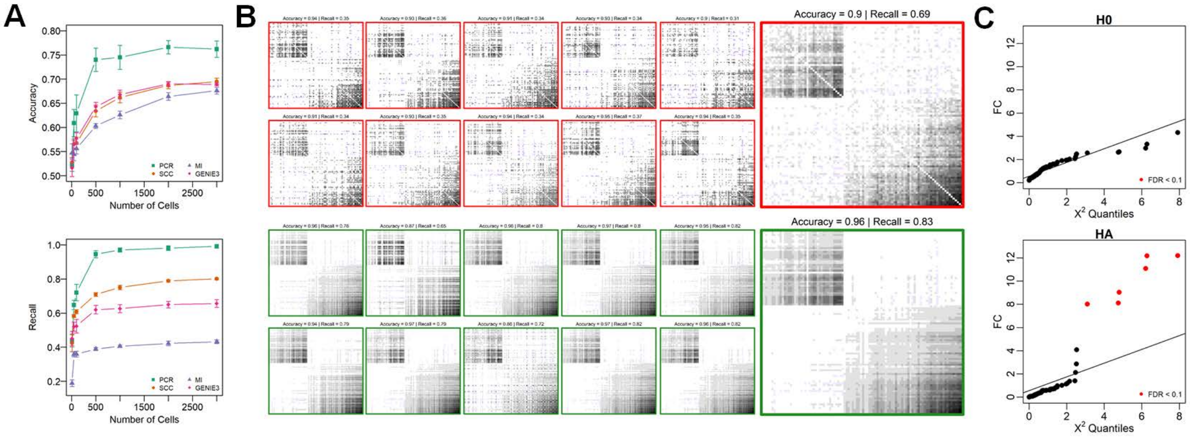


Figure 5.2: Benchmarking the performance of scTenifoldNet using simulated data. (A) The accuracy and recall of scGRN construction using different methods: PC-regression, SCC, MI, and GENIE3, as functions of the number of cells used in the analysis. Error bar is the standard deviation of the computed values after 10 bootstrapped evaluations. PCR – PC-regression; SCC – Spearman correlation coefficient; MI – mutual information; GENIE3 – a random forest-based network construction method. (B) Visualization of the effect of tensor denoising on accuracy and recall of multilayer scGRNs. Each subpanel is a heatmap of a 100×100 adjacency matrix constructed using PC-regression over the counts of 500 randomly subsampled cells. Grayscale indicates the relative strength of regulatory relationships between genes. Top part includes networks before tensor denoising (adjacency matrices in heatmap with red box); bottom part includes corresponding networks after tensor denoising (adjacency matrices in heatmap with green box). In each part, adjacency matrices of networks of 10 subsamples (10 small heatmaps) and their average adjacency matrix (one large heatmap) are shown. (C) Evaluation of the sensitivity of scTenifoldNet in identifying punctual changes in the regulatory profiles. Top panel: evaluation of the original data matrix against itself; bottom panel: evaluation of the original matrix against the perturbed matrix. Significant genes identified using the differential regulation test (FDR < 0.1, B–H correction) are indicated in red. All significant genes are perturbed in simulation and thus are expected to be identified.

elsewhere [205]. (3) What if the number of cells is too small? We expect that each input matrix contains a sizable number of cells (e.g., $n > 2,000$). If this is the case, the jackknife method (subsampling without replacement) is adapted by default: $m = 500$ cells are subsampled each time. Alternatively, an m-out-of-n bootstrap method (subsampling with replacement) can be used [193]. When the number of cells is small (e.g., $n = 500$), a full bootstrap method can be used, i.e., resampling 500 cells each time out of 500 given cells with replacement [193, 206]. scTenifoldNet is robust against imbalanced cell numbers in the two samples for comparison (Supplementary Fig. S4). (4) What is the relationship between scTenifoldNet analysis and DE analysis? scTenifoldNet analysis should be used as a complementary analysis method in addition to DE analysis, rather than replacing DE analysis. DE analysis (using e.g., MAST [89], edgeR [207] or SCDE [208]) is still a widely used method for understanding the difference between two scRNAseq samples [209]. scTenifoldNet is designed based on a principle different from that underlying DE analysis. Thus, the results of scTenifoldNet analysis and DE analysis are not supposed to be compared side by side. It is not uncommon that scTenifoldNet and DE analyses report the same genes to be significant. This is because the change of the regulatory pattern of a gene in scGRNs may be associated with the change of the gene's expression level. To evaluate the influence of gene expression level on scGRN construction, we calculated the correlation between the average gene expression level and the average weighted degree of nodes in scGRNs, which are constructed using scTenifoldNet/PC-regression and other algorithms in the BEELINE package [22]. If the weighted degree of nodes in a scGRN constructed using a method is correlated with the expression level of genes, then it indicates that the method is likely to be biased towards highly expressed genes during the process of scGRN construction. We found that all evaluated algorithms produced results showing a certain level of such a correlation (Supplementary Fig. S5). However, compared with all other algorithms, scTenifoldNet/PC-regression produced the smallest correlation value and thus is most robust against the bias towards highly expressed genes.

Study	**Reference**	**Species**	**Cell type**	**Perturbation type**	**Number of genes included in analysis**	**Number of cells in two groups**	**Number of DR genes**	**Enriched functions of the DR gene list**
1	[210]	Mouse	Neurons	Morphine	$8,136$	Mock-treated = $8,912$ Morphine-treated = $7,972$	56	Opioid signaling Signaling by G protein-coupled receptors Reduction of cytosolic calcium levels Morphine addiction
2	[211]	Human	Carcinoma cell line	Cetuximab	$11,140$	Untreated = $5,217$ Treated = $4,507$	125	EGFR1 pathway Regulation of apoptosis Cell cycle checkpoints G1 cell cycle arrest Regulation of apoptosis
3	[212]	Mouse	Lung alveolar cells	*Nkx2-1* gene knockout	$7,842$	Wild-type = 638 Knockout = $2,397$	29	Gastrointestinal marker genes *Sox2* target genes
4	[117]	Human	Dermal fibroblasts	dsRNA immune stimulus	$7,904$	Unstimulated = $2,553$ Stimulated = $2,130$	29	Interferon signaling Immune system Interleukin-1 regulation of extracellular matrix
5	[213]	Mouse	Neurons	Alzheimer's disease	$2,869$	Wild-type = $4,561$ Knockout = $2,423$	29	Functions of *Apoe* and *Bin1* Regulation of neuron projection development Positive regulation of cell projection organization Phosphatidylserine metabolic process Protein acylation Potassium channel activity Methylation-dependent protein binding Integrin signaling pathway Serotonin HTR1 group and FOS pathway Glutamate neurotransmitter release cycle

Table 5.1: Summary of real-data applications of scTenifoldNet analysis

5.3.8.2 Analysis of transcriptional responses of neurons to acute morphine treatment

To illustrate the use of scTenifoldNet, we first applied scTenifoldNet to a scRNAseq data set from [210]. This is a study on transcriptional responses of mouse neural cells to morphine (Fig. 5.3 A.). In the study [210], Avey and colleagues performed scRNAseq experiments with the nucleus accumbens (NAc) of mice after four hours of the morphine treatment, using mice treated with saline as mock controls. Single-cell expression data was obtained for $11,171$ and $12,105$ cells from four morphine- and four mock-treated mice, respectively [210]. The measured cells were clustered to identify neurons ($7,972$ and $8,912$ from morphine- and mock-treated samples, respectively); the identified neurons were then sub-grouped into 11 clusters, including major clusters of D1 and D2 medium spiny neurons (MSNs), comprising $\approx 95\%$ of the neurons in the NAc. Using differential expression (DE) analysis implemented in SCDE [208], Avey *et al.* identified several hundred genes that are differentially expressed between morphine- and mock-treated samples (Supplementary Table S2 of [210]). Although this result is intriguing, we argue that it seems that when so many genes are identified as 'significant players', it is difficult to interpret the result and to pinpoint the specific regulatory mechanism underlying the true response. Indeed, instead of performing functional enrichment analysis with identified DE genes, the subsequent analyses in the study of [210] were re-focused on a tiny portion of D1 MSNs, called activated MSNs. It is only when activated MSNs were compared to all other D1 MSNs that 256 DE genes were identified (SCDE, P < 0.001, Supplementary Table S2 of [210]). These genes were then found to be associated with several terms related to *opioid addiction*, including *morphine dependence* and *opioid-related disorders* (Supplementary Table S3 of [210]). In the morphine-treated sample, less than 4.5% of D1 MSNs are activated MSNs; in the mock-treated sample, less than 2% (see Fig. 2B of [210]). In view of these, we point out here that while relevant signals can be detected using traditional DE analysis, the analytical method involves extensive human intervention – i.e., an iterative clustering procedure is needed to identify a final population of cells (in this case, activated MSNs). The cell population size is small, making the analysis result potentially variable.

We were motivated by these considerations and set out to re-analyze the data. We first re-produce results of DE analysis. We found that the mock- and morphine-treated neurons indeed exhibited a striking similarity. For example, mock- and morphine-treated neurons are indistinguishable in a tSNE plot (Fig. 5.3 B.); expression levels of several known morphine responsive genes, e.g., *Adcy5*, *Ppp1r1b*, and *Ppp3ca*, show no difference (Fig. 5.3 C.). Thus, a direct comparison of gene expression between neurons using the DE method may have limited power to identify relevant genes involved in the morphine response.

Next, using scTenifoldNet, we identified 56 genes showing significant differences in their transcriptional regulation between mock- and morphine-treated neurons (Supplementary Table S2). Compared to other genes, these genes have significantly greater distance between their positions in two scGRNs aligned into the manifold (FDR < 0.05, Chi-square test with B-H multiple test adjustment, see 5.2 for details). GSEA analysis [178] showed that these DR genes are enriched for *opioid signaling*, *signaling by G protein-coupled receptors*, *reduction of cytosolic Calcium levels*, and *morphine addiction* (Fig. 5.3 D. inset). It is known that morphine binds to the opioid receptors on the neuronal membrane. The signal is then transmitted through the G-protein signaling system, inhibiting the adenylyl cyclase in the cytoplasm and decreasing the levels of cAMP and the calcium-channel conduction [214, 215, 216]. Furthermore, 21 out of 56 (38%) identified DR genes (Supplementary Table S2) were found to be targets of *RARB* (adjusted p-value < 0.01, Enrichr enrichment test [123, 124] based on ChIP-seq data [177]). *RARB* plays a role in synaptic transmission in dopaminergic neurons and the adenylate cyclase-activating dopamine receptor signaling pathway [217, 218]. Thus, these enriched functions are relevant to the morphine stimulus, which is known to induce the disinhibition of dopaminergic neurons by GABA transmission, enhance dopamine release, and cause addiction [219, 220]. Using the constructed scGRN, we were able to trace DR genes back to their topological positions in the network and examine their interacting genes. Fig. 5.3 E. shows such a network module, including multiple DR genes.

In this case, scTenifoldNet is used as an unsupervised tool, and no human interference is needed to operate. This feature is critical when referring to this specific set of data because where the signal is limited to rare types of cells, there is a chance that a less sensitive approach would miss the signal, especially when human interference is not provided. It is ideal to have an unsupervised tool that is sensitive to signals, and robust to variation between cells at the same time. We note that scTenifoldNet is a different tool to conventional DE analysis tool – scTenifoldNet reported less DR genes in terms of the number of genes, compared with DE genes identified in the original study [210]. Among the 56 DR genes that scTenifoldNet detected, 11 (*Actb*, *Adcy5*, *Akap9*, *D430041D05Rik*, *Eif1*, *Pcp4l1*, *Penk*, *Phactr1*, *Rasd2*, *Scn4b* and *Ubb*) are among the 256 DE genes reported in Supplementary Table S2 of [210]. The number of overlap genes is not significantly higher than expected by random according to a hypergeometric test (P $= 0.29$) with a total of $1,432$ genes (from Supplementary Table S2 of [210]) included in the test. Fig. 5.3 C. shows expression levels of three representative genes: *Pde1b*, *Adcy5* and *Gabrg1*, in neurons from mock- and morphine-treated mice. All three genes are known to be involved in morphine response [221, 222, 223], but only when DE and DR tests are applied jointly, all three genes are identified: *Pde1b* is a DE but not a DR gene, *Adcy5* a DR and DE gene, and *Gabrg1* a DR but not DE gene.

5.3.8.3 Analysis of transcriptional responses of a carcinoma cell line to cetuximab

To further illustrate the power of scTenifoldNet in identifying genes associated with specific perturbations, we applied scTenifoldNet to another published scRNAseq data [211]. In this study, Kagohara et al. [211] use scRNAseq to study mechanisms that lead the development of resistance to cetuximab in head and neck squamous cell carcinoma (HNSCC)(Fig. 5.4 A.). Cetuximab is a human-murine chimeric monoclonal antibody used for the treatment of metastatic colorectal cancer, metastatic non-small cell lung cancer, and head and neck cancer. In conjunction with the radiotherapy, cetuximab improves the objective response rate in first-line treatment of recurrent or metastatic squamous cell carcinoma of the head and neck [224]. Cetuximab binds to the extracellular domain of the epidermal growth factor receptor (*EGFR*) on both normal and tumor cells [225]. *EGFR* is over-expressed in many cancers. Competitive binding of cetuximab to *EGFR* blocks the phosphorylation and activation of receptor-associated kinases and their downstream targets, e.g., *MAPK*, *PI3K/Akt*, and *Jak/Stat* pathways [226], thereby reducing their effects on cell growth and metastatic spread. It is known that blocking *EGFR* activation also affects cellular processes such as

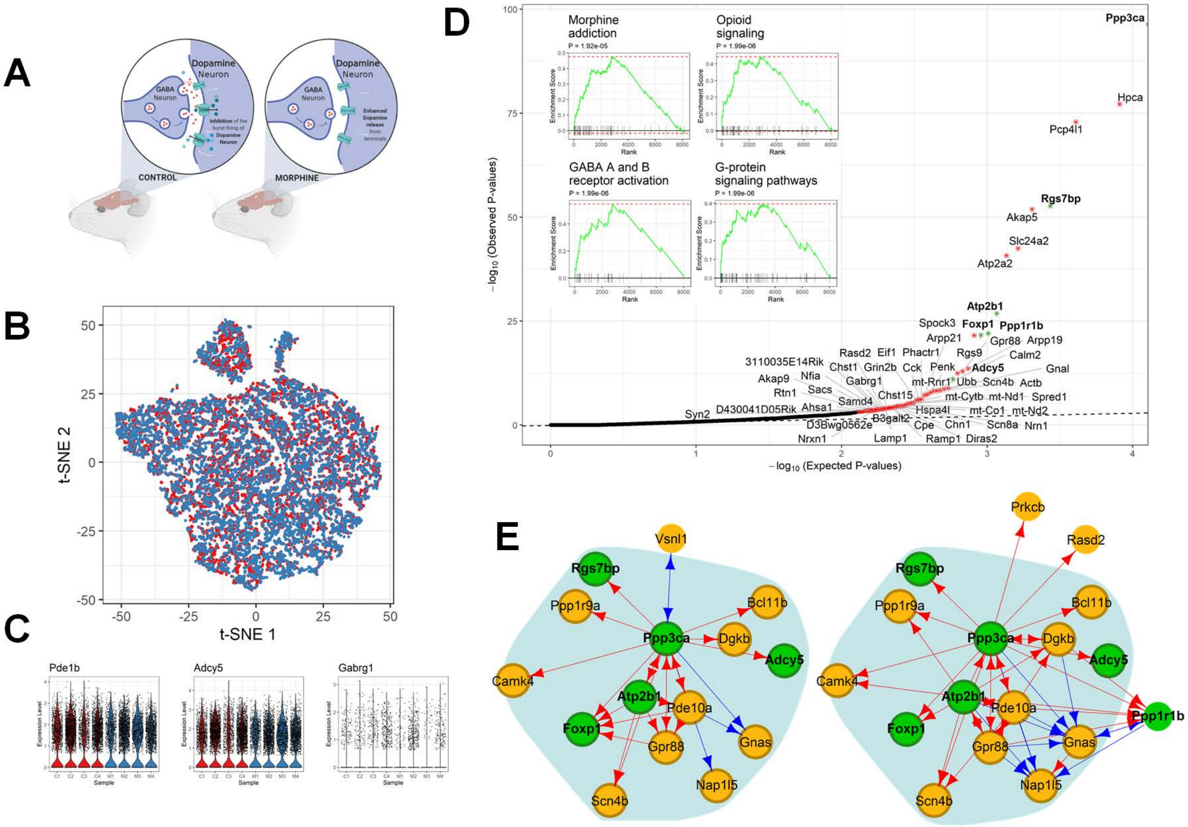


Figure 5.3: Analysis of transcriptional responses to morphine in mouse cortical neurons. (A) Illustration of experimental design and data collection of the morphine response study [210]. (B) A t-SNE plot of $7,972$ and $8,912$ neurons from morphine-treated (blue) and mock-treated (red) mice, respectively. (C) Violin plots show the log-normalized expression levels of representative DR and/or DE genes in four (M) morphine- and four (C) mock-treated mice. (D) Quantile-quantile (Q-Q) plot for observed and expected p-values of the $8,138$ genes tested. Genes ($n = 65$) with FDR < 0.1 are shown in red; genes ($n = 56$) with FDR < 0.05 are labeled with asterisk. Inset shows results of the GSEA analysis for genes ranked by their distances in manifold aligned scGRNs from morphine- and mock-treated mice. (E) The module enriched with DR genes and the corresponding subnetworks in two scGRNs. For illustrative purposes, the module is centered on the DR gene, *Ppp3ca*. Significant DR genes (FDR < 0.05) in the module are highlighted in green. Edges are color-coded: red indicates a positive association, and blue indicates negative. Weak edges are filtered out by thresholding for clear visualization, and the background shadow indicates the shared portion of the module in the two scGRNs.

apoptosis, *cell growth*, and *vascular endothelial growth factor (VEGF) production* [227]. Cetuximab is also known to cause degradation of the antibody-receptor complex and the downregulation of *EGFR1* expression [228].

Kagohara et al. [211] sequenced the transcriptome profile of cells before and after Cetuximab treatment for 120 hours in three different HNSCC cell lines: SCC1, SCC6, SCC25. They found that SCC6 is the most sensitive to the cetuximab treatment, reporting $8,389$ genes as differentially expressed (including $4,166$ upregulated and $4,223$ downregulated ones with P < 0.05; Supplementary Table S4 of [211]). Such a large number of differentially expressed genes makes it difficult to identify genes directly associated with the molecular mechanism through which cetuximab acts. We extracted scRNAseq data for $4,507$ and $5,217$ SCC6 cells treated with and without cetuximab, respectively (Fig. 5.4 B.). Expression levels of three genes: *DuSP4*, *TIGA3* and *LIF*, in cells of two treatment groups are shown in Fig. 5.4 C. All three genes are in the *EGFR* pathway. We used scTenifoldNet to re-analyze the data and identified 125 DR genes (FDR < 0.05, Fig. 5.4 D., Supplementary Table S3). These genes are enriched with those (39 out of 125) that are under the regulation of TFs: *SMAD2* and *SMAD3*. GSEA analysis [178] showed that these DR genes are associated with *EGFR1 pathway*, *regulation of apoptosis*, *cell cycle checkpoints*, *G1 cell cycle arrest*, and *regulation of apoptosis* (Fig. 5.4 D. inset, Supplementary Table S3). Once again, scTenifoldNet identified a much smaller set of significant genes compared to those reported in the original paper [211]: 125 DR genes vs. $8,389$ DE genes. Nevertheless, functional analyses show that scTenifoldNet identified a more specific gene set relevant to cetuximab's mechanism of action. Further scrutinization of enriched molecular functions of these DR genes will help to identify more regulatory targets induced by cetuximab in HNSCC cells.

5.3.8.4 Analysis of transcriptional responses of alveolar type 1 cells to Nkx2-1 gene knockout

In the third example, we applied scTenifoldNet to another published scRNAseq data from type 1 alveolar (AT1) cells [212]. AT1 cells are responsible for gas exchange, the physiological function of the lung [229]. Little *et al.*, [212] found that NK homeobox 2-1 (*Nkx2-1*) is expressed in AT1 cells and thought *Nkx2-1* might be essential to the development and maintenance of AT1

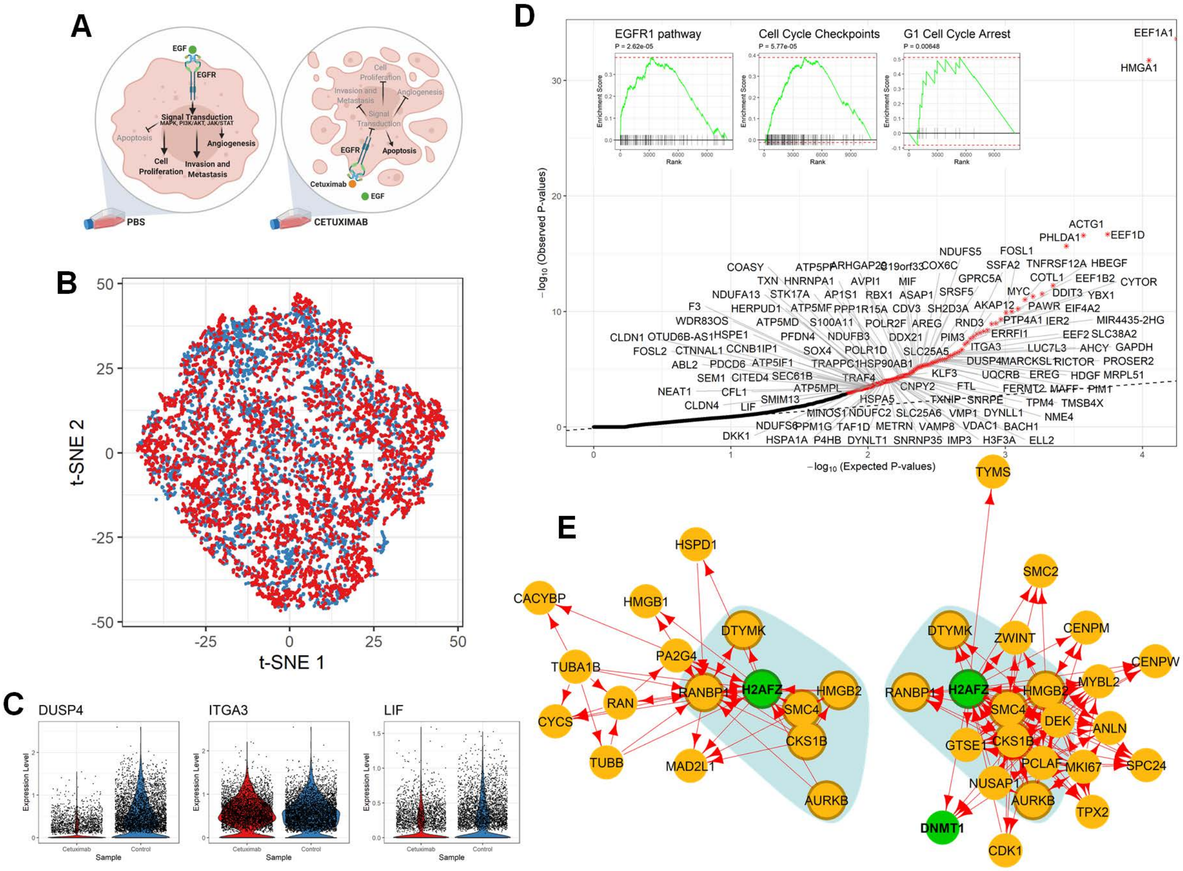


Figure 5.4: Analysis of transcriptional responses of a carcinoma cell line to cetuximab. (A) Illustration of experimental design, including sample groups and the known mechanism of drug action, in the study of cetuximab resistance of HNSCC cell lines [211]. (B) t-SNE plot of $5,217$ and $4,507$ HNSCC-SCC6 cells treated with cetuximab (red) and PBS (blue), respectively. (C) Violin plots show the log-normalized expression levels of selected DR genes in SCC6 cells with and without cetuximab treatment. (D) Q-Q plot for observed and expected p-values of the $7,503$ genes tested. Genes ($n = 25$) with FDR < 0.05 are labeled with asterisk. Inset shows the results of the GSEA analysis for genes ranked by their distances in manifold aligned scGRNs from young and old mice. (E) A representative module with DR genes and corresponding subnetworks in two scGRNs. The module is enriched with DR genes and the corresponding subnetworks in two scGRNs. For illustrative purposes, the module is centered on the DR gene, *H2AFZ*. The colors, edges, and marks are presented as in Fig. 5.3 E.

cells. To determine the function of *NKX2-1* during the development of AT1 cells, they performed scRNAseq experiment to obtain transcriptome profile of cells from the lungs of *Nkx2-1*$^{CKO/CKO}$; *Aqp5*$^{Cre/+}$ mutant mice (i.e., knockout [KO] mice) and littermate controls (i.e., wild-type [WT] mice). They used early infant mice (postnatal day 10, P10) because P10 represents an intermediate time point when individual AT1 cells in the mutant lung are expected to collectively feature the full range of transcriptomic phenotypes. They reported $3,622$ DE genes ($2,105$ upregulated and $1,517$ downregulated, Supplementary Dataset S1 of [212]) between the KO and WT mice. Their analyses suggest that, without *Nkx2-1*, developing AT1 cells lose their molecular markers, morphology, and cellular quiescence, leading to aberrant expression of gastrointestinal (GI) genes, alveolar simplification and lethality (Fig. 5.5 A.).

To evaluate the power of scTenifoldNet in identifying regulatory changes caused by gene KO, we re-analyzed the transcriptional profiles of $2,397$ mutant AT1 cells from the *Nkx2-1*$^{CKO/CKO}$; *Aqp5*$^{Cre/+}$ mice and 638 AT1 cells from the WT mice (Fig. 5.5 B.). Expression levels of *Cd24a*, *Fau* and *Eef1a1* in AT1 cells of KO and WT mice are shown in Fig. 5.5 C. *Cd24a* is a marker gene for AT1 cells; *Fau* and *Eef1a1* are GI genes, known to be highly expressed in the GI tissues. Using scTenifoldNet, we identified 29 genes exhibiting significant difference in their regulation between the two samples: KO vs. WT (FDR < 0.05, Fig. 5.5 D.). These 29 genes are: ***Cd24a***, *Clu*, ***Muc1***, *Stard10*, *Glul*, ***Fxyd3***, ***Gsto1***, *Eef1a1*, *Bag1*, ***Atp1b1***, *Txnip*, *Csrp2*, ***Tspan1***, *Nr2f2*, ***Elf3***, *Sepp1*, *Pabpc1*, *Lurap1l*, *Gnb2l1*, *Eef2*, *Smim6*, *Cox7a2l*, ***Tpt1***, *Fau*, *Eef1b2*, *Eif3f*, *Atpif1*, ***0610040J01Rik*** and ***Krt19***. Targets of *Sox2* [230] are in bold. As reported [212], this gene list is enriched with genes highly expressed in the intestine. Using GSEA analysis [178], we showed the significant enrichment of gastrointestinal marker genes [29] (Fig. 5.5 D., insets), which confirmed the effect of *Nkx2-1* KO on the cellular identity of AT1 cells.

5.3.8.5 *Analysis of transcriptional responses of human dermal fibroblasts to the double-stranded RNA stimulus*

Next, we show the use of scTenifoldNet to a scRNAseq data set from human dermal fibroblasts [117]. In the original paper, Hagai *et al.* [117] focused on single-cell transcriptional responses

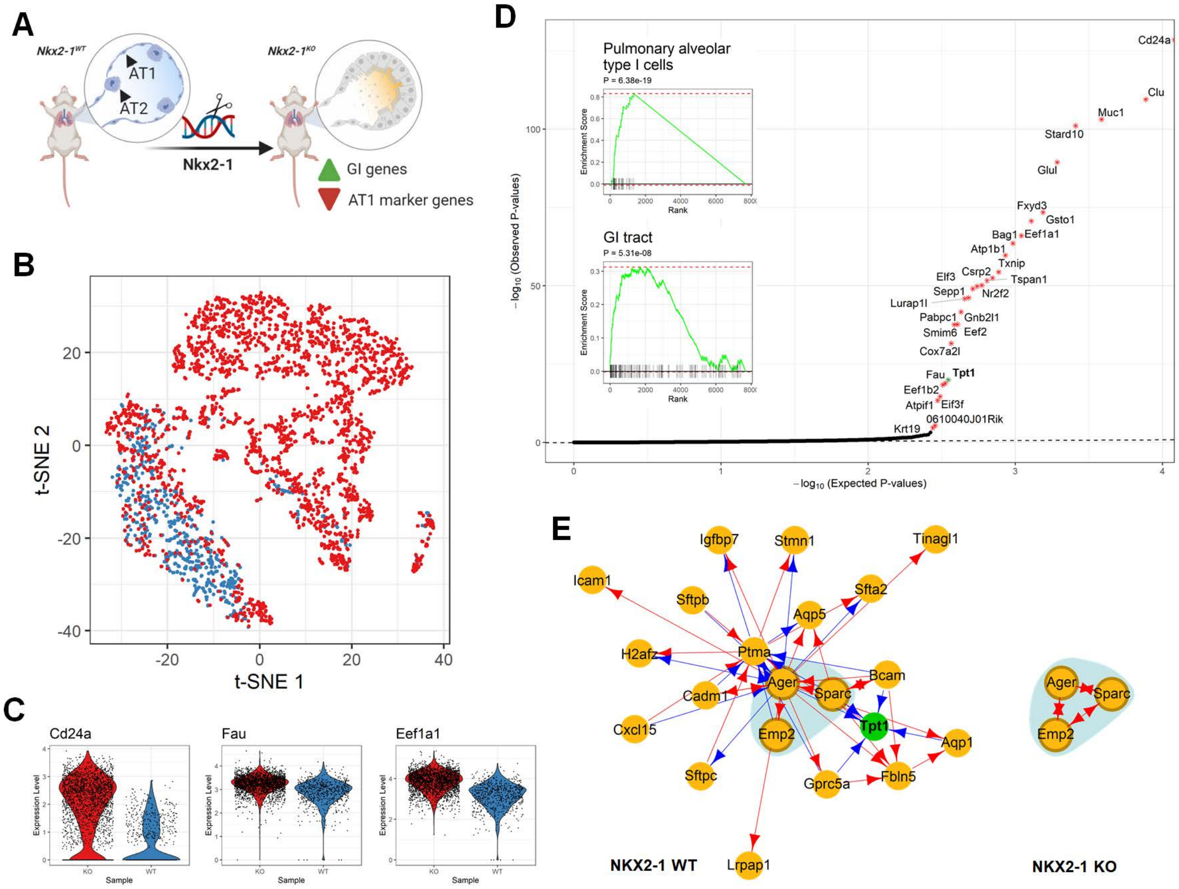


Figure 5.5: Analysis of transcriptional responses of alveolar type 1 cells to *Nkx2-1* gene knockout. (A) Illustration of experimental design and data collection of the KO experiment [212]. (B) t-SNE plot of $2,397$ and 638 AT1 cells from *Nkx2-1* KO mice (red) and WT mice (blue). (C) Violin plots show the log-normalized expression levels of selected DR genes in KO (red) and WT (blue) mice. (D) Q-Q plot for observed and expected p-values of tested genes. Genes ($n = 29$) with FDR < 0.05 are labeled with asterisk. Inset shows the results of the GSEA analysis for genes ranked by their distances in manifold aligned scGRNs. (E) A representative module that contains DR gene, *Tpt1*, in the WT mice. Most parts of the module disappear in the KO mice. The colors, edges and marks are presented as in Fig. 5.3 E.

induced by the stimulus of polyinosinic-polycytidylic acid (polyI:C), a synthetic double-stranded RNA (dsRNA)(Fig. 5.6 A.). They obtained and compared transcriptomes of $2,553$ unstimulated and $2,130$ stimulated cells and identified 875 DE genes (Supplementary Table S3 of [117]). These DE genes include IFNB, TNF, IL1A, and CCL5, encoding antiviral and inflammatory gene products, and are enriched for inflammatory response, positive regulation of immune system process, and response to cytokine, among many others biological processes and pathways. We found the original scRNAseq data has a batch effect between two samples, but the global batch effect can be removed using Harmony [231], as shown in the tSNE-plot of cells of two samples (Fig. 5.6 B.). Nevertheless, the differences in the expression level between samples can still be detected in selected genes with Harmony-processed data (Fig. 5.6 B.). Applying scTenifoldNet to the processed data, we identified 29 DR genes: ***SOD2***, ***GBP1***, *WARS*, *ZC3HAV1*, ***EGR1***, ***BBC3***, *ISG15*, ***HLA-B***, *ZFP36*, ***PPP1R15A***, ***JUN***, *IFI6*, ***JUNB***, ***B2M***, *APOL2*, ***HLA-A***, ***IER3***, *SAT1*, ***NFKBIA***, ***NNMT***, ***FN1***, *IFITM3*, *MEG3*, *NEAT1*, *COL1A1*, *PLEKHA4*, *EEF1A1*, *SOCS1*, and *SERF2* (Fig. 5.6 D., Supplementary Table S4). Among them, 14 (highlighted in bold) are targets of TF, *RELA* [232] (48%, adjusted p-value < 0.01, enrichment test by Enrichr [123, 124]). These DR genes are functionally enriched for *interferon signaling*, *immune system*, *interleukin-1 regulation of extracellular matrix*, and among others (Supplementary Table S4).

Once again, scTenifoldNet reports fewer genes than DE analysis does in the original paper [117]. Through comparing DR genes with the DE genes, we found that enriched functions of DE genes reflect the differences between unstimulated cells and cells that have completed an initial response to dsRNA stimulus and reached a final phase of the response, whereas the enriched functions of DR genes reflect ongoing activities associated with regulatory changes and immune responses to the stimulus. In this sense, DR genes are valuable for informing of mechanisms, through which the dsRNA acts to induce immunological responses [233, 234, 235]. For example, it is known that the dsRNA inhibits the translation of mRNA to proteins [234] and leads the synthesis of interferon, which induces the synthesis of ribosomal units that are able to distinguish between cell mRNA and viral RNA [235]. Interferon also promotes cytokine production that activates the immune responses and induces inflammation [233]. To further illustrate the changes in the regulatory patterns between samples, we plotted the GRN module around *EEF1A1*. It can be seen that, before and after the dsRNA treatment, the interacting partnership of the genes is changed substantially (Fig. 5.6 E.). Two scatter plots show the change of correlation between *TPT1* and *ANXA2*, as an example (Fig. 5.6 F.). The negative correlation between the two genes' expression among cells disappears after the dsRNA treatment and, thus, the two genes are not linked in the scGRN constructed using the after-treatment data.

5.3.8.6 Analysis of transcriptional responses of mouse neurons in Alzheimer's disease

Lastly, we applied scTenifoldNet to scRNAseq data of isolated single nuclei from the brains of the WT and $5\times$FAD mice [213]. The $5\times$FAD strain recapitulates the major features of Alzheimer's disease amyloid pathology. The genotype of these mice contains several Familial Alzheimer's Disease (FAD) mutations in *APP* and *PSEN1*, causing the overexpression of mutant human amyloid-beta (Aβ) precursor protein and human presenilin 1. The $5\times$FAD model rapidly develops amyloid pathology, with high levels of intraneuronal Aβ accumulation beginning around 1.5 months of age, and extracellular Aβ deposition beginning around two months [236].

In the original paper [213], Zhou et al. compared single-cell gene expression between 6-month-old WT mice with 6-month-old $5\times$FAD mice. They found that neurons show limited responses to Aβ peptides—compared to microglia and oligodendrocytes, neurons show minimal transcriptional changes (149 DE genes) between WT and $5\times$FAD mice. To test whether scTenifoldNet can detect genes whose expression is differentially regulated between WT and $5\times$FAD mice, we decided to apply our method to this scRNAseq data, exclusively in neurons. We downloaded expression data matrices from the GEO database using accession number GSE140511 and extracted expression data of neurons from two samples: WT (GSM4173505) and WT $5\times$FAD (GSM4173511)(Fig. 5.7 A.).

After re-analyzing the data using scTenifoldNet, we identified 18 DR genes: *Zdhhc17*, *Chl1*, *Abhd17b*, *Rchy1*, *Stmn2*, *Tjp1*, *Nrbp2*, *Ly6h*, *Smarcd1*, *Rhbdd2*, *Ndfip1*, *Mark2*, *Icam5*, *Fam92a*, *Rgl1*, *Gmcl1*, *Daam1*, and *Fxr1* (FDR < 0.05, Fig. 5.7 D.). For functional enrichment analy-

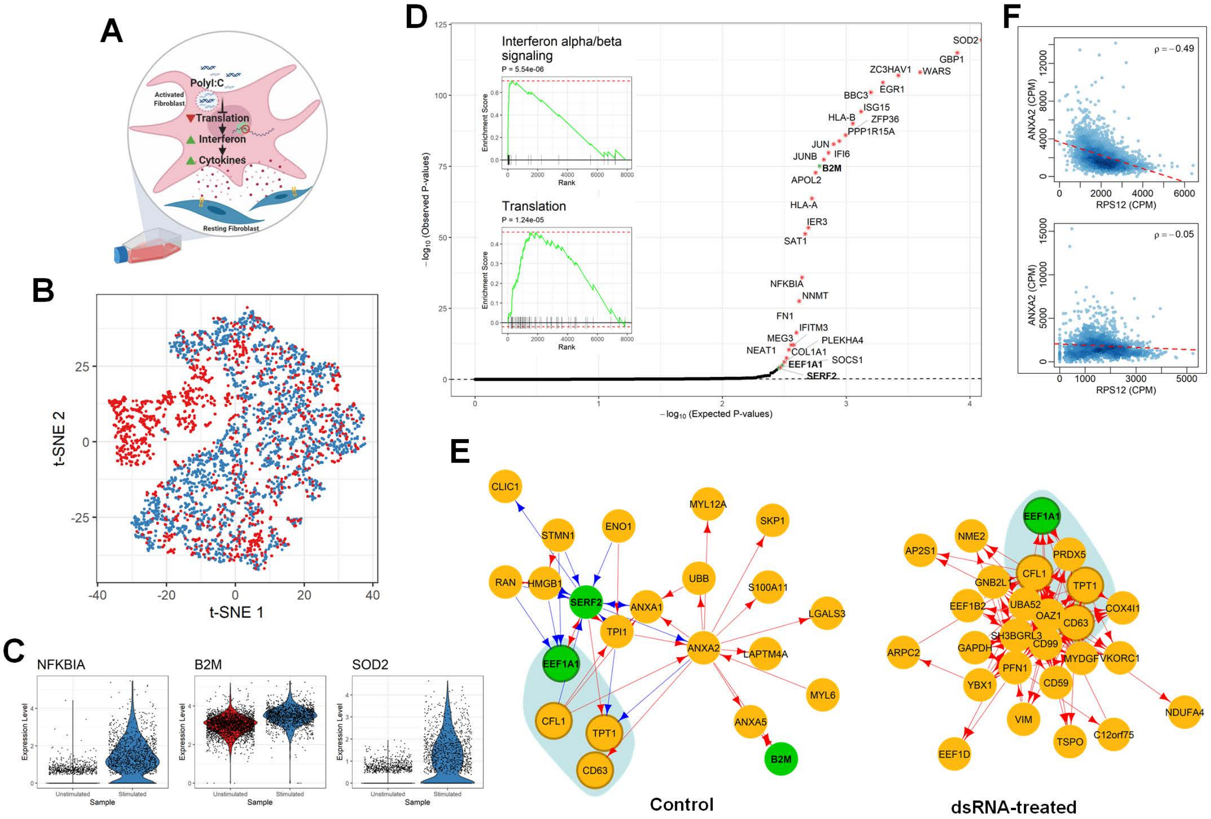


Figure 5.6: Analysis of transcriptional responses of human dermal fibroblasts to the double-stranded RNA stimulus. (A) Illustration of experimental design and tested mechanism of transcriptional responses [117]. (B) t-SNE plot of human dermal fibroblasts before (blue) and after (red) dsRNA stimulus. (C) Violin plots show the log-normalized expression levels of selected DR genes before (blue) and after (red) stimulus. (D) Q-Q plot for observed and expected p-values of tested genes. Genes ($n = 29$) with FDR < 0.05 are labeled with asterisk. Inset shows the results of GSEA analysis for genes ranked by their distances in manifold aligned scGRNs. (E) Comparison of a representative module that contains three DR genes in the control sample. The colors, edges and marks are presented as in Fig. 5.3 E. (F) Scatter plots show the correlation between *TPT1* and *ANXA2* before (top) and after (bottom) dsRNA stimulus.

sis, we relaxed the significant-gene cutoff to include 57 additional genes with FDR ≥ 0.05 but nominal p-value < 0.05. These additional genes include: *Apoe* and *Bin1*. *Bin1* encodes bridging integrator 1 (also known as amphiphysin 2), which is the second most important risk locus (after *Apoe*) for the late onset Alzheimer's disease [237, 238]. *Apoe* and *Bin1* rank 25th and 61th, respectively, in the list of 75 significant genes (18 genes with FDR < 0.05 followed by 57 genes with nominal p-value < 0.05), both play a role in *negative regulation of amyloid precursor protein catabolic process* and *tau protein binding*. Enrichr analysis [124, 123] reported following top GO terms: *regulation of neuron projection development*, *positive regulation of cell projection organization*, *phosphatidylserine metabolic process*, and *protein acylation*, *potassium channel activity* and *methylation-dependent protein binding*. GSEA analysis [178] showed that regulatory changes are associated with *integrin signaling pathway*, *serotonin HTR1 group* and *FOS pathway*, and *glutamate neurotransmitter release cycle* (Fig. 5.7 D., insets).

5.4 Discussion

We present scTenifoldNet, a robust, unsupervised machine learning workflow that streamlines comparative GRN analyses with data from scRNAseq. The key feature of scTenifoldNet is to apply comparative network analysis with scRNAseq data. It detects differences in the cell population's state between two samples in a sensitive and scalable manner. It provides the function of differential regulation (DR) analysis, which can be used to reveal subtle regulatory shifts of genes.

Today, differential expression (DE) analysis is still the primary method for the purpose of comparative analysis between scRNAseq samples (see, e.g., [210, 117, 239]). As scRNAseq data sets are becoming widely available, there will be more and more interest in comparing between samples. The scTenifoldNet-based DR analysis is expected to be adapted in more scenarios wherever DE analysis is applicable. scTenifoldNet learns and contrasts high-dimensional features of genes in scGRNs by examining global interactions between the genes. scTenifoldNet is more suitable for comparing highly similar samples, such as two populations of cells of the same type. scTenifoldNet is built as a robust, sensitive tool that can capture signals that are even confined to rare cell types.

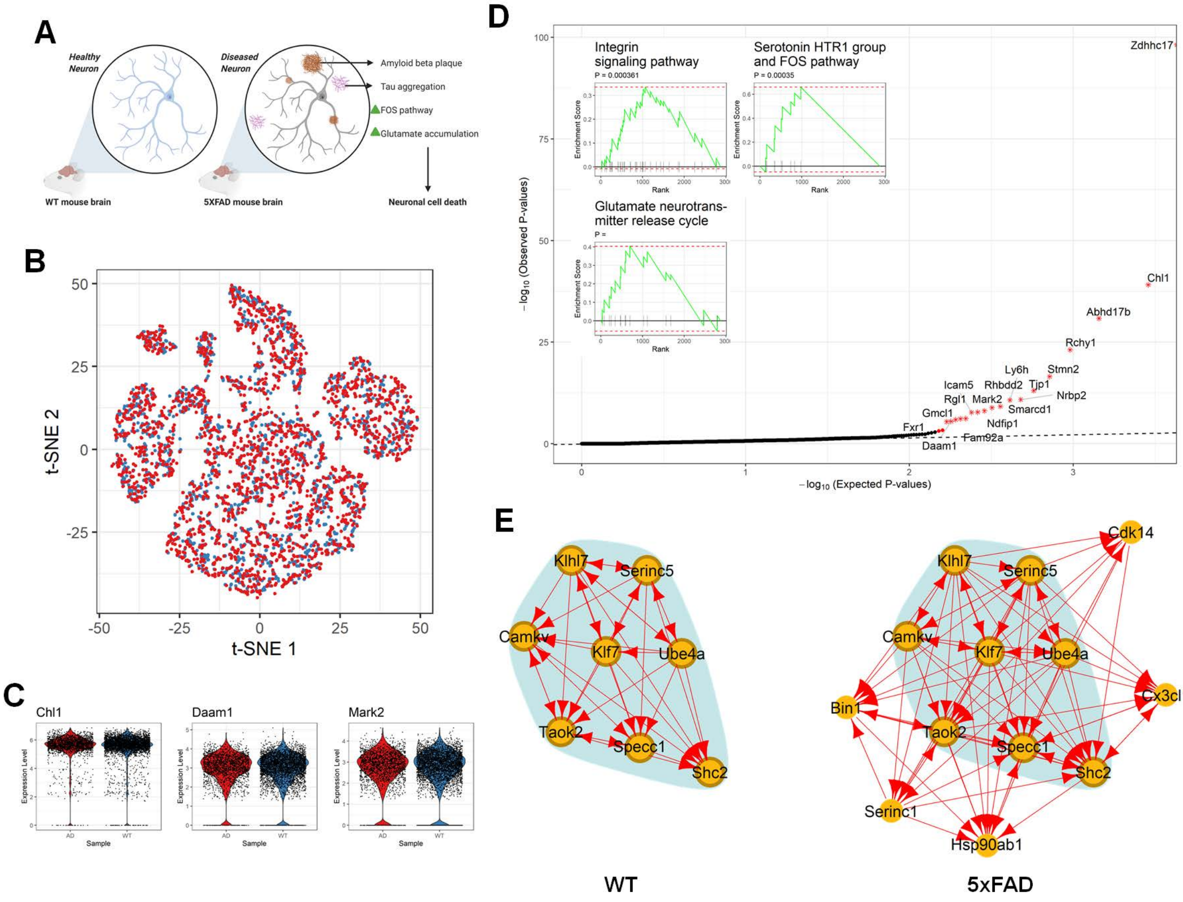


Figure 5.7: Analysis of transcriptional responses of neurons to amyloid-beta (Aβ) peptides in the 5×FAD mice, a model of Alzheimer's disease. (A) Illustration of experimental design and data collection of the 5×FAD mice study [213]. (B) t-SNE plot of neurons of the 5×FAD (red) and WT (blue) mice. (C) Violin plots show the log-normalized expression levels of selected DR genes in neurons of the 5×FAD (red) and WT (blue) mice. (D) Q-Q plot for observed and expected p-values of tested genes. Genes ($n = 18$) with FDR < 0.05 are labeled with asterisk. Inset shows the results of the GSEA analysis for genes ranked by their distances in manifold aligned scGRNs. (E) Comparison of a representative module that contains top-ranked DR genes between the two scGRNs. The colors, edges and marks are presented as in Fig. 5.3 E.

To achieve technical requirements, we overcome several analytical barriers in developing scTenifoldNet. First, constructing scGRN from scRNAseq data, which consists of cells in many different states, is challenging at present. It is also difficult to control for technical noise in the data. To address these issues, we let scTenifoldNet begin with random cell subsampling. It is worth noting that random cell subsampling can not only help dealing with the problem of cell heterogeneity, additional information of cells can be incorporated into subsampling schema. More specifically, in addition to the random subsampling using jackknife and bootstrap methods, we can adapt a semi-random subsampling schema, if cells in an input matrix are sorted according to pseudotime [240]. These cells can be subsampled using a pseudotime-guided method, with which sorted cells are sampled along the pseudotime trajectory. In such a way, the subsamples contain pseudotime information, and the multilayer scGRN constructed from these subsamples will contain the pseudotime-series information. In machine learning, many multilayer network analysis algorithms have been proposed [241, 242, 243]. With our pseudotime-series scGRN data, these algorithms will be relevant and applicable. Second, regulatory relationships between genes from scRNAseq data are difficult to establish, even though the data may theoretically capture a complete picture of the regulatory gene landscape. We consider PC-regression to stand out as a crucial method of building scGRNs. PC-regression significantly outperforms the other GRN construction algorithms in all aspects of methodology metrics, including specificity, sensitivity, computational efficiency, and the required minimum number of cells. Importantly, PC-regression explicitly projects thousands of gene expression measurements into a low dimensional space to capture much of the observed variation. PC-regression, therefore, establishes the relationship for each pair of genes after controlling for the most important background interactions. Third, in scTenifoldNet, the tensor denoising procedure effectively smooths edge weights across all networks in multilayer scGRNs. Fourth, scTenifoldNet performs nonlinear manifold alignment to align two networks. As such, two networks can be contrasted directly, and DR genes could be detected using distance in new coordinates of data in a low-dimensional space.

We validate the power of scTenifoldNet using real data sets coming from various studies and demonstrate that scTenifoldNet is sensitive to signals. Five real scRNAseq data sets are involved (Table 5.1). These five data sets have one thing in common: they all have two sets of scRNAseq data – one from a treated group and the other from a control/untreated group. More importantly, in all five cases, we have sufficient prior knowledge about the biological system, from which the data is collected. Therefore, we have hypotheses about what transcriptional changes are expected to see before doing the analysis. For example, in the morphine response analysis, the causal factor of transcriptional responses, i.e., the morphine stimulus, is known and thus, we know what should be recovering through the analysis. Similarly, we had some clues in the examples of cetuximab and fibroblasts about what transcriptional changes we might be able to retrieve. By compiling all the findings from scTenifoldNet applications, we tested scTenifoldNet and showed that scTenifoldNet provides findings that are precise, specific and relevant to the biological systems and questions in the test. It is of significance to building a specific and sensitive tool like scTenifoldNet for the purpose of molecular mechanism studies using scRNAseq. This is because causal factors and their target genes remain unknown in many biological systems studied. If this is the case, it is crucial to apply the sensitive approach like scTenifoldNet, which may be in addition to the DE analysis, to unveil more gene candidates. Only then will we be able to scrutinize identified genes further to learn the mechanisms behind their actions in the whole system. We face such a challenge in many studies from unknown factors that cause the disorder. It is therefore critical that we adopt tools such as scTenifoldNet, instead of relying solely on conventional DE analysis, to tackle this big data analysis problem.

In summary, scRNAseq enables the study of cellular, molecular components, and dynamics of complex biological systems at single-cell resolution. To unravel the regulatory mechanisms underlying cell behaviors, novel computational methods are essential for understanding the complexity in scRNAseq data (e.g., scGRNs) that surpasses human interpretative ability. We anticipate that, when applied to real scRNAseq data, our machine learning workflow implemented in scTenifoldNet, can help achieve breakthroughs by deciphering the full cellular and molecular complexity of the data through constructing and comparing scGRNs.

6. SCTENIFOLDKNK: A MACHINE LEARNING WORKFLOW PERFORMING VIRTUAL KNOCKOUT EXPERIMENTS ON SINGLE-CELL GENE REGULATORY NETWORKS

6.1 Introduction

Gene knockout (KO) experiments are a proven approach for studying gene function. A typical KO experiment involves the phenotypic characterization of organisms that carry the gene KO. For example, in KO mice, a gene is made inoperative or knocked out using genetic techniques. Deletion of one or more alleles provides mechanistic insight as to how the target gene functions within a particular biological context. Gene function may be predicted from phenotypic differences between the KO and wild-type (WT) samples, and its expression is a molecular phenotype that can be quantitatively measured. Gene expression is regulated in a coordinated manner in all living organisms. Typically, regulation is observable as synchronized patterns of transcription in a gene regulatory network (GRN). Co-regulation and co-expression are often seen among genes associated with the same biological processes and pathways, or regulated by the same master transcription factors [25]. If one gene is knocked out, functionally related genes can mediate a homeostatic response. Thus, in unraveling regulatory mechanisms and synchronized patterns of the cellular transcriptional activities, network analysis of gene expression data in KO experiments provides mechanistic insights.

Here we present a machine learning workflow, called scTenifoldKnk that can be used to perform virtual KO experiments from the topology of the constructed GRNs. scTenifoldKnk takes expression data from scRNA-seq of wild-type (WT) samples as input and constructs a denoised single-cell GRN (scGRN) using principle component (PC) regression and tensor decomposition. The WT scGRN is copied and then converted to a pseudo-KO scGRN by artificially zeroing out the target gene in the adjacency matrix. To compare the two scGRNs (WT vs. pseudo-KO), a quasi-manifold alignment method is adapted [200, 199]. Through comparing the two scGRNs, scTenifoldKnk predicts changes in transcriptional programs and assesses the impact of KO on the

WT scGRN. This information is then used to elucidate functions of the KO gene in analyzed cells through enrichment analysis.

scTenifoldKnk is computationally efficient enough to allow the method to be applied to systematic KO experiments. In such a systematic study, we assume that thousands of genes in analyzed cells will be knocked out one by one. As mentioned, due to the experimental and biological limitations, such systematic KO experiments would never be possible in a real-animal experimental setting. The other features of scTenifoldKnk include: (1) scTenifoldKnk requires no data from KO animals or cell lines, as it only utilizes the scRNA-seq data from WT samples; and (2) scTenifoldKnk can be used to perform double-KO experiments – that is, to knock out more than one gene at a time. The remainder of this paper is organized as follows. We first present an overview of the workflow of scTenifoldKnk and use simulated data to demonstrate its basic functions. We then use existing data generated from authentic-animal KO experiments to highlight the use of scTenifoldKnk. These existing data sets contain scRNA-seq expression matrices from both WT and KO samples. Although the KO data sets were available, they were not used by scTenifoldKnk as input. Instead, the KO data sets were specifically used as positive controls to show that scTenifoldKnk can produce expected results. Next, we show applications of scTenifoldKnk to delete genes that cause three different Mendelian diseases with known phenotypic outcomes. Finally, we conduct our own KO animal experiments to validate the utility of scTenifoldKnk as a predictive tool and discuss the contribution of scTenifoldKnk to the future of predictive biology.

6.2 Materials and Methods

scTenifoldKnk is a machine learning workflow for virtual KO experiments with scRNA-seq data. It utilizes a scRNA-seq expression matrix from a WT sample as input, without using any data from a KO sample, to predict regulatory network changes and perturbed genes caused by the KO of a gene.

6.2.1 Construction of the WT scGRN

scTenifoldKnk uses a method we proposed previously – scTenifoldNet [244] – to construct scGRNs. The procedure of scGRN construction consists of three steps: cell subsampling, principal component (PC) regression, and tensor decomposition. In cell subsampling, cells are randomly sampled to form subsets. This strategy assures that the outliers in the sample are not included by many sub-sample sets and reduce the influences of outliers. When constructing the scGRN for each sub-sample set, both the accuracy and computational efficiency of the algorithms should be considered. Among existing methods, PCNet based on PC regression is an efficient approach to construct scGRN. It can also surpass the multicollinearity problem when explanatory variables are linearly correlated and provides stable constructing results. Finally, to aggregate multiple GRNs, scTenifoldNet uses the CP tensor decomposition algorithm. CP decomposition factorizes the multilayer scGRN and regenerates all adjacency matrices using top components, which achieves denoising and enhancing, i.e., making main signals stronger and other signals weaker. Overall, scTenifoldNet provides relatively stable and accurate results in constructing GRN. We briefly describe each step below. Additional details can be found in a previous publication [244].

1. Cell subsampling: Initially, scTenifoldNet builds several subsets of cells via random sampling. Denote $X \in \mathbb{R}^{n \times p}$ as the scRNA-seq data matrix that reflects gene expression levels for p genes in n cells. A sub-sample set of cells is constructed via randomly sampling m ($< n$) cells in X. By repeating this subsampling process for t times, t sub-sample sets of cells are derived, denoted as $X'_1, \dots, X'_t \in \mathbb{R}^{m \times p}$.

2. Network construction: For each X'_i, scTenifoldNet builds a GRN with an adjacency matrix W^i via PC regression, where a PC analysis (PCA) is applied to the original explanatory variables and then the response variable is regressed on a few leading PCs. Since PC regression only utilizes d PCs ($d \ll n$) as the covariates in regression, it mitigates over-fitting and reduces the computation time. To build an scGRN, each time scTenifoldNet focuses on one gene (referred to as the response gene) and applies PC regression. The expression level of

the response gene is used as the response variable, and the expression levels of other genes are used as the explanatory variables in PC regression. scTenifoldNet repeats this process for another $N-1$ times, with one different gene as the response gene each time. In the end, scTenifoldNet collects the coefficients of N regression models together and forms a $p \times p$ adjacency matrix W^i, whose (i, j) entry saves the coefficient of the ith gene on the jth gene. W^i could reflect the interaction strengths between each pair of genes.

3. Network denoising: The adjacency matrices of the t networks $W^1, \ldots, W^t$ can be stacked to form a third-order tensor $\Xi \in \mathbb{R}^{(p\times p\times t)}$. To remove noise and construct an overall adjacency matrix, scTenifoldNet applies CANDECOMP/PARAFAC (CP) tensor decomposition to Ξ to extract important latent factors. More specifically, scTenifoldNet approximates Ξ by Ξ_R: $\Xi \approx \Xi_R = \sum_{r=1}^{R} \lambda_r a_r \circ b_r \circ c_r$, where $\circ$ denotes the outer product, $a_r \in \mathbb{R}^p$, $b_r \in \mathbb{R}^p$, and $c_r \in \mathbb{R}^t$ are unit-norm vectors, and λ_r is a scalar. The reconstructed tensor $\Xi_R \in \mathbb{R}^{(p\times p\times t)}$ includes t denoised adjacency matrices, and by taking the average of them, scTenifoldNet obtains the overall stable adjacency matrix. After further normalizing its entries by dividing them by their maximum absolute value, scTenifoldNet generates the final adjacency matrix of scGRN for the given sample. For later use, denote it as W_d.

4. Obtaining enhanced directed network: We provided an option to get the strictly directed scGRN. Given the denoised network W_d, for each pair (i, j) and (j, i), only the entry whose absolute value is large will be left. More specifically, we defined the (i, j) entry for the strictly directed network W_s by $W_s\,(i,\ j) = W_d(i,j)$ if $|W_d(i,j)| > |W_d(j,i)|$ and $W_s(i,j) = 0$ otherwise. Note that if $|W_d(i,j)| < |W_d(j,i)|$, then $W_s(i,j) = 0$, but it is believed that $W_d(i,j)$ still contains some information. Instead of using restricted directed network, we can also come up with the ‘interpolated network’ W_i which contains part of the information of $W_d\,(i,j)$. Given the hyperparameter , the “interpolated network” is $W_i = \lambda W_s + (1-\lambda)\, W_d$, where $\lambda \in [0,1]$. It is easy to check that when $\lambda = 0$, we get the original denoised network W_d and when $\lambda = 1$, it is to go back to the strictly directed network W_s.

6.2.2 Deletion of the KO gene from the WT scGRN

We propose the virtual KO method that directly works on the WT scGRN (Figure 6.1 B.). The adjacency matrix W_d represents the scGRN constructed using the WT data. In the virtual KO method, the entire row of the adjacency matrix W_d for the target gene is set to zero. We denote the adjacency matrix of the scGRN generated as $\tilde{W}_d$.

6.2.3 Comparison between the WT and pseudo-KO scGRNs

After obtaining W_d and $\tilde{W}_d$, two comparable low-dimensional feature vectors of each gene in the two networks are built and then compared to detect affected genes. Our approach of creating low-dimensional feature vectors is inspired by manifold alignment [203, 199] and its application [201]; our approach is referred to as quasi-manifold alignment because the adjacency matrices used here are not symmetric matrices while they are required to be symmetric in the original procedure. Here W_d and $\tilde{W}_d$ serve as the inputs for manifold alignment and the outputs are the low-dimensional features $F \in \mathbb{R}^{p\times k}$ and $\tilde{F} \in \mathbb{R}^{p\times k}$ of genes before and after knocking out the target gene, where $k \ll n$. Before giving the details of the alignment procedure, we point out that W_d and $\tilde{W}_d$ may include negative values, which reflect the negative correlation between genes. Before doing alignment, we add 1 to all entries in W_d and $\tilde{W}_d$, and the range of W_d and $\tilde{W}_d$ is transformed from $[-1, 1]$ to $[0, 2]$.

To perform quasi-manifold alignment, we first construct a joint adjacency matrix W by combining W_d and $\tilde{W}_d$ together, where $W = \begin{bmatrix} W_d & \frac{\lambda}{2}I \\ \frac{\lambda}{2}I & \tilde{W}_d \end{bmatrix}$. We can treat W as the adjacency matrix of a joint network formed by linking the corresponding genes in two networks. The off-diagonal block of this matrix reflects the corresponding genes between two networks. λ is a tuning parameter. In practice, we select λ as the mean of the row summations of W_d and $\tilde{W}_d$. We further build $\mathbb{F} = [\begin{smallmatrix} F \\ \tilde{F} \end{smallmatrix}] \in \mathbb{R}^{2p\times k}$, and the manifold alignment problem of two networks characterized by the adjacency matrices W_d and $\tilde{W}_d$ is equivalent to the manifold learning problem that finds the low dimensional features $\mathbb{F}$ for the joint network characterized by the adjacency matrix W. For the sake of convenience, we denote $\mathbb{F}_i \in \mathbb{R}^k$ as the ith row of $\mathbb{F}$ that reflects the projection corresponding to the ith gene in the large network. The next step is to build a 'Laplacian' matrix $L = D - W$, where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} (W)_{j,i}$. Denote $f_1, f_2, \ldots, f_d$ as the eigenvectors corresponding to the d smallest nonzero eigenvalues of L. Note that L is not a symmetric matrix. We found that the usual solution of symmetrizing L does not work well with either simulated or real data. We therefore use asymmetric matrix L in our quasi-manifold alignment procedure. Since L is not symmetric, there may be imaginary parts in the eigen decomposition. Based on our experiment, taking only the real part of eigenvectors with respect to the eigenvalue that has the smallest real part will give better overall results. The final low dimensional representation is $\mathbb{F} = [Re(f_1), Re(f_2), \ldots, Re(f_d)]$, where $Re(f)$ means the real part of f.

6.2.4 Test for significance of virtual-KO perturbed genes

The virtual-KO perturbed genes are identified as genes with significant difference in their regulatory patterns in two scGRNs constructed from the WT and KO data. The method for testing the significance of the difference for each gene is described here. With $\mathbb{F} = \begin{bmatrix} F \\ \tilde{F} \end{bmatrix} = [f_1,\ f_2,\ \ldots,\ f_d]$ obtained in manifold alignment, for each gene, we calculate the distance d_j between its two projected feature vectors from two networks. The rankings of d_j are used to help identify significant genes. To avoid arbitrariness in deciding the number of selected genes, we proposed a χ^2 test. Specifically, since d_j^2 is calculated by taking the summation of squares of the differences of projected representations of two samples, its distribution could be approximately χ^2. Instead of d_j^2, we use the scaled fold-change $\frac{df \cdot d_j^2}{\bar{d^2}}$ as the test statistic for each gene j to adjust the scale of the distribution, where df is the degree of freedom. $\frac{df \cdot d_j^2}{\bar{d^2}}$ approximately follows a χ^2 distribution with the degree of freedom (df) if the gene does not perform differently before and after knocking out the target gene. By using the upper tail ($P[X > x]$) of the χ^2 distribution and the Benjamini–Hochberg (B–H) FDR correction [122] for multiple testing correction, we assign a P-value for each gene. To determine df, since the number of the selected significant genes will increase as df increases, we choose $df = 1$ to make a conservative selection of genes with high confidence.

6.2.5 Gene functional annotation and enrichment tests

Identified virtual-KO perturbed genes were evaluated and assessed for their molecular functions and biological processes using Enrichr functional enrichment analysis. The enrichment test results of Enrichr were derived from a large number of reference gene sets. In this study, we reported results derived from the following gene set collections: KEGG 2019 for Human, KEGG 2019 for Mouse, GO Biological Process 2018, GO Cellular Component 2018, GO Molecular Function 2018, BioPlanet 2019, WikiPathways 2019 for Human, WikiPathways 2019 for Mouse, and Reactome 2016. GSEA analysis was conducted using a ranked list of all genes as input and comparing against either KEGG 2019 for Mouse or KEGG 2019 for Human gene sets. The genes were ranked according to their distance between corresponding projection on the aligned manifold between WT and pseudo-KO scGRNs. To remove redundant hits, identified gene sets were compared pairwise. For each pair of identified gene sets, e.g., A and B, if Jaccard index, $J = \frac{|A \cap B|}{|A \cup B|}$, is greater than 0.8, then this pair of gene sets were merged into one. In this way, non-redundant sets of identified gene sets were created and used for gene functional inference. GSEA analysis was also conducted against gene sets of marker genes. The marker genes were obtained from PanglaoDB [29]. The protein interaction enrichment tests were conducted using the web tool of STRING database [245].

6.2.6 Systematic KO analysis using scRNA-seq data in microglia

The scRNA-seq data used in the systematic KO analysis was obtained from the brain immune atlas (https://www.brainimmuneatlas.org), which is a scRNA-seq resource for assessing and capturing the diversity of the brain immune compartment, as published in [246]. The data was generated using the 10× Genomics Chromium platform, including more than $61,000$ CD45+ immune cells from whole brains or isolated dura mater, subdural meninges and choroid plexus of mice. The downloaded data is called the full aggregate data set (combining cells of whole brain and choroid plexus cells from WT + *Irf8* KO mice). The downloaded matrix was processed and the sub-matrix contained $5,271$ microglia from the WT mice. For all genes, the KO perturbation profile of each gene, i.e., a vector of DR distances, was transformed using Box-Cox transformation and then was standardized using z-score transformation. The processed KO perturbation profiles of all genes were combined into one matrix for t-SNE embedding.

6.2.7 *Ahr*$^{-/-}$ KO mouse strain and crypt cell isolation

Animals were housed under conventional conditions, adhering to the guidelines approved by the Institutional Animal Care and Use Committee at Texas A&M University. Stem cell targeted *Lgr5*-EGFP-IRES-Cre^{ERT2}, *AhR*$^{f/f}$ and tdTomato$^{f/f}$ mouse strains have all been previously described [247]. The mouse genotypes used in this study include: tamoxifen inducible - *Lgr5*-EGFP-Cre^{ERT2} $\times$ Tomato$^{f/f}$ (WT, control), *AhR*$^{f/f}$ $\times$ *Lgr5*-EGFP-Cre^{ERT2} $\times$ Tomato$^{f/f}$ (KO). Male mice ($n = 5$ per genotype, $8 - 10$ weeks of age) were intraperitoneally injected with 2.5mg of tamoxifen (Sigma, T5648) dissolved in corn oil (25mg/ml) once a day, for four consecutive days. Mice were maintained on an AIN-76A, semi-purified diet (Research Diets, D12450B), fed ad libitum and housed on a 12h light-dark cycle. For all experiments, littermate controls were cohoused with the KO mice. Two weeks post tamoxifen injection, colons were removed, washed with cold PBS without calcium and magnesium (PBS-/-), everted on a disposable mouse gavage needle (Instech Laboratories) and incubated in 15mM EDTA in PBS-/- at 37°C for 35 min as previously described [247]. Subsequently, following transfer to chilled PBS-/-, crypts were mechanically separated from the connective tissue by vigorous vortexing. Cell suspensions were then filtered through a 40μm mesh and Tomato-expressing cells (includes GFP^{+}/Tom^{+} as well as GFP negative/Tom^{+}) were collected using a MoFlo Astrios Cell Sorter (Beckman Coulter) or a Bio-Rad S3e Cell Sorter. Tomato positive cells represent colonic stem cells and their progeny. Dead cells were excluded by staining with propidium iodide or 7-AAD. Samples were subsequently processed using the 10× Genomics scRNA-seq pipeline described below. A total of $62,741$ cells from 10 mice were sequenced. These included $34,889$ sorted colonocytes from the WT and $27,852$ from the KO mice. The average number of genes detected per sample was $20,141$. From all sequenced cells, $40,690$ ($21,263$ from WT and $19,427$ from KO samples) were removed with the scRNA-seq quality control procedure. These cells had either a high proportion of mitochondrial reads (greater than 10%), or exhibited an extremely large or small library size.

6.2.8 *Malat1*$^{-/-}$ KO mouse strain and pancreatic islet cell isolation

Malat1 null mice were obtained from Texas A&M Institute for Genomic Medicine (TIGM). Islets were isolated based on published procedures [248, 249, 250]. Briefly, mice were killed by cervical dislocation. The pancreas was inflated with 3mL of cold collagenase solution (0.3mg/mL) (Collagenase type 4, Worthington Biochemical Corporation) through the common bile duct with a 20G needle, starting at the gallbladder. The pancreas was then removed from the body and placed in a siliconized vial containing 2mL of 1mg/mL collagenase solution, and digested at 37°C in a water bath for $\approx 12-15$ min. After three washes of the digested pancreas with centrifugation (97g at 4°C for 1 min) to collect the tissues, islets were purified by density gradient centrifugation (cold polysucrose/sodium diatrizoate solution 1.1119g/ml, 560g at 4 °C for 15 min) and then dispersed in 96 well plates and cultured in RPMI (Gibco) with 10% FBS for further experiments. To determine islet function and confirm success of isolation of the functional islets, the glucose-stimulated insulin secretion level was measured using a highly sensitive sandwich ELISA assay kit following manufacturer's instructions (EMD Millipore, EZRMI-13K). After confirming the function of isolated islets, the islets were washed in 10ml of PBS and dissociated using 2ml of Accutase (Sigma-Aldrich, CatA6964) and incubated at room temperature for 20mins with gentle intermittent mixing by pipetting. The dissociated islets were confirmed using microscopy and were stopped with islet culture media (RPMI with 10% FBS). The dissociated islets were then filtered through a 40mm filter to remove disassociated tissue and obtain a single-cell suspension of pancreatic islet cells. A total of approximately 20,000 cells were resuspended in 30mL PBS with 0.05% BSA. Samples were subsequently processed using the 10× Genomics scRNA-seq pipeline described below.

6.2.9 10× Genomics scRNA-seq

Single-cell sample preparation was conducted according to Sample Preparation Demonstrated Protocol provided by 10× Genomics as follows: 1mL of cell suspension from each mouse genotype was pelleted in Eppendorf tubes by centrifugation (400g, 5min). The supernatant was discarded, and the cell pellets were then resuspended in 1× PBS with 0.04% BSA, followed by two washing procedures by centrifugation (150g, 3min). After the second wash, cells were resuspended in $\approx 500\mu$L $1\times$ PBS with 0.04% BSA followed by gently pipetting $10 - 15$ times. Cells were counted using an Invitrogen Countess automated cell counter (Thermo Fisher Scientific, Carlsbad, CA) and the viability of cells was assessed by Trypan Blue staining (0.4%). Subsequently, single cell GEMs (Gel bead in EMulsion) and sequencing libraries were prepared using the 10× Genomics Chromium Controller in conjunction with the single-cell 3' v3 kit. Cell suspensions were diluted in nuclease-free water to achieve a targeted cell count of $5,000$ for each cell line. The cDNA synthesis, barcoding, and library preparation were then carried out according to the manufacturer's instructions. Libraries were sequenced in the North Texas Genome Center facilities on a NovaSeq 6000 sequencer (Illumina, San Diego). For the mapping of reads to transcripts and cells, sample demultiplexing, barcode processing, and unique molecular identifier (UMI) countings were performed using the 10× Genomics pipeline CellRanger v.2.1.0 with default parameters. Specifically, for each library, raw reads were demultiplexed using the pipeline command `'cellranger mkfastq'` in conjunction with `'bcl2fastq'` (v2.17.1.14, Illumina) to produce two fastq files: the read 1 file containing 26bp reads, consisting of a cell barcode and a unique molecule identifier (UMI), and the read 2 file containing 96bp reads including cDNA sequences. Reads then were aligned to the mouse reference genome (mm10), filtered, and counted using `'cellranger count'` to generate the gene-barcode matrix.

6.2.10 Data and code availability

The scRNA-seq data of KO mice reported in this paper will be deposited in the NCBI's Gene Expression Omnibus database. scTenifoldKnk has been implemented in R, as well as in Matlab. The source code is available at https://github.com/cailab-tamu/scTenifoldKnk. The R package is also available at the CRAN repository, https://cran.r-project.org/web/packages/scTenifoldKnk/.

6.3 Results

6.3.1 The scTenifoldKnk workflow

scTenifoldKnk takes a single gene-by-cell count matrix from the WT sample as input. Eventually, scTenifoldKnk employs a network comparison method to compare a pseudo-KO scGRN to the WT scGRN to identify differentially regulated genes. These genes are called virtual-KO perturbed genes. From the enriched function of these perturbed genes, the function of the KO gene (i.e., the gene that is virtually knocked out) can be inferred. scTenifoldKnk is implemented with a modular structure containing three core modules illustrated in Figure 6.1 and briefly summarized in three steps as follows.

6.3.1.1 Constructing scGRN with scRNA-seq data from a WT sample

With the scRNA-seq data from a WT sample, scTenifoldKnk first constructs a scGRN using a pipeline we proposed previously, namely scTenifoldNet [244]. This network construction step contains three sub-steps (Figure 6.1 A.):

1. Subsampling cells randomly: Denote X as the scRNA-seq expression data matrix, which contains the expression levels for p genes and n cells. Then, m ($< n$) cells in X are randomly sampled to form X' using an m-out-of-n bootstrap procedure. This subsampling process repeats t times to create t subsets of cells $X'_1, \ldots, X'_t$.

2. Constructing a GRN for each subsampled set of cells: For each X'_i, principal component (PC) regression is run p times to construct a GRN. Each time the expression level of one gene is used as the response variable and the expression levels for the remaining genes as the dependent variables. The constructed GRN from X'_i is stored as a signed, weighted and directional graph, represented with a $p \times p$ adjacency matrix W_i, each of whose columns stores the regression coefficients for the PC regression of a gene. W_i is then normalized via dividing by the maximal absolute value.

3. Denoising adjacency matrices to obtain the final GRN: Tensor decomposition is used to

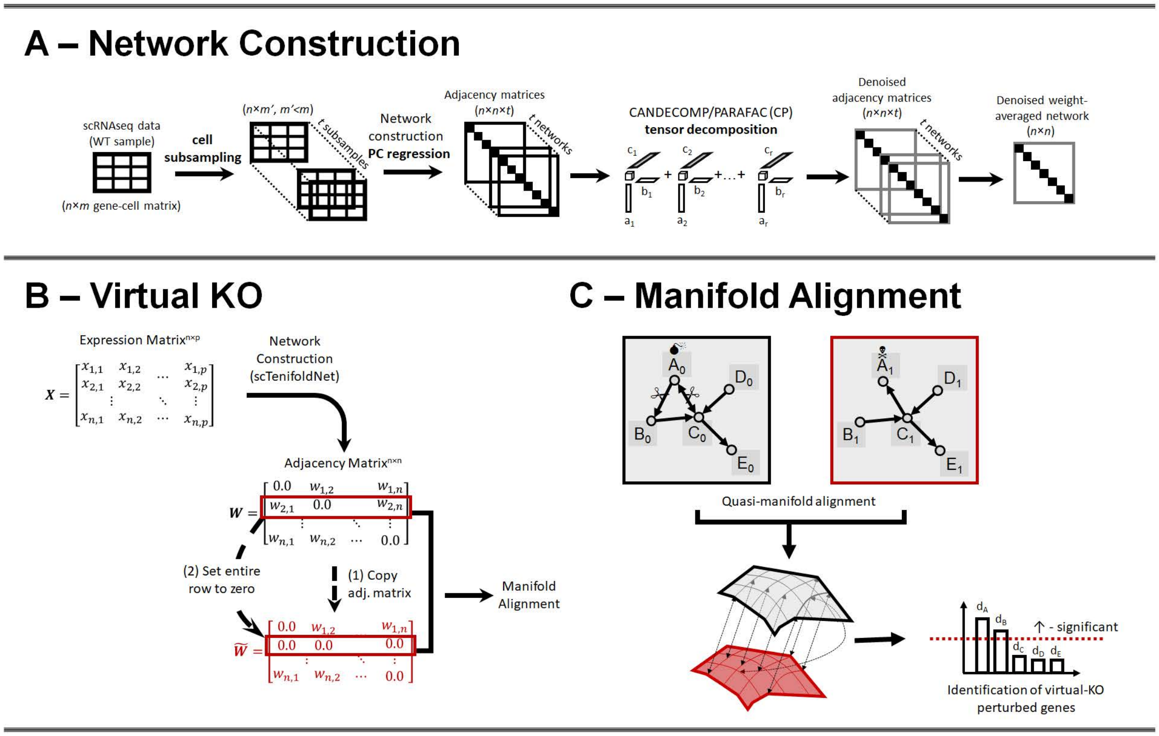


Figure 6.1: Overview of the scTenifoldKnk workflow. scTenifoldKnk is a machine learning framework designed to perform virtual KO experiments with data from scRNA-seq. It consists of three main analytical modules: network construction, virtual KO, and manifold alignment. (A) Network construction. This module consists of three steps: cell subsampling, principal component regression, and tensor decomposition/denoising. (B) Virtual KO. This module starts by duplicating the WT adjacency matrix, W^0, to make $W^1 = W^0$. Then, the entire row of W^1 corresponding to the KO gene is set to zero. The modified W^1 is called the pseudo-KO scGRN. (C) Quasi-manifold alignment. This method is used to learn latent representations of two networks, W^0 and W^1, and align them based on their underlying manifold structures. The distance between a gene's projections with respect to the two scGRNs on the low-dimensional latent representation is used to measure the level of differential regulation of the specific gene. The significantly differentially regulated genes are identified as virtual-KO perturbed genes.

denoise the adjacency matrices W_i obtained from the PC regression step. First, the collection of W_i for t GRNs is processed as a third-order tensor Ξ, containing $p \times p \times t$ elements. Next, the CANDECOMP/PARAFAC (CP) decomposition is applied to decompose Ξ into components. Then, Ξ is reconstructed using top d components to obtain denoised tensors Ξ_d. Denoised W_i in Ξ_d are collapsed by taking the average of edge weights to obtain the final averaged matrix, W_d.

6.3.1.2 Generating pseudo-KO scGRN by virtually knocking out a gene

In the last step, the scRNA-seq expression data matrix from the WT sample, X, is first used to construct the WT scGRN. In this step named virtual KO, the adjacency matrix of the WT scGRN, W_d, is copied, and then the entire row of W_d corresponding to the target gene is set to 0 (Figure 6.1 B.). In this way, the virtual KO operation is performed on W_d directly. The modified W_d is denoted as $\tilde{W}_d$, that is, the adjacency matrix of pseudo-KO scGRN.

6.3.1.3 Comparing scGRNs to identify virtual-KO perturbed genes

In this step, we assume that WT and pseudo-KO scGRNs, W_d and $\tilde{W}_d$, have been obtained. A quasi-manifold alignment method is then used to align W_d and $\tilde{W}_d$ (Figure 6.1 C., see Methods for details). All genes included in the two scGRNs are projected in a k-dimensional space, where $k \ll p$. After the projection, each gene has two low-dimensional representations: one is in respect to W_d and the other $\tilde{W}_d$. For each gene j, d_j is the Euclidean distance between the gene's two projections. The greater d_j, the more significant the differential regulation. Finally, a χ^2 test is applied to detect significantly differentially regulated (DR) genes, i.e., virtual-KO perturbed genes.

A more detailed description of scTenifoldKnk modules is in the section 6.2.

6.3.2 Virtual knockout experiments using simulated scRNA-seq data

We first used the simulated data to validate the relevance of our method. We generated a synthetic scRNA-seq data set using the simulator SERGIO [191]. The generated data produces a sparse matrix (70% zeros) of $3,000$ cells and 100 genes. To simulate the data, we supplied SERGIO with a predefined GRN with five modules of different sizes (containing 5, 10, 25, 40, and 20

genes, respectively) representing modules of functionally related genes under the same regulation. As expected, genes in the simulated data were clustered using the output of the PC regression algorithm into five distinct modules (Figure 6.2 A.), mirroring the predefined modules given to the generative model of SERGIO. We regarded this simulated data as the WT data. We applied the virtual KO method of scTenifoldKnk to artificially delete the 20th gene, or gene #20, in the list. This gene is one of 25 genes in the third module. The removal of the gene produced the pseudo-KO data set. We then used scTenifoldKnk to compare the pseudo-KO with the WT data to identify genes significantly differently regulated before and after the KO. These genes were predicted likely to be perturbed. Because we knew the KO effect was due to the deletion of gene #20, we expected that the identified genes to be those closely correlated with gene #20. Indeed, as expected, scTenifoldKnk showed all significant genes were from the third module (Figure 6.2 B. 6.2 C., top), in which gene #20 is located. We repeated the analysis using genes #50 and #100 as another two examples. Again, the results were as expected (Figure 6.2 B. 6.2 C., middle and bottom). Thus, we concluded that when a member gene is knocked out from a tightly regulated module, othe member genes in the same module should be detected by scTenifoldKnk. Algorithmically, genes in the same module with the KO gene were detected because their projected positions in low-dimensional latent representations of WT and pseudo-KO networks changed more than other genes not in the same module (see Section 6.2 for details).

To further increase our approach's specificity, we included a lambda parameter in our network construction algorithm that maximizes the constructed gene regulatory networks' directionality. This parameter allows us to move from a close to the symmetric matrix like the one shown in Fig 6.2 A. to one totally directed as the one shown in Supplementary Figure S1A. The lambda parameter ranges from 0 to 1 and has a multiplicative $(1 - \lambda * weight)$ effect over the edges' weaker weight connecting two genes. When lambda is equal to 1, only the strongest edge with the largest weight is kept in the network. We applied the same approach described above to knock out the gene 20, gene 50, and 100, respectively, in the totally directed adjacency matrix. Our method's specificity remains stable in all the cases (Supplementary Figure S1B). The identified

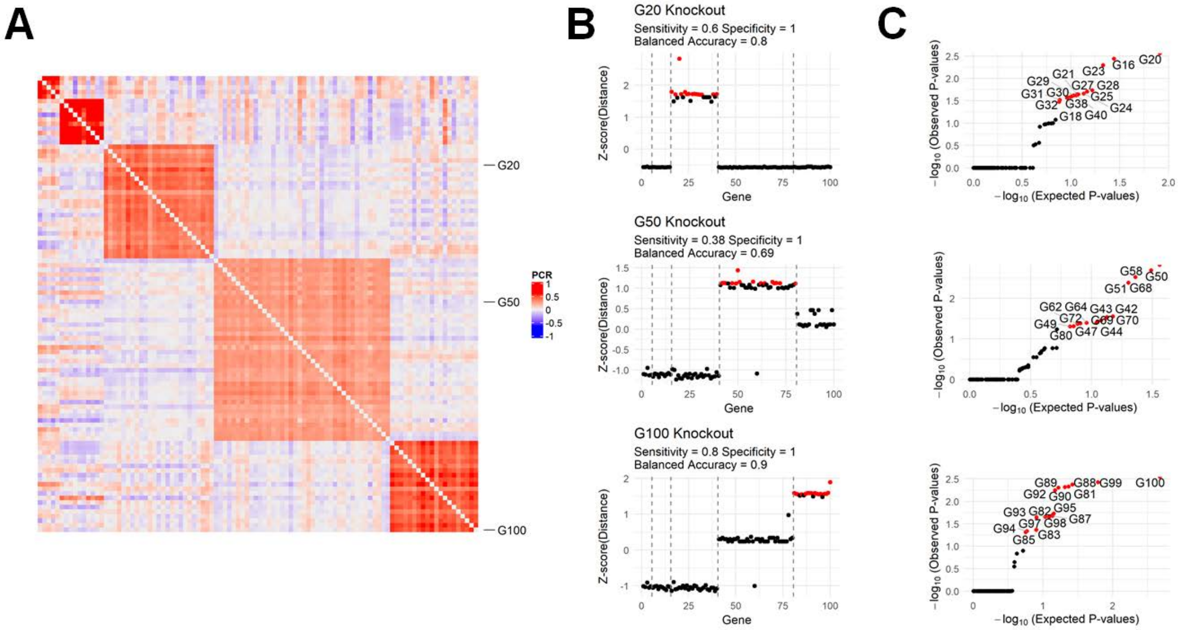


Figure 6.2: Simulations show that scTenifoldKnk specifically detects regulatory modules that include the KO gene. (A) Heatmap of a 100×100 adjacency matrix of scGRN constructed from a simulated scRNA-seq data of 100 genes and $3,000$ cells. The color is scaled according to the normalized PC regression coefficient values between genes. The network contains 5 predefined co-regulated modules of different sizes, indicated by the blocks of gene pairs with a high correlation. The number of genes of each module is 5, 10, 25, 40, and 20, respectively. (B) Significance of scTenifoldKnk DR test of genes in the simulated network is shown as -log(p-value). Red and black dots indicate whether genes are significant or not after false discovery correction. Assuming genes in the same module should be detected as significance genes, sensitivity is defined as = $\frac{TP}{TP+FN}$, and specificity as = $\frac{TN}{TN+FP}$, where T, P, F, and N stands for true, positive, false and negative, respectively. Balanced accuracy is defined as the average of sensitivity and specificity values. (C) QQ-plots of expected vs. observed p-values of genes, given by the DR tests.

genes predicted to be perturbed belonged to the same cluster from where the knocked-out gene was located in the SERGIO simulation. Additionally to the identified genes, the genes next in the rank of topmost perturbed genes belong to the same cluster, confirming the power of scTenifoldKnk to identify modules of perturbed genes after a gene knockout in directed scenarios.

To cross-validate that the directionality retained in the network after using $\lambda = 1$ follows the real regulatory direction, we tested the inbound and outbound directionality of the transcription factors and their target genes reported in the ENCODE database. Briefly, for each transcription factor and associated target genes, we compared the interquartile range of the weights from the transcription factor to their target genes against the interquartile range in the inverse direction (from targets to the transcription factors) in a gene regulatory network constructed for microglial cells [251]. First, we tested the difference in the distribution of the weights using the Kolmogorov–Smirnov test and then cross-validated the difference in the weights using a paired t-test (Supplementary Figure S1C). Both tests agreed that the outbound direction between transcription factors and their target genes is favored by our approach using $\lambda = 1$, confirming the power of scTenifoldKnk to provide the right directionality of the gene regulatory networks.

6.3.3 scTenifoldKnk virtual KO analysis recapitulates results of real KO experiments

As a virtual KO tool, scTenifoldKnk is expected to recapitulate results obtained using real KO experimental data. To prove this, we applied scTenifoldKnk to three scRNA-seq data sets from three real KO experiments. In all three cases, the scRNA-seq data sets of the original studies contained expression matrices from both WT and KO samples. When applying scTenifoldKnk to each of these data sets, we did not use the data from the KO samples, because the virtual KO does not need it. Instead, we used scTenifoldKnk to generate pseudo-KO scGRN, which was used as a surrogate to replace the scGRN generated from the real KO data.

KO gene	**Cell type**	**Main results from original and related studies**	**Ref.**	**scTenifoldKnk predicted results**	**Figure and table in this article**
Real-animal KO experiments with published data					
Nkx2-1	Pulmonary alveolar cells	Decreased expression of marker genes of alveolar type I (AT1) and II (AT2) cells Increased expression of marker genes of gastrointestinal cells *Nkx2-1* regulates expression of genes related to membrane composition, extracellular matrix, and cytoskeleton Mutations in *Nkx2-1* interrupts AT2 cell function and identity	[212]	Marker genes of AT1 and AT2 cells; Intestinal microvillus; Cell cycle and cytoskeleton; Epithelial to mesenchymal transition led by the *WNT* signaling pathway members; Surfactant homeostasis; Lamellar body; Cell adhesion molecules	Figure 6.3 A.. Suppl. Table S1 (171 genes)
Trem2	Microglial cells	*Trem2* regulates expression of genes related to lipid transport and catabolism Lipid metabolism Microglial cell damage response Lysosome and phagosome function Alzheimer's disease Oxidative phosphorylation *Trem2* interacts with signaling transducer *Hcst* and adaptor *Tyrobp*	[251]	Alzheimer's disease; Oxidative phosphorylation; Lysosome; *TYROBP* causal network; Microglia pathogen phagocytosis pathway;	Figure 6.3 B. Suppl. Table S2 (128 genes)
Hnf4a and *Hnf4g*	Intestinal villus epithelial cells	Increased expression of Goblet cell enriched genes Genes in *BMP/SMAD* signaling pathway Decreased expression of enterocyte-enriched genes Genes involved in lipid metabolism Microvillus and absorption Genes related to cytoplasm	[252]	Enterocyte marker genes Electron transport chain Fat digestion and absorption Cholesterol metabolism Chylomicron assembly Cytoplasmic vesicle lumen	Figure 6.3 C. Suppl. Table S3 (65 genes)
Mendelian diseases					
Cftr	Pulmonary alveolar cells	Expressed in epithelial cells of many organs Mutations disrupt the function of the chloride channels	[253]	ABC transporter disorders and surfactant metabolism Ion transmembrane transporter activity Abnormal surfactant secretion and alveolus morphology	Figure 6.4 A. Suppl. Table S4 (17 genes)
Dmd	Skeletal myocytes	Mutation disrupts the linkage between the cytoskeleton and the glycoproteins of the extracellular matrix Impairment of muscle contraction Muscle cell necrosis	[254]	Beta-1 integrin cell surface interaction Contractile actin filament bundle Actomyosin, extracellular matrix receptor interaction Abnormal skeletal muscle morphology	Figure 6.4 B. Suppl. Table S4 (190 genes)
Mecp2	Neurons	Repressing the transcription factor *REST* Altered synapses and synaptic vesicle proteins Defects of GABAergic synapses Syntaxin-1 mutant phenotype	[255]	Affected *REST* target genes GABA synthesis, release, reuptake and degradation Syntaxin binding Synaptic vesicle cycle	Suppl. Figure S2 Suppl. Table S5 (211 genes)
Real-animal KO experiments (data generated in this study)					
Ahr	Enterocytes	Unreliable barrier and uncontrolled proliferation cells Enterocyte marker genes		Myc active pathway; Nuclear receptors meta-pathway.	Figure 6.5 Suppl. Table S6 (53 genes)
Malat1	Pancreatic beta cells	Adenocarcinoma associated genes Reactive Oxygen Species (ROS) Cell cycle regulation Hyperglycemia		Cori cycle; Glycolysis and gluconeogenesis; Regulation of insulin secretion, Cellular carcinomas; HIF signaling pathway	Figure 6.6 Suppl. Table S7 (167 genes)

Table 6.1: Summary of real-data applications of scTenifoldKnk analysis.

6.3.3.1 Nkx2-1 is required for the transcriptional control, development, and maintenance of alveolar type-1 and type-2 cells

NK homeobox 2-1 (*Nkx2-1*) is a transcription factor expressed in lung epithelial cells of alveolar type I (AT1) and type II (AT2). AT1 cells cover 95% of the surface of gas exchange and are 0.1μm thick to allow passive oxygen diffusion into the bloodstream. *Nkx2-1* is essential at all developmental stages of AT1 cells. Loss of *Nkx2-1* results in the impairment of three main defining features of AT1 cells, molecular markers, expansive morphology, and cellular quiescence [212]. AT2 cells are cuboidal and secrete surfactants to reduce surface tension. Mutations in *Nkx2-1* interrupt the expression of *Sftpb* and *Sftpc*, two genes related to AT2 cell function and molecular identity [212, 256].

To examine the molecular and cellular changes underlying the mutant phenotypes in mouse lungs caused by the *Nkx2-1*$^{-/-}$ KO, Little, Gerner-Mauro [212] generated a comprehensive set of data using the lung samples from WT and *Nkx2-1*$^{-/-}$ KO mice. Using bulk RNA-seq and immunostaining assays, they observed that the expression of marker genes of AT1 and AT2 cells were downregulated in the *Nkx2-1* mutant cells. They also found that the expression of marker genes for gastrointestinal cells was upregulated in *Nkx2-1*$^{-/-}$ mutant AT1 cells, which form dense microvilli-like structures apically. Using ChIP-seq, they found that *Nkx2-1* binds to a set of genes implicated in regulating the cytoskeleton, membrane composition, and extracellular matrix. Little *et al.* [212] also generated scRNA-seq data for $2,312$ and $2,558$ epithelial cells from lung samples of the WT and *Nkx2-1*$^{-/-}$ KO mice, respectively.

We obtained the scRNA-seq data, generated by Little *et al.* [212], and used the expression matrix of $8,647$ genes $\times$ $2,312$ cells from WT mice as the input for scTenifoldKnk. We constructed the WT scGRN and then knocked out *Nkx2-1*. The final report of scTenifoldKnk analysis contained 171 significant genes (FDR $<$ 0.05, Supplementary Table S1). These virtual-KO perturbed genes included 7 marker genes of AT1 cells (*Egfl6*, *Ager*, *Cldn18*, *Icam1*, *Crlf1*, *Gprc5a*, and *Aqp5*) and 25 marker genes of AT2 cells (highlighted in Supplementary Table S1). Enrichr functional enrichment test [123] indicated that these genes were enriched for the following functional categories: *epithelial to mesenchymal transition led by the WNT signaling pathway members*, *surfactant homeostasis*, *lamellar body*, and *cell adhesion molecules*. These enriched function annotations were known to be related to the function of AT2 cells. Next, we applied the interaction enrichment analysis to the 171 significant genes. The interaction enrichment analysis was based on the STRING database [245]. We found that these significant genes appeared in a fully connected component in the STRING interaction network, which is unexpected by chance (P < 0.01, STRING interaction enrichment test), indicating a closely related functional relationship between those genes. We subsequently performed the gene set enrichment analysis (GSEA [178]) to evaluate the extent of perturbation caused by the *Nkx2-1* KO at the transcriptome-wide level. The GSEA analysis identified gene sets containing marker genes of AT1 and AT2 cells (FDR < 0.01 in both cases). That is to say, AT1 and AT2 marker genes were among the topmost perturbed genes, caused by the deletion of *Nkx2-1*. GSEA analysis also showed that the *Nkx2-1*$^{-/-}$ KO impacted genes with functions related to *intestinal microvillus* (Figure 6.3 A.), *cell cycle*, and *cytoskeleton*. These results are consistent with those reported in the original study [212]. Again, we emphasize that scTenifoldKnk did not use data generated from the KO mice. Instead, all results of scTenifoldKnk were derived *in-silico* using data generated from the WT mice. Overall, the results of our scTenifoldKnk analysis are consistent with those of the original study [212], in which scRNA-seq data was experimentally derived from KO mice and combined with other experimental techniques to characterize the KO individual.

6.3.3.2 Trem2 regulates microglial cholesterol metabolism

The Triggering Receptor Expressed on Myeloid cells 2 (*Trem2*) is a single-pass transmembrane immune receptor, selectively expressed in microglia within the central nervous system. *Trem2* is known to be involved in late-onset Alzheimer's disease, and plays a role in *modulating proliferation*, *survival*, *immune response*, *calcium mobilization*, *cytoskeletal dynamics*, *mTOR signaling*, *autophagy*, and *energy metabolism* [257]. The function of *Trem2* is known to be mediated via signaling transducer *Hcst* and adaptor *Tyrobp* [258]. *Trem2* is also known to play a role in regulating *lipid metabolism*, with most studies focusing on lipids in the form of either lipoprotein particles

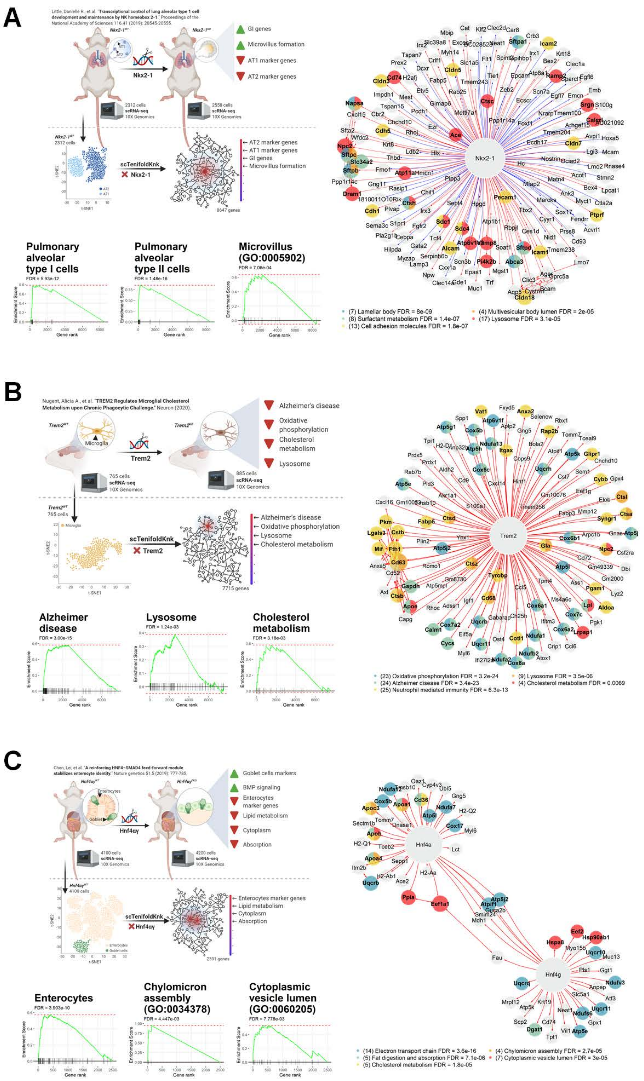

A
Little, Danielle R., et al. "Transcriptional control of lung alveolar type 1 cell development and maintenance by NK homeobox 2-1." Proceedings of the National Academy of Sciences 116.41 (2019): 20545-20555.
Nkx2-1WT
AT1
AT2
Nkx2-1
Nkx2-1KO
GI genes
Microvillus formation
AT1 marker genes
AT2 marker genes
2312 cells
scRNA-seq
10X Genomics
2558 cells
scRNA-seq
10X Genomics
Nkx2-1WT
2312 cells
t-SNE2
t-SNE1
AT2
AT1
scTenifoldKnk
Nkx2-1
← AT2 marker genes
← AT1 marker genes
← GI genes
← Microvillus formation
8647 genes
Pulmonary alveolar type I cells
FDR = 5.93e-12
Pulmonary alveolar type II cells
FDR = 1.48e-16
Microvillus (GO:0005902)
FDR = 7.06e-04
Enrichment Score
Gene rank
Nkx2-1
(7) Lamellar body FDR = 8e-09
(8) Surfactant metabolism FDR = 1.4e-07
(13) Cell adhesion molecules FDR = 1.8e-07
(4) Multivesicular body lumen FDR = 2e-05
(17) Lysosome FDR = 3.1e-05
B
Nugent, Alicia A., et al. "TREM2 Regulates Microglial Cholesterol Metabolism upon Chronic Phagocytic Challenge." Neuron (2020).
Trem2WT
Microglia
Trem2
Trem2KO
Alzheimer's disease
Oxidative phosphorylation
Cholesterol metabolism
Lysosome
765 cells
scRNA-seq
10X Genomics
885 cells
scRNA-seq
10X Genomics
Trem2WT
765 cells
Microglia
t-SNE2
t-SNE1
scTenifoldKnk
Trem2
← Alzheimer's disease
← Oxidative phosphorylation
← Lysosome
← Cholesterol metabolism
7715 genes
Alzheimer disease
FDR = 3.00e-15
Lysosome
FDR = 1.24e-03
Cholesterol metabolism
FDR = 3.18e-03
Enrichment Score
Gene rank
Trem2
(23) Oxidative phosphorylation FDR = 3.2e-24
(24) Alzheimer disease FDR = 3.4e-23
(25) Neutrophil mediated immunity FDR = 6.3e-13
(9) Lysosome FDR = 3.5e-06
(4) Cholesterol metabolism FDR = 0.0069
C
Chen, Lei, et al. "A reinforcing HNF4–SMAD4 feed-forward module stabilizes enterocyte identity." Nature genetics 51.5 (2019): 777-785.
Hnf4αγWT
Enterocytes
Goblet
Hnf4αγ
Hnf4αγDKO
Goblet cells markers
BMP signaling
Enterocytes marker genes
Lipid metabolism
Cytoplasm
Absorption
4100 cells
scRNA-seq
10X Genomics
4200 cells
scRNA-seq
10X Genomics
Hnf4αγWT
4100 cells
t-SNE2
t-SNE1
Enterocytes
Goblet cells
scTenifoldKnk
Hnf4αγ
← Enterocytes marker genes
← Lipid metabolism
← Cytoplasm
← Absorption
2591 genes
Enterocytes
FDR = 3.903e-10
Chylomicron assembly (GO:0034378)
FDR = 4.447e-03
Cytoplasmic vesicle lumen (GO:0060205)
FDR = 7.778e-03
Enrichment Score
Gene rank
Hnf4a
Hnf4g
(14) Electron transport chain FDR = 3.6e-16
(5) Fat digestion and absorption FDR = 7.1e-06
(5) Cholesterol metabolism FDR = 1.8e-05
(4) Chylomicron assembly FDR = 2.7e-05
(7) Cytoplasmic vesicle lumen FDR = 3e-05

Figure 6.3: scTenifoldKnk virtual KO analysis recapitulates the findings of real KO experiments. (A) Virtual KO of *Nkx2-1* identifies gene expression program changes reported in the original study. GSEA analysis identifies significant gene sets including pulmonary alveolar type I cells, pulmonary alveolar type II cells, and microvillus (GO:0005920 under cellular component actin-based cell projection). Egocentric plot shows virtual-KO perturbed genes connected to *Nkx2-1*. Nodes are color-coded by each gene's membership association with significantly enriched functional groups. (B) Virtual KO of *Trem2* identifies gene expression program changes reported in the original study. GSEA analysis identifies significant gene sets including Alzheimer's disease, lysosome, and cholesterol metabolism. Egocentric plot shows virtual-KO perturbed genes connected to Trem2. Nodes are color-coded by each gene's membership association with significantly enriched functional groups. (C) Virtual double KO of *Hnf4a* and *Hnf4g* identifies gene expression program changes reported in the original study. GSEA analysis identifies significant gene sets including enterocytes, chylomicron assembly, and cytoplasmic vesicle lumen. Egocentric plot shows virtual-KO perturbed genes connected to *Hnf4a* and *Hnf4g*. Nodes are color-coded by each gene's membership association with significantly enriched functional groups.

or cell surface-exposed signals, such as candidate *Trem2* ligands [259]. By comparing between WT and *Trem2*$^{˘/˘}$ KO mice, Poliani *et al.* [260] showed that *Trem2* regulates many genes, such as *Apoe* and *Lpl*, which control lipid transport and catabolism in microglia. *Trem2* was also found to modulate gene expression of macrophages in adipose and control blood cholesterol metabolism in obese mice [261], further linking the function of *Trem2* to lipid metabolism. To examine whether *Trem2* mediates myelin lipid processing in microglia, Nugent *et al.*[251] isolated and characterized *Cd11b*$^{+}$/*Cd45*^{low} microglial cells from *Trem2*$^{+/+}$, *Trem2*$^{+/˘}$, and *Trem2*$^{˘/˘}$ mice, fed with a 0.2% demyelinating cuprizone diet for 12 weeks. They analyzed a comprehensive set of analytical data using FACS, RNA-seq, scRNA-seq, and lipidomics. They reported that *Trem2* upregulates *Apoe* and other genes involved in *cholesterol transport and metabolism*, causing robust intracellular accumulation of a storage form of cholesterol upon chronic phagocytic challenge. *Trem2* was also shown to regulate the expression of genes associated with *cell damage response*, *lysosome* and *phagosome function*, *Alzheimer's disease*, and *oxidative phosphorylation* [251].

To perform the virtual KO analysis, we obtained scRNA-seq data from WT (*Trem2*$^{+/+}$) mice [251]. The expression matrix contained data of $7,715$ genes and 765 *Cd11b*$^{+}$/*Cd45*^{low} microglial cells. We used this WT expression matrix as the input and used scTenifoldKnk to knock out *Trem2*. The final results of scTenifoldKnk analysis contained 128 virtual-KO perturbed genes (FDR < 0.05, Supplementary Table S2). The Enrichr analysis [123] showed that these genes were enriched with: *Alzheimer's disease*, *oxidative phosphorylation*, *lysosome*, *TYROBP causal network*, *metabolic pathway of LDL, HDL and TG*, and *microglia pathogen phagocytosis pathway*. Such an enrichment indicates that the proteins were at least partially biologically connected as a group. These virtual-KO perturbed genes were highly interactive with each other, as shown by their positions on the STRING interaction network. The network of virtual-KO perturbed genes had significantly more interactions than expected (P < 0.01, STRING interaction enrichment test), which means that gene products exhibited more interactions among themselves than what would be expected for a random set of proteins of similar size, drawn from the genome. This result suggests that the identified, virtual-KO perturbed genes are closely related with shared functions.

Collectively, these scTenifoldKnk findings provide insight into understanding *Trem2* function by revealing the list of genes perturbed following *Trem2* deletion (Figure 6.3 B.). We show that the scTenifoldKnk results recapitulate those reported in the original study [251], demonstrating the utility of scTenifoldKnk.

6.3.3.3 Hnf4a and Hnf4g stabilize enterocyte identity

Hepatocyte nuclear factor 4 alpha and gamma, *Hnf4a* and *Hnf4g*, are transcription factors which regulate gene expression in the gut epithelium. *Hnf4a* and *Hnf4g* function redundantly, and thus, an independent deletion of one paralog causes no gross abnormalities [252, 262]. *Hnf4a* and *Hnf4g* double-KO *Hnf4ag*^{DKO} mice exhibit fluid-filled intestines indicative of intestinal malfunction [252]. Epithelial cells in the *Hnf4ag*^{DKO} mutants fail to differentiate. Using the bulk RNA-seq data, Chen *et al.* [252] compared gene expression in duodenal epithelial cells isolated from WT mice and *Hnf4a*^{KO}, *Hnf4g*^{KO} and *Hnf4ag*^{DKO} mutants. They identified $2,892$ DE genes in the *Hnf4ag*^{DKO} mutant but only 560 and 77 in the *Hnf4a*^{KO} and *Hnf4g*^{KO} mutants, respectively (FDR < 0.05, absolute log2(fold change) > 1). The DE genes identified in the *Hnf4ag*^{DKO} enterocytes were enriched for functions in digestive metabolisms such as *lipid metabolism*, *microvillus*, and *absorption*, as well as *enterocyte morphology*, *cytoplasm*, *Golgi apparatus*, and *immune signaling*. *Hnf4ag*^{DKO} epithelium exhibited a robust shift in the transcriptome away from differentiated cells toward proliferating and Goblet cells, suggesting that *Hnf4ag*^{DKO} impair enterocyte differentiation and destabilize enterocyte identity. To validate their findings, Chen *et al.* [252] used scRNA-seq to measure gene expression in intestinal villus epithelial cells. They obtained scRNA-seq data for $4,100$ and $4,200$ cells from *Hnf4ag*^{WT} and *Hnf4ag*^{DKO} respectively, and confirmed that, compared to the WT, mutant epithelial cells show increased Goblet cell-enriched genes, such as *Agr2*, *Spink4*, *Gcnt3* and *S100a6*, and decreased expression of enterocyte-enriched genes, such as *Npc1l1*, *Apoc3*, *Slc6a19* and *Lct* (Figure 6.3 C. left panel). The *Hnf4ag*^{DKO} mutant cells also showed increased expression for genes in the *BMP/SMAD* signaling pathway and decreased expression of genes involved in lipid metabolism, microvillus and absorption, and genes related to cytoplasm [252].

We obtained the scRNA-seq expression matrix of $4,100$ cells from the *Hnf4ag*^{WT} samples used as input for scTenifoldKnk. We constructed the WT scGRN of $2,591$ genes and then virtually knocked out both *Hnf4a* and *Hnf4g* genes at the same time. The final result of scTenifoldKnk contained 65 virtual-KO perturbed genes (FDR < 0.05, Supplementary Table S3). These genes were enriched with *electron transport chain*, *fat digestion and absorption*, *cholesterol metabolism*, *chylomicron assembly*, and *cytoplasmic vesicle lumen*. A search of STRING database [245] indicated that all virtual-KO perturbed genes form a fully connected network module. STRING is a database of known and predicted protein-protein interactions. The interaction enrichment test, a statistical test provided by the STRING database, indicated that such a full interconnection of 65 genes is less likely to be expected by chance (P < 0.01). Furthermore, GSEA analysis revealed that these virtual-KO perturbed genes were enriched with canonical marker genes of enterocytes (Figure 6.3 C.), which is consistent with the finding of the original study [252].

6.3.4 scTenifoldKnk reveals functions of Mendelian disease genes

Mendelian diseases are a family of diseases caused by the loss or malfunctioning of a single gene. For many Mendelian diseases, we know a great deal about their genetic basis and pathophysiological phenotypes [263]. Thus, we decided to use three Mendelian diseases as ‘positive controls’ to test the performance of scTenifoldKnk. We attempted to determine whether scTenifoldKnk can accurately predict the molecular phenotypic consequences of gene deletion. The three chosen Mendelian diseases were Cystic fibrosis, Duchenne muscular dystrophy, and Rett syndrome. As described in more detail below, in each case, we performed scTenifoldKnk analysis using existing scRNA-seq data generated from cell types that are most relevant to the disease conditions (Table 6.1).

6.3.4.1 Cystic fibrosis

Cystic fibrosis (CF) is one of the most common autosomal recessive diseases [264]. It is caused by mutations in *CFTR*, a gene encoding for a transmembrane conductance regulator [265, 266], which functions as a channel across the membrane of cells that produce mucus, sweat, saliva, tears, and digestive enzymes. The *CFTR* protein also regulates the function of other channels. *CFTR* is expressed in epithelial cells of many organs, including lung, liver, pancreas, and digestive tract [267]. The most common *CFTR* mutation that causes CF is the deletion of phenylalanine 508 (ΔF508), which disrupts the function of the chloride channels, preventing patients from regulating the flow of chloride ions and water across cell membranes [265]. The truncated *CFTR* protein leads to a reduction of surfactant, and causes the build-up of sticky, thick mucus that clog the airways, increasing the risk of bacterial infections and permanent lung damage [268].

To test scTenifoldKnk, we obtained scRNA-seq data from $7,326$ pulmonary alveolar type II (AT2) cells in the GEO database (access number: GSM3560282). The original data sets were generated by Frank *et al.* [253] to study the lineage-specific development of alveolar epithelial cells in mice. The original study was not directly focused on CF. Nevertheless, with the downloaded data, we constructed a WT scGRN that contained $7,107$ genes. We then used scTenifoldKnk to knock out *Cftr*. The final results of scTenifoldKnk contained 17 virtual-KO perturbed genes: *Cftr*, *Birc5*, *Cldn10*, *Cxcl15*, *Dcxr*, *Hmgb2*, *Lamp3*, *Mgst1*, *Npc2*, *Pclaf*, *Pglyrp1*, *Sftpa1*, *Sftpb*, *Sftpc*, *Smc2*, *Tspan1* and *Tubb5* (FDR < 0.05, Figure 6.4A, Supplementary Table S4). Among them, *Sftpa1*, *Sftpb*, and *Sftpc* have functions associated with *ABC transporter disorders* and *surfactant metabolism*. *Sftpb* and *Sftpd* encode for surfactant proteins implicated in CF and innate immunity [269, 270]. *Ctsh* encodes for cathepsin H, a cysteine peptidase, which is involved in the processing and secretion of the pulmonary surfactant protein B [271]. GSEA analysis of the ranked gene list was conducted using gene sets of the MGI mammalian phenotypes database as reference. The result showed that the gene list scTenifoldKnk produced was significant in terms of ion transmembrane transporter activity, abnormal surfactant secretion, and alveolus morphology (FDR < 0.01 in all cases, Figure 6.4 A.). These results are consistent with the known pathophysiological changes resulting from the loss of *Cftr* function in the lungs.

6.3.4.2 Duchenne muscular dystrophy

Duchenne muscular dystrophy (DMD) arises as a result of mutations that affect the open reading frame of *DMD* [272, 273]. The *DMD* gene encodes dystrophin, a large cytoskeletal structural

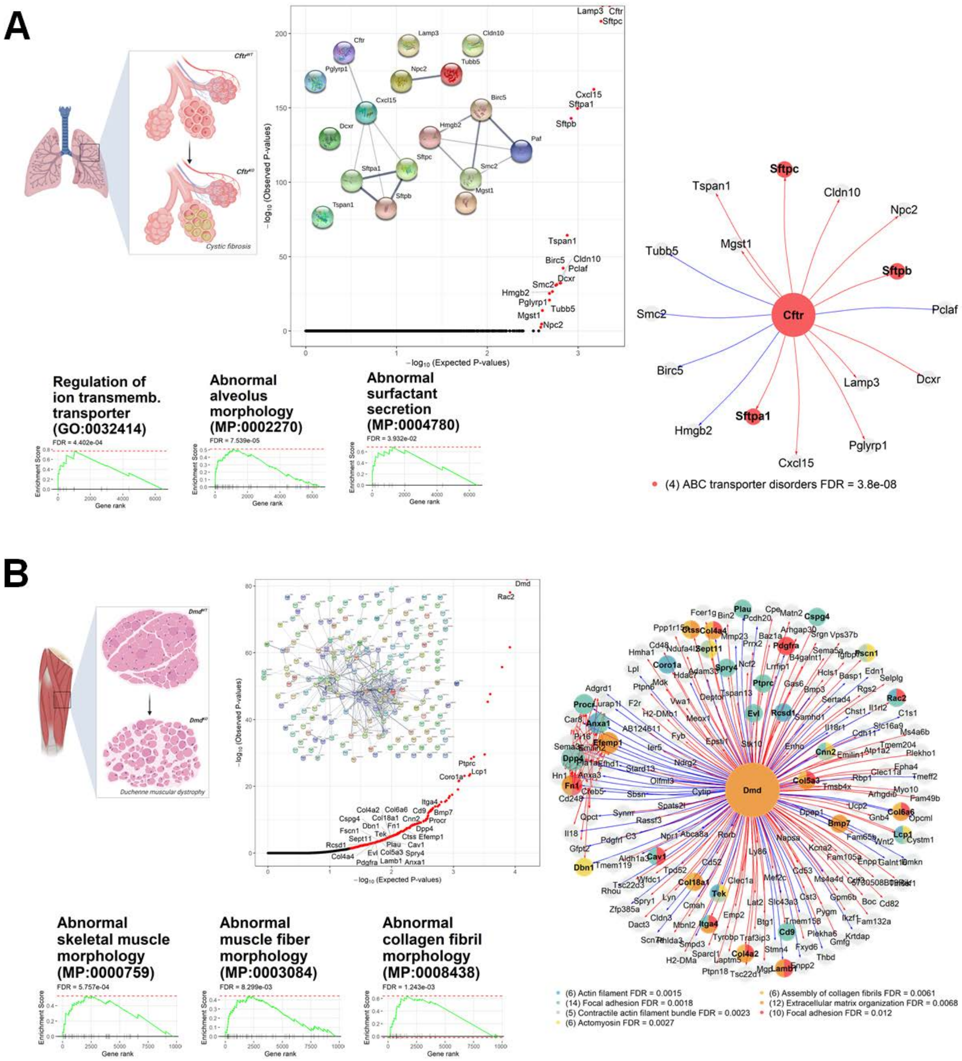


Figure 6.4: scTenifoldKnk virtual KO analysis reveals the function of Mendelian disease genes in relevant cell types. (A) Virtual KO of *Cftr* in pulmonary alveolar type II cells identifies gene expression program changes associated with cystic fibrosis. GSEA analysis identifies significant gene sets including regulation of ion transmembrane transporter, abnormal alveolus morphology, and abnormal surfactant secretion. Egocentric plot shows virtual-KO perturbed genes connected to *Cftr*. Nodes are color-coded by each gene's membership association with significantly enriched functional groups. QQ-plot of genes and interconnection of virtual-KO perturbed genes in STRING are given. (B) Virtual KO of *Dmd* in muscular cells identifies gene expression program changes associated with Duchenne muscular dystrophy. GSEA analysis identifies significant gene sets including abnormal skeletal muscle morphology, and abnormal collagen fibril morphology. Egocentric plot shows virtual-KO perturbed genes connected to Dmd.

protein, which is mostly absent in DMD patients [274]. The absence of dystrophin results in a disturbance of the linkage between the cytoskeleton and the glycoproteins of the extracellular matrix, generating an impairment of muscle contraction, eventually leading to muscle cell necrosis [274, 275] (Figure 6.4 B. left). We obtained the scRNA-seq data of $5,159$ muscle cells from the mouse limb (quadriceps) in the GEO database (Access number: GSM4116571). The original data was generated to study gene expression patterns in skeletal muscle cells [254]. The original study was not focused on DMD. We used scRNA-seq data from normal tissue to construct the WT scGRN of $9,783$ genes. We subsequently performed the scTenifoldKnk virtual KO analysis to predict the molecular phenotype due to the impact of the *Dmd* KO. The final results of scTenifoldKnk included 190 virtual-KO perturbed genes (FDR < 0.05, Supplementary Table S5). These genes were enriched with functions related to *beta-1 integrin cell surface interaction*, *contractile actin filament bundle*, *actomyosin*, *extracellular matrix receptor interaction*, and *extracellular matrix organization* (Figure 6.4 B. middle). The GSEA analysis against the MGI mammalian phenotype database gave the following top hits (FDR < 0.01): *abnormal collagen fibril morphology*, *abnormal skeletal muscle morphology*, and *abnormal skeletal muscle fiber morphology* (Figure 6.4 B. right). These phenotype terms represent consistently with known effects of the loss of *DMD* function in muscle cells, verifying that scTenifoldKnk can predict phenotypic effects caused by gene KO that are pertinent for the biological context.

6.3.4.3 Rett syndrome

The third Mendelian disease we considered was the Rett syndrome (RTT, MIM 312750), which is a severe neurodevelopmental disease [276, 277]. RTT is known to be caused by mutations in *Mecp2*, a transcriptional repressor required for the maintenance of normal neuronal functions [278, 279]. *Mecp2* deficiency in the brain decreases the expression level of genes involved in the *Bdnf* signaling pathway, mediated by repressing the transcription factor *Rest* [280]. We obtained scRNA-seq data generated from mouse neurons (SRA database access numbers: SRX3809326 and SRX3809327) for the mouse brain atlas project [255]. The two data sets contain $2,054$ and $2,156$ neurons respectively, derived from two CD1 P19 female mice that served as biological replicates.

Here, we analyzed the two data sets independently, to see whether scTenifoldKnk could, as expected, produce similar results with data gathered from biological replicates. Two scGRNs containing $8,652$ and $8,555$ genes were constructed first. Then, the scTenifoldKnk analysis of virtual KO of *Mecp2* produced 377 and 322 virtual-KO perturbed genes, respectively (FDR < 0.05, Supplementary Table S6 and S7), including 211 shared genes. The number of shared genes was significantly higher than the random expectation (P $< 1 \times 10^{-5}$, hypergeometric test), indicating a high overlap rate between results of the scTenifoldKnk analysis when applied to the two data sets from biological replicates.

Many of these 211 genes were found to be targets of the transcription factor *Rest* (FDR < 0.01). The functions of these genes were enriched (FDR < 0.01) and include: *axon*, *synaptic vesicle cycle*, *GABA synthesis, release, reuptake and degradation*, *syntaxin binding*, and *transmission across chemical synapses*. These results are consistent with previous experimental results. For example, it is known that the most prominent alterations in gene products due to *Mecp2* KO are related to synapses and synaptic vesicle proteins [281]. At the phenotypic level, the *Mecp2* KO causes early defects in GABAergic synapses [282] and mediates autism-like stereotypies and RTT phenotypes [283]. Mutations in syntaxin-1 are known to illicit phenotypes similar to those found in RTT [284]. In addition to using the overlapped genes to show the reproducibility of scTenifoldKnk analysis, we also compared the two ranked gene lists generated from the two replicate mice. If scTenifoldKnk results are robust, we expected the relative positions of the same genes in the two ranked lists, generated from biological replicates, to show stronger correlation than those in two random lists of genes. Indeed, we found the correlation in the rank of genes between two reported lists was highly significant (P $< 1 \times 10^{-12}$), with a Spearman correlation coefficient of $\rho = 0.68$. Finally, GSEA analyses with both lists of genes showed that *BDNF signaling pathway* was highly significant (FDR < 0.01, for both replicates, Supplementary Figure S2).

Since validating the reproducibility of scTenifoldKnk findings is important even when the available experiments with public biological replicates containing sufficient cells (over 500) for the same cell type are still uncommon. We decided to evaluate the results' stability using ten random subsamplings (bootstrap) of cells within the dataset used to assess the impact of the *Trem2* knockout on microglial cells. For each random subsampling of 500 cells, we constructed an scGRN, performed the virtual knockout of *Trem2* using scTenifoldKnk, and obtained a perturbation profile to test the recovered functional enrichment and the similarity in the ranking of the predicted to be perturbed genes. We found that all the computed perturbation profiles returned by scTenifoldKnk are positively correlated with an average Spearman correlation coefficient of 0.55 ± 0.06 (Supplementary Figure S3A). This result strengthens the power of scTenifoldKnk to identify similar patterns of perturbation, surpassing the differences that the sparsity of the data may induce. We also evaluated the power to detect the functional enrichment of the gene sets reported in the paper across each subsampling. For that purpose, we used single-sample gene set enrichment analysis using the gene set reported by the KEGG database. We found that the gene sets reported associated with *Oxidative phosphorylation*, *Alzheimer's disease*, *Cholesterol metabolism*, *Phagosome*, and *Lysosome* are among the top predicted to be perturbed gene sets in all the cases (Supplementary Figure S3B). This result confirms the stability of the perturbation predictions provided by scTenifoldKnk.

6.3.5 scTenifoldKnk predicts real KO experimental results

In this section, we describe two in vivo KO experiments that we performed. In each of the experiments, we generated scRNA-seq data from both WT and KO mice. The WT data was used to perform scTenifoldKnk virtual KO analysis and predict the consequences of gene KO. The comparative studies between the data sets from WT and KO mice were conducted to validate the prediction made.

6.3.5.1 scTenifoldKnk predicts transcriptional changes in enterocytes of Ahr$^{-/-}$ KO mice

The Aryl hydrocarbon receptor (AhR) is a transcription factor that regulates the gene expression of metabolic enzymes, such as cytochrome P450s. It acts primarily as a sensor for both dietary and gut microbial-derived ligands. These include structurally diverse phytochemicals, pharmaceuticals, and endogenous compounds, including tryptophan metabolites and xenobiotic chemicals.

AhR also plays a role in regulating immunity, stem cell maintenance, and cellular differentiation [285, 247, 286, 287, 288]. To study the role of Ahr in intestinal stem cell fate determination, we generated inducible, stem cell-specific Ahr$^{-/-}$ KO mice (see Section 6.2.7). We then used scRNA-seq to generate data for five WT and five KO mice, and obtained $13,626$ and $8,425$ sequenced cells from WT and KO mice, respectively. Data was pooled and projected to low-dimensional embedding, showing that there was no global shift in overall gene expression profiles between cells from WT and KO mice (Figure 5A). We subsequently clustered cells into groups and identified seven distinct cell types: non-cycling Lgr5+ colonic stem cells, Goblet cells, cycling Lgr5+ colonic stem cells, pre-enterocytes, transit-amplifying cells, enteroendocrine cells, and Tuft cells (Figure 5B). Since enterocytes are the most abundant epithelial cell lineage in the intestine, we decided to conduct scTenifoldKnk using data from enterocytes (indicated by dashed oval in Figure 6.5 B.) to assess the function of Ahr.

We conducted scTenifoldKnk analysis to knock out *Ahr* from the scGRN of WT mice. The scGRN was constructed from an input expression matrix of $1,029$ cells (i.e., WT enterocytes) and $8,478$ genes. The output of scTenifoldKnk was a ranked gene list including 53 top genes, i.e., virtual-KO perturbed genes (FDR < 0.05, Supplementary Table S8). Enrichr functional enrichment tests indicated that these genes were associated with *Myc active pathway*, *nuclear receptors meta-pathway*, *MHC class II protein complex binding*, *Chaperone-mediated protein complex assembly*, *response to unfolded protein*, *RNA binding*, *translation elongation factor activity*, *cytoplasmic vesicle lumen*, *T cell receptor regulation of apoptosis*, and *secretory granule lumen* (Figure 6.5 C.). Next, a STRING database [245] search indicated that identified virtual-KO perturbed genes form two network modules (both $P < 0.01$, STRING interaction enrichment tests, Supplementary Figure S4). Gene members in each module are fully connected, further confirming closely related functional associations among identified genes. Finally, GSEA analysis against a comprehensive collection of gene sets (see Section 6.2) identified 60 non-redundant functional gene sets (Supplementary Table S9), including many related to *cell cycle* and *cell proliferation* signaling pathways.

GSEA analysis against the canonical marker gene sets (retrieved from PanglaoDB, see Section

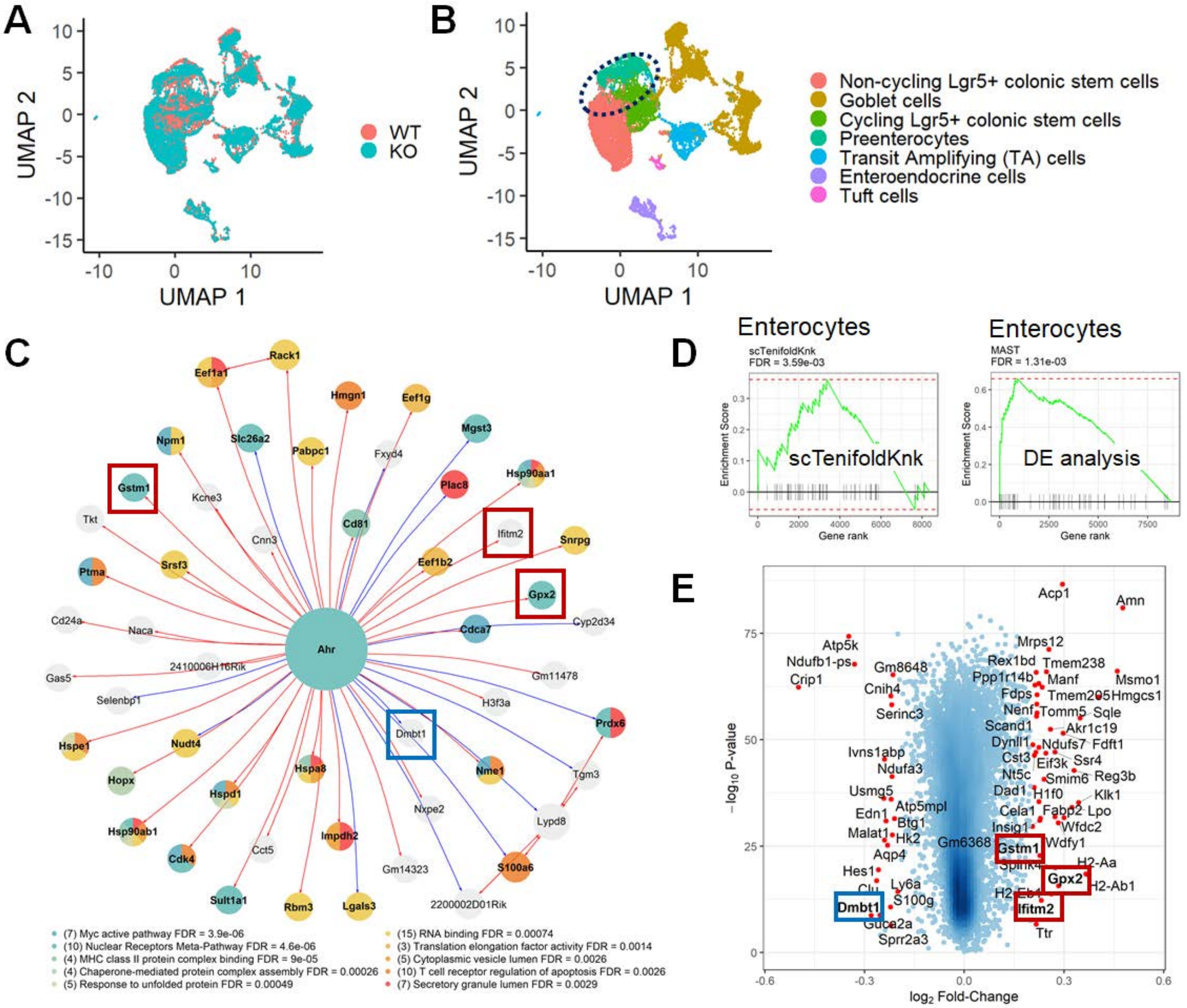


Figure 6.5: scTenifoldKnk predicts transcriptional changes in the enterocytes of *Ahr*$^{-/-}$ KO mice (A) UMAP plot of cells distinguished by experimental groups: WT or KO. (B) Cell clustering and cell type assignment. Enterocytes included in the analysis are highlighted with dashed oval. (C) Egocentric plot showing virtual-KO perturbed genes connected to *Ahr*. Nodes are color-coded by each gene's membership association with the top 10 significantly enriched functional groups. Proliferative genes (associated with the *Myc* pathway) were identified as the expected topmost perturbed genes after the deletion of *Ahr* in enterocytes. (D) GSEA analyses with the DE gene list and scTenifoldKnk gene list produce the same result – gene sets of enterocytes. (E) Volcano plot from the DE analysis comparing gene expression between WT and KO samples. In the plot, -log2(FC) and -log10(P-value) of genes are plotted against each other. Names of genes are given for those with a fold change (FC) > 10% and FDR < 0.05. Among them, 4 virtual-KO perturbed genes are highlighted with a red (*Gstm1*, *Gpx2* and *Ifitm2*) or a blue (*Dmbt1*) box, indicating up-regulated or down-regulated in the KO sample.

6.2) showed that enterocytes was the most significant hit (Figure 6.5 D., left), suggesting that *Ahr* KO induces a change in the expression of enterocyte marker genes, which may be responsible for the alteration of enterocyte differentiation fate [289]. Taken together, our comprehensive functional annotation for the virtual-KO perturbed genes revealed the multifactorial functions of *Ahr* as a key player of enterocyte cell differentiation and proliferation. These results are consistent with the known functions of *Ahr*. For example, a previous study showed that *Ahr* deletion in intestinal epithelial cells resulted in an abnormal barrier function and uncontrolled proliferation of intestinal cells, promoting malignant transformation [285]. Organoids with dysregulated *Ahr* had substantially compromised differentiation to goblet and enterocyte cell fate [289]. Intestinal-specific *Ahr* KO increases basal stem cell and crypt injury-induced cell proliferation and promotes colon tumorigenesis [247]. *Myc* is a key transcriptional effector of the *Wnt* signaling pathway through its actions to promote cellular proliferation [290].

Without scTenifoldKnk, we would have to use differential expression (DE) analysis. Indeed, DE analysis has been a dominant method for assessing the impact of gene KO (e.g., as in [251, 212, 252]). Here, we put ourselves in such a hypothetical scenario – that is, scTenifoldKnk is not available. We resorted to comparing gene expression levels between the WT and Ahr KO mice, using an established scRNA-seq DE analysis tool, MAST [89]. A total of 68 DE genes were identified (FDR < 0.05 and log(FC) > 0.25), including 46 upregulated and 22 downregulated genes in the *Ahr* KO enterocytes (Figure 6.5 E.). Three up-regulated DE genes, *Gstm1*, *Gpx2*, and *Ifitm2*, and one down-regulated DE genes, *Dmb1t*, are also among the list of 53 virtual-KO perturbed genes identified by scTenifoldKnk (highlighted with red and blue boxes, respectively, in Figure 6.5 C. 6.5 E.). Note that, in the WT scGRN, the sign of edges from *Ahr* to the three up-regulated genes is positive (shown as red edges in Figure 6.5 C.), while the sign of edge from *Ahr* to the down-regulated *Dmb1t* is negative (shown using the blue edge between *Ahr* and *Dmb1t* in Figure 6.5C). The GSEA analysis of the DE gene list, with the comprehensive gene sets as the references (Section 6.2), reported 35 non-redundant significant gene sets, with 16 overlaps with the 60 non-redundant significant gene sets in Supplementary Table S9, which were identified using scTenifoldKnk ranked gene list. The overlap is significantly larger than expected by chance (P $< 10\times^{-20}$, hypergeometric test, assuming $n = 109$ of non-redundant gene sets). In addition, GSEA analysis of the DE gene list with marker gene sets as references showed that enterocytes was one of the significant hits (Figure 6.5 D., right), which is the same as the result of GSEA analysis of virtual-KO perturbed genes (Figure 6.5 D., left). In summary, we obtained overall similar results using scTenifoldKnk and DE analysis. The difference is that scTenifoldKnk is a virtual KO analysis, whereas DE analysis requires data from the real KO experiment.

6.3.5.2 *scTenifoldKnk predicts transcriptional changes in pancreatic beta cells of Malat1$^{-/-}$ KO mice*

Malat1 is a long, non-coding RNA gene that acts as a transcriptional regulator for numerous genes, which are involved in the generation of reactive oxygen species, cell cycle regulation, cancer metastasis, and cell migration [291, 292, 293, 294]. *Malat1* upregulation is also associated with hyperglycemia [295]. We have previously shown that the ablation of *Malat1* changes insulin responses in vivo by suppressing *Jnk* activity and inducing the activation of the insulin receptor substance 1 (*IRS-1*) [294]. The ablation of *Malat1* also induces the phosphorylation of *Akt* in a process mediated by increased reactive oxygen species (ROS) [294].

To explore the regulatory role of *Malat1* in pancreatic cell functions, we generated *Malat1*$^{-/-}$ KO mice (see Section 6.2.8). We performed scRNA-seq experiments with isolated pancreatic cells from WT (C57BL/6) and *Malat1*$^{-/-}$ KO mice (Figure 6.6 A.). The beta cells were the most abundant cells in both samples, $1,142$ WT and $1,695$ KO (Figure 6.6 B.). We used scTenifoldKnk to knock out *Malat1* over the WT scGRN, and identified 167 virtual-KO perturbed genes (FDR $<$ 0.05, Supplementary Table S10). Enrichr functional enrichment analysis indicated that these genes were significantly associated with glucose metabolisms functions such as *Cori cycle*, *glycolysis* and *gluconeogenesis*, and the *regulation of insulin secretion*, as well as with pathways related to *cellular carcinomas* (Figure 6.6 C.). This result is consistent with the known function of *Malat1*. For example, *Malat1*, which stands for Metastasis Associated Lung Adenocarcinoma Transcript 1, is known to be overexpressed in lung carcinoma. In addition, *Malat1* is known to have pleiotropic

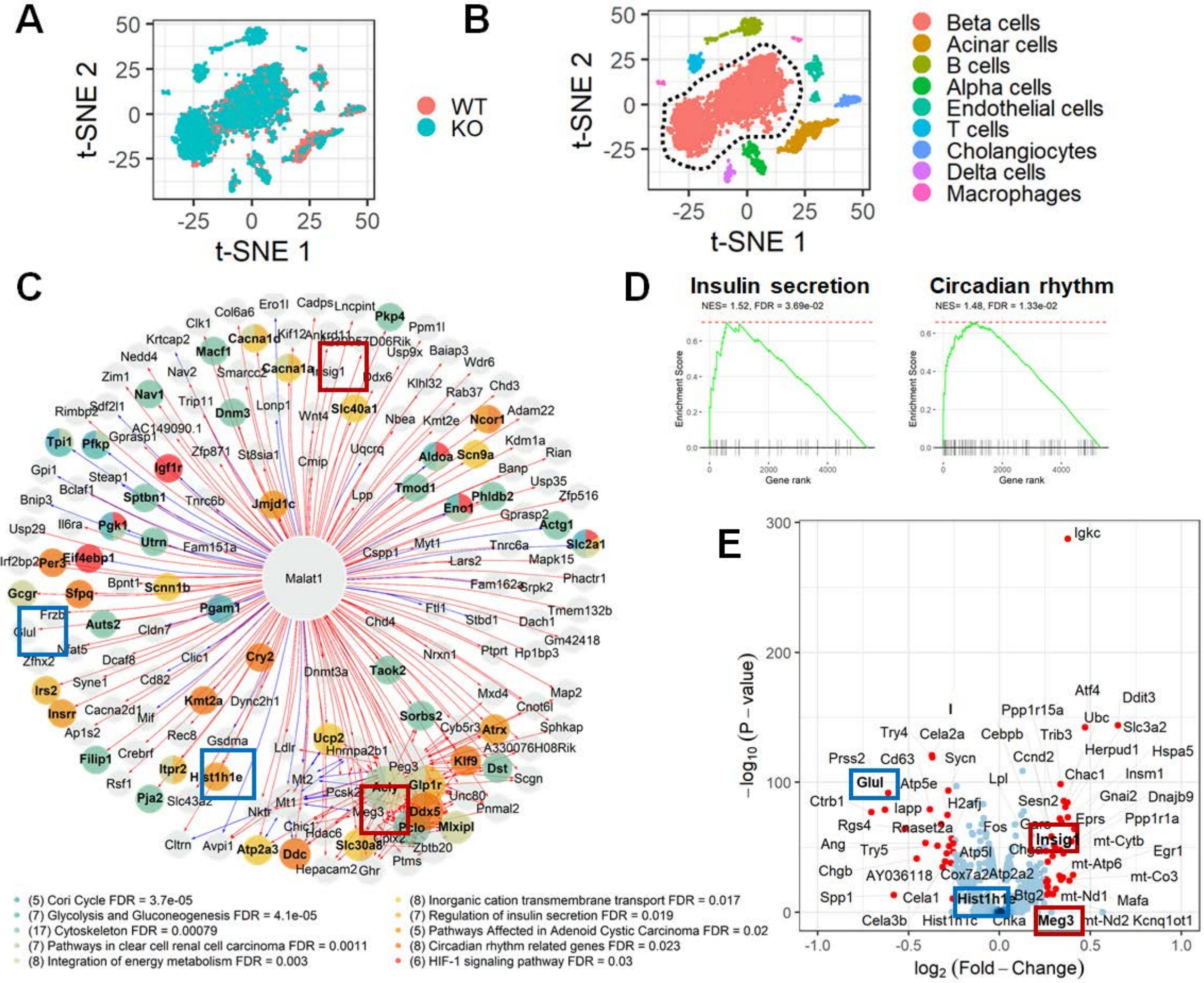


Figure 6.6: scTenifoldKnk predicts transcriptional changes in beta cells of *Malat1*$^{-/-}$ KO mice. (A) UMAP plot of cells distinguished by experimental groups: WT or KO. (B) Cell clustering and cell type assignment. Beta cells included in the analysis are highlighted with a dashed outline border. (C) Egocentric plot showing virtual-KO perturbed genes connected to *Malat1*. Nodes are color-coded by each gene's membership association with the top 10 significantly enriched functional groups. (D) GSEA analysis with scTenifoldKnk gene list identifies significant gene sets, including insulin secretion and circadian rhythm. (E) Volcano plot from the DE analysis comparing gene expression between WT and KO samples. In the plot, -log$_2$(FC) and -log$_{10}$(P-value) of genes are plotted against each other. Names of genes are given for those with a fold change (FC) $> 10\%$ and FDR < 0.05. Among them, 4 virtual-KO perturbed genes are highlighted with a red (*Insig1* and *Meg3*) or a blue (*Glu1* and *Hist1h1*) box, indicating up-regulated or down-regulated in the KO sample.

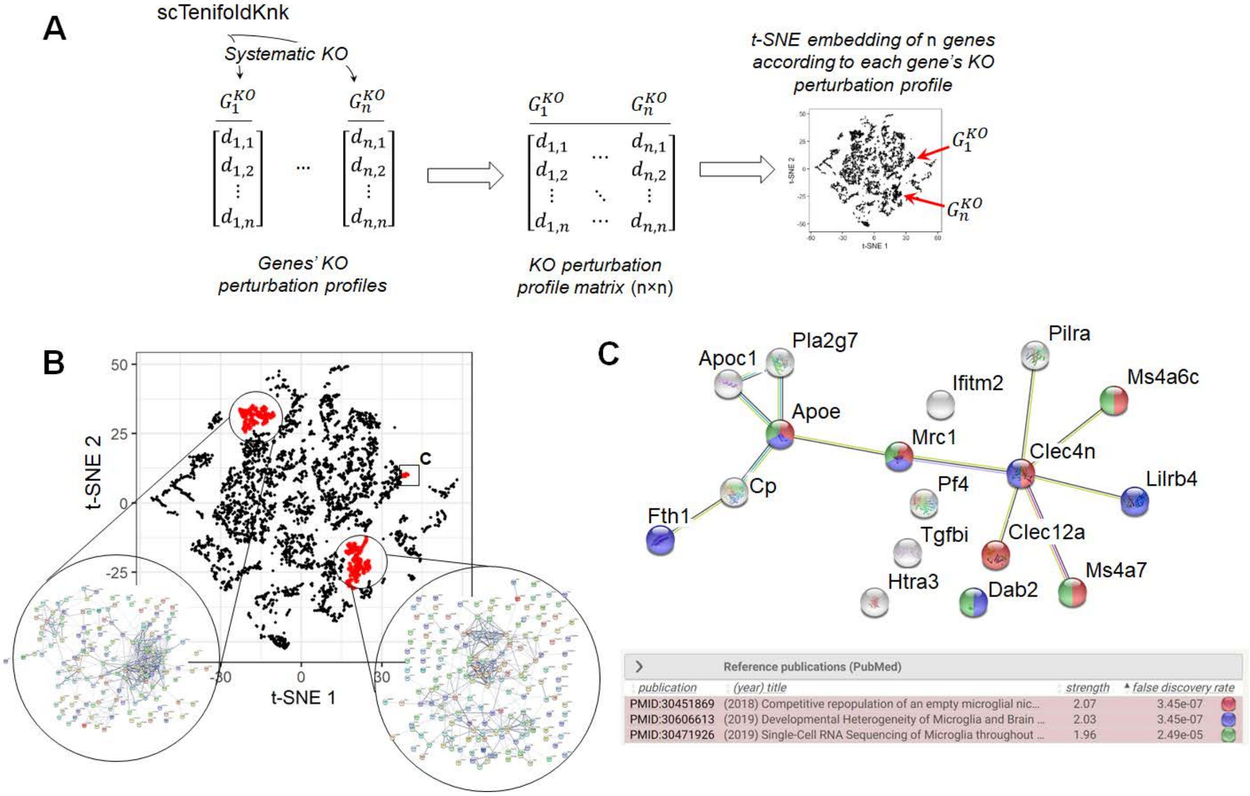
functions in various physiological and pathophysiological processes [294, 296, 297, 298, 299, 300, 301, 302], including cytoskeleton, circadian rhythm, and hypoxia-inducible factor (HIF) signaling pathway, as predicted (Figure 6.6 C.). STRING database search showed that virtual-KO perturbed genes form a highly connected interaction network (P < 0.01, STRING interaction enrichment test, Supplementary Figure S5). GSEA analysis with scTenifoldKnk-reported gene list indicated that 79 non-redundant gene sets were enriched significantly (FDR < 0.05, Supplementary Table S11), including insulin secretion and circadian rhythm (Figure 6D).

To validate these predictions made by using scTenifoldKnk, we performed DE analysis using MAST to compare gene expression in beta cells between WT and KO mice. We identified $1,695$ DE genes (FDR < 0.05, Figure 6.6 E.). Two up-regulated genes (*Insig1* and *Meg3*) and two down-regulated genes (*Glu1* and *Hist1h1*) were among the 167 virtual-KO perturbed genes (Figure 6.6 C.). GSEA analysis with the ranked list of DE genes sorted by the fold-change produced only 17 non-redundant enriched gene sets including insulin secretion.

6.3.6 Systematic KO of all genes in a given sample to generate the landscape of genes' perturbation profiles

As mentioned, scTenifoldKnk is designed and implemented to be computationally efficient. Indeed, we benchmarked the performance of scTenifoldKnk when the WT scGRN is given. Running on a standard workstation, scTenifoldKnk could complete a virtual KO analysis in one minute per gene. That is to say, for a given scGRN of $6,000$ genes, scTendifoldKnk could knock out all of these genes individually in 4 to 5 days in a workstation. The process can also be splitted and run in paraelle to increase the speed of computing. The outcome of such a systematic KO experiment is a collection of perturbation profiles of all genes. For each gene (e.g., gene δ), the perturbation profile of the gene (i.e., gene δ) is a vector of distances of all other genes ($\{\delta^-\}$) produced by scTenifoldKnk (see Figure 6.1 C.). The distance value quantifies the level of a gene (i.e., a gene in $\{\delta^-\}$) being perturbed by the deletion of the KO gene (i.e., gene δ). Figure 6.7 A/ illustrates an analytic flowchart when scTenifoldKnk is used in a systematic KO experiment. For a given WT scGRN with n genes, scTenifoldKnk can be used to delete individual genes from 1 to n. For the ith gene deletion, G_i^{KO}, scTenifoldKnk produces a KO perturbation profile for the ith gene. The KO perturbation profile is a vector of distances: $[d_{i,1}, d_{i,2}, ...d_{i,n}]^T$, where $i = 1, 2, \ldots, n$. Combining all genes' KO perturbation profiles into a $n \times n$ matrix, called KO perturbation profile matrix, is followed by t-SNE embedding and clustering of genes. As shown in Figure 6.7 B., each predicted perturbation profile caused by the deletion of a gene is unique (each point representing a perturbation profile is located in a different position), supporting the power of scTenifoldKnk to provide specific results for each gene's functional relationship in a cell-type-specific manner. Additionally, as reported before, genes with similar perturbation profiles will be closely located in the low dimensional embedding [303]. Therefore, examining genes closely located will allow us to discover potential functional associations between them.

To demonstrate the use of the systematic KO functionality, we downloaded the scRNA-seq data from the brain immune atlas and obtained the expression matrix of $6,853$ genes and $5,271$ microglial cells (see Section 6.2). These microglial cells were derived from WT homeostatic mice [246]. After knocking out all genes, we obtained the KO perturbation landscape of all $6,853$ genes, as shown in the t-SNE embedding (Figure 6.7 B.). We selected three clusters of different sizes to examine the member of genes in each cluster and the functional relationships between these genes. We found that all three clusters contain genes that show a significantly higher level of functional associations than expected by chance (P < 0.01, STRING interaction enrichment tests). We then focused on the smallest cluster, which contains 16 genes (Figure 6.7 C.). According to the STRING database, 12 of these genes (*Apoc1*, *Apoe*, *Clec12a*, *Clec4n*, *Cp*, *Fth1*, *Lilrb4*, *Mrc1*, *Ms4a6c*, *Ms4a7*, *Pilra* and Pla2g7) are functionally associated with each other. These associations are supported by evidence from published literature [304, 305, 306]. All these supporting references are related to the microglia study. The remaining 4 genes (*Htra3*, *Tgfbi*, *Pf4* and *Ifitm2*) are not connected to this 12 gene sub-network. Nevertheless, a new study [307] showed that *Htra3* is overexpressed in repopulating microglia. Therefore, these 'isolated' genes may be worthy of further scrutiny for their functions in microglia with the potential links to the other genes in the 12 gene sub-network.


A
scTenifoldKnk
Systematic KO
G_1^{KO}
G_n^{KO}
$\begin{bmatrix} d_{1,1} \\ d_{1,2} \\ \vdots \\ d_{1,n} \end{bmatrix}$ ... $\begin{bmatrix} d_{n,1} \\ d_{n,2} \\ \vdots \\ d_{n,n} \end{bmatrix}$
Genes' KO perturbation profiles
$\begin{bmatrix} d_{1,1} & \cdots & d_{n,1} \\ d_{1,2} & & d_{n,2} \\ \vdots & \ddots & \vdots \\ d_{1,n} & \cdots & d_{n,n} \end{bmatrix}$
KO perturbation profile matrix (n×n)
t-SNE embedding of n genes according to each gene's KO perturbation profile
t-SNE 1
t-SNE 2
B
50
25
0
-25
-30
0
30
t-SNE 1
t-SNE 2
c
C
Apoc1
Pla2g7
Apoe
Pilra
Ifitm2
Ms4a6c
Mrc1
Clec4n
Lilrb4
Cp
Pf4
Fth1
Tgfbi
Clec12a
Htra3
Dab2
Ms4a7

›	Reference publications (PubMed)			
publication	(year) title	strength	false discovery rate	
PMID:30451869	(2018) Competitive repopulation of an empty microglial nic...	2.07	3.45e-07	🔴
PMID:30606613	(2019) Developmental Heterogeneity of Microglia and Brain ...	2.03	3.45e-07	🔵
PMID:30471926	(2019) Single-Cell RNA Sequencing of Microglia throughout ...	1.96	2.49e-05	🟢

Figure 6.7: scTenifoldKnk enables systematic KO experiment in microglia and establishment of KO perturbation profile landscape. (A) An illustration of systematic virtual-KO analysis using scTenifoldKnk. Two KO genes are shown as an example. Due to the difference in their profiles, two genes are embedded in two different locations, indicated by the red arrows, in the dimensionality reduction visualization. (B) t-SNE embedding of $6,853$ genes expressed in microglia based on these genes' perturbation profiles. Genes in three clusters in the embedding are highlighted in red. The connections between genes are retrieved from the STING database. Clusters are zoomed to show the STING sub-networks. The rectangle indicates a sub-network shown in C. (C) STING sub-network of 12 genes (*Apoc1*, *Apoe*, *Clec12a*, *Clec4n*, *Cp*, *Fth1*, *Lilrb4*, *Mrc1*, *Ms4a6c*, *Ms4a7*, *Pilra* and *Pla2g7*) in the cluster highlighted with the rectangle in B. The references [304, 305, 306], from which the associations between genes are established, are given in the table.

Instead of using genes' KO perturbation profiles, using genes' expression information can also produce a landscape of genes. Indeed, we applied t-SNE to the UMI count matrix, which was transformed and standardized using the same procedure (see Section 6.2), and obtained the embedding plot of genes (Supplementary Figure S6B). The difference between this embedding plot derived from gene expression and the one derived from the gene's KO perturbation profile is obvious. The former has no structure among cells, but is just a cloud of data points. Performing clustering using the embedding derived from gene expression ought to not produce clusters containing more true interactions than using the embedding derived from gene expression profile across cells. To test whether similar results were obtained using expression and perturbation profiles, we calculated the average distance between genes that belong to the same KEGG gene set. We found that, across all KEGG gene sets, the average distance in the embedding derived from a KO perturbation profile was significantly smaller than that in the embedding derived from expression profile (Supplementary Figure S6C). Genes in the three example clusters, as shown in Figure 6.7 B., were found to be scattered in different clusters in the embedding derived from gene expression (Supplementary Figure S6B). We also conducted the same test using the embedding derived from the genes' profile of the weight of edges in the WT scGRN. The same pattern was uncovered – that is, the average distance is smaller in the embedding derived from the KO perturbation profile than in that derived from the scGRN edge weight. These results suggested that gene sets identified using the gene's KO perturbation profile were more likely to functionally connected than gene sets identified using other types of gene profiles.

6.3.7 Systematic KO of a gene across many different cell types

To precisely understand a gene's role requires the knockout of the gene in all the cell types across the body. It is, however, an intractable task if the role of the gene is necessary for cell survival as well as for the technical difficulties and high expenses of culturing different types of cells. In this case, using scTenifoldKnk to conduct virtual knockout of genes in multiple cell types will allow us to understand more confidently both the global and cell-type specific functions of a gene.

MYDGF (Myeloid Derived Growth Factor), also known as *C19orf10*, is a 142-residue protein broadly expressed in multiple tissues and cell types [308, 309]. In mouse model, *Mydgf* has shown to enhance cardiac myocyte survival, tissue repair and angiogenesis caused by myocardial infarction [310]. Although the structure and functionally important moieties of *MYDGF* have been identified, further elucidation of *MYDGF*'s functional roles is required. To find the primary targets of *MYDGF*, we conducted virtual knockout of the gene in 45 cell types. Datasets were downloaded from the PanglaoDB database [29]. For each cell-type we recovered a cell-type specific perturbation profile with a number of genes ranging from 4642 to 12386. Using the perturbation profile computed by scTenifoldKnk for endothelial cells (experimental surrogate used by Wang to study the functional role of *Mydgf* in mice heart) and gene set enrichment analysis (GSEA), we cross-validated the associated functional roles reported before in mice, the identified pathways include *Cell cycle*, *VEGFA-VEGFR2 pathway*, *Intra-Golgi traffic and activation* (Figure 6.8 A.) [311].

When all the perturbation profiles were concatenated to identify the broad targets of *Mydgf* we recovered a total of $1,288$ perturbed genes expressed in at least the 5% of the cells of a given cell-type and shared across all cell-types. Correlation analysis showed that the most similar patterns of perturbations to the ones recovered from endothelial cells are from T-cells, hepatocytes, cardiomyocytes, and kidney cells (Figure 6.8 B.). The enrichment analysis using ssGSEA analysis showed that the perturbation profiles are posititively enriched (ES > 0) with genes associated with *AKT signaling pathway*, *endothelial NOS activation pathway*, *apoptosis pathway*, *muscle contraction pathway*, and *ALK2 pathway* using the BioPlanet 2019 database as gene sets reference (Figure 6.8 C.) [312]. These results are consistent with previous results from mice reporting that overexpression or addition of *Mydgf* can also increase *AKT phosphorylation* and *cell proliferation* via *AKT/MAPK signaling pathways* [313]. Taken together, these pathways are closely related to the promoting survival and growth functions of *MYDGF* found in previous studies [308, 309, 310, 311, 313]. Thus, scTenifoldKnk can be used to conduct virtual KO of a gene across multiple cell types to identify common as well as cell type-specific transcriptional targets of the KO gene.

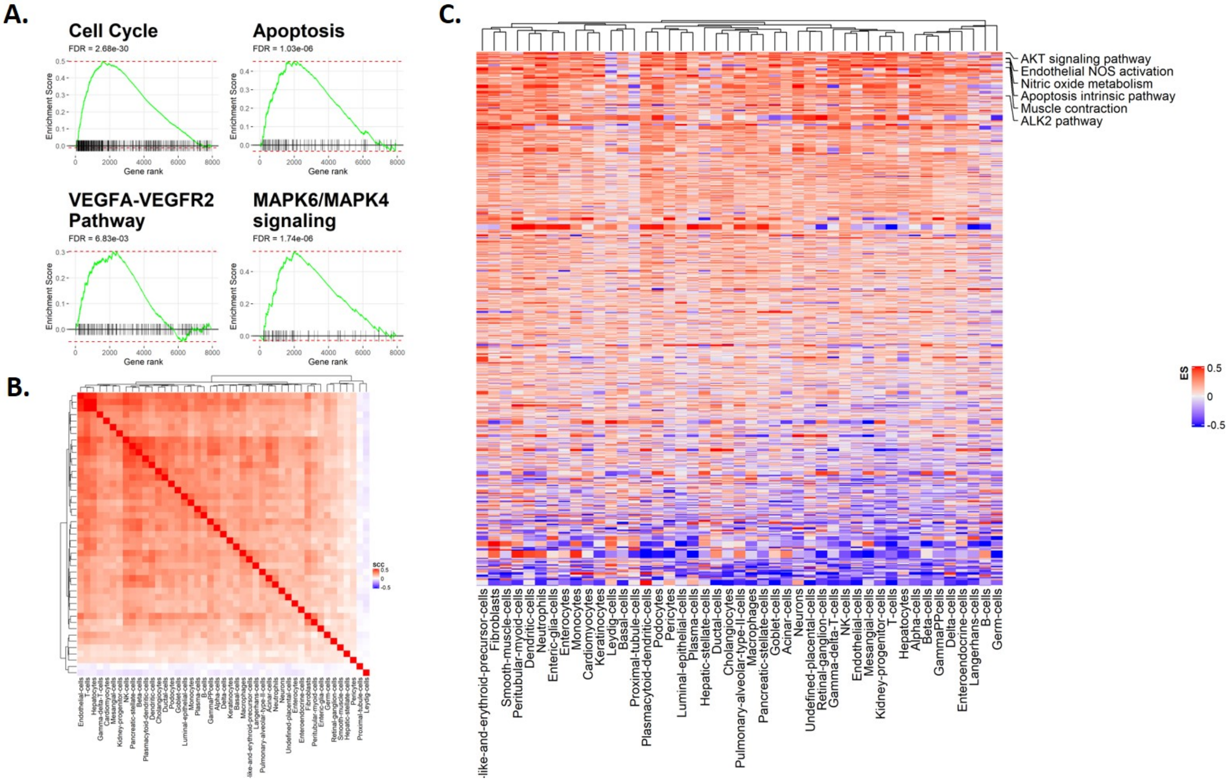


Figure 6.8: scTenifoldKnk enables the identification of universal and cell-type-specific functional roles of *Mydgf* across human cell-types. (A) GSEA analysis with scTenifoldKnk gene list identifies significant gene sets predicted to be perturbed in endothelial cells, including cell cycle, apoptosis, *VEGFA-VEGFR2* and *MAPK6/4* signaling pathways. (B) Correlation plot showing the similarity between perturbation profiles predicted for the selected 45 cell-types. (C) Heatmap showing the enrichment score (ES) for the gene sets reported in the BioPlanet database sorted by the average enrichment score across cell-types.

6.4 Discussion

Gene expression is almost always under coordinated regulation in cells of living organisms. Inferring GRNs is the key to a better understanding of such coordinated regulation. However, inferring GRNs is a very difficult process – there are always a large number of unknown variables in the system, and the power of inference is limited by the sample size. The development of single-cell technology has brought new 'oil' to network science. We have previously shown that scRNA-seq information can be leveraged to fuel the machine learning algorithms for reliable scGRN construction [244].

In a GRN, the regulatory effect manifests as observable synchronized patterns of expression between genes. These genes are associated with the same biological process, pathway, or under the control of the same set of transcription factors [25]. When a gene involved in a process is perturbed (e.g., knocked out), the expected first responders for such perturbation are those functionally closely related to the KO gene. Thus, modeling influence patterns in a GRN, such as using topological models to approximate perturbation patterns [26], can be used to bypass expensive experimental measurement. Thus, in principle, GRN-based perturbation analysis can contribute to the planning and design of real-animal experimental work for testing or validating hypotheses.

Our contribution is to provide such an scGRN-based perturbation analytical system – scTenifoldKnk. By performing analyses with a series of existing scRNA-seq data, we showed that scTenifoldKnk can be used to identify KO responsive genes. The reported genes were found to be enriched with molecular functions that are similar to those enriched in genes identified by the authentic KO experiments. We tested scTenifoldKnk using data from three different cell types, which are affected by three Mendelian diseases. While these diseases represent conditions involving distinct causal genes and different dysregulated molecular processes, scTenifoldKnk demonstrated its value in all cases. The top hits reported by scTenifoldKnk are highly interpretable, suggesting scTenifoldKnk can be used to reveal molecular mechanisms underlying less characterized pathophysiological processes. The predictive power of scTenifoldKnk was further demonstrated with scRNA-seq data from our own KO experiments, involving enterocytes in $Ahr^{-/-}$ mice and islet cells in *Malat1*$^{-/-}$ mice. Finally, we showed a case study of a systematic KO experiment using microglia data as an example.

Despite some obvious limitations associated with the virtual KO method, we start by discussing its advantages. First, the virtual KO method, as we implemented in scTenifoldKnk, is species agnostic – it works with scRNA-seq data from humans and animals alike. This feature gives the method a huge advantage for KO experiments focusing on human samples. In the lack of human KO samples, the KO animals are used as surrogates. The evolutionary divergence between humans and model animals is assumed to play a neglectable role in shaping orthologue gene function – but we know this is not always the case. While applying scTenifoldKnk to human scRNA-seq data, researchers can avoid many pitfalls caused by extending the conclusions from animal KO experiments to humans. Second, scTenifoldKnk allows for any gene to be knocked out for functional analysis as long as the gene expression is detectable in the WT sample. One may want to knock out all genes or a set of genes one by one to obtain an effect profile for each of the KO genes. The effect profile can be as simple as a virtual-KO disrupted gene list produced by scTenifoldKnk. Genes have similar effect profiles that are most likely to share molecular functions, or are involved in the same signaling pathways. Third, using scTenifoldKnk, it is feasible to virtually delete more than one gene from a system at the same time to study synergistic KO effects. An example is the double-KO settings for *Hnf4ag*^{DKO}. This is significant because the number of possible combinations of multiple KO targets is often too large to allow researchers to explore the impact of even a small fraction of the multi-gene KO possibilities. In this circumstance, scTenifoldKnk will be an invaluable tool to predict the consequences of these KO combinations, and provide experimental design guidance by suggesting the gene combinations that should be prioritized. Fourth, scTenifoldKnk can be used to study the effects of gene KO across multiple cell types. A typical scRNA-seq experiment generates expression data for multiple cell types. In this case, the scTenifoldKnk analysis will allow researchers to study diverse phenotypes associated with the KO gene in all cell types as long as the gene is expressed. Finally, scTenifoldKnk can be used to study the function of essential genes, for which the gene KO causes lethal outcomes, making it impossible to establish the KO animals. When scTenifoldKnk is applied to these essential genes, especially with embryonic expression data of the genes from the WT samples, developmental functions of these genes can be studied.

scTenifoldKnk can be used in extended research areas other than KO experiments. For example, biologists often need to know whether a manipulation has an effect or not. scTenifoldKnk can be applied not only to detect novel targets, but also to prioritize known targets prior to in vivo or in vitro studies with different experimental techniques. One may use drugs to block the transcription of a gene in a candidate pathway. If the drug has an effect, one would conclude that the drug works on that pathway involved; otherwise, the pathway is not affected. The scTenifoldKnk-based analysis represents a novel analytical method that can be applied to genome-wide association studies (GWAS). Throughout the last decade, GWAS have been successful in detecting associations between variants and phenotypes; however, a phenotypic trait is usually associated with many variants presumably influencing gene expression regulation. scTenifoldKnk may be used to help geneticists to assess functional consequences in order to prioritize actionable gene targets. We point out that the prediction of scTenifoldKnk does not have to be perfect. As we can see, with many example data sets, scTenifoldKnk could recapitulate major findings reported in real-KO experiments. We showed the landscape of KO perturbation profiles of all genes in microglia. Experimentally, such systematic perturbation analysis has only been done in yeast [314]. With scTenifoldKnk, systematic KO analysis in silico can be performed in a cell-type-specific manner in organismal systems.

Limitations of scTenifoldKnk are inherited from being a virtual KO method. scTenifoldKnk cannot be used to predict the consequence of gene overexpression, which is also a commonly used method for gene function study. Also, as the power of scTenifoldKnk is rooted from the WT scGRN, the regulatory network from the WT sample, the prediction that scTenifoldKnk may ‘favor’ regulatory rather than structural genes. Nevertheless, it is still possible to make adjustments to some details in the implementation of scTenifoldKnk to make it better fit user analytical needs. For example, instead of knocking out a target gene by setting its values to 0 in the adjacency matrix, randomly shuffling the count values in the expression matrix to construct a network that mimics the effect of gene dysregulation and comparing against the original may also provide insights about the functional role of the gene.

Prediction of gene expression responses to perturbation using scRNA-seq data is an active research area. To the best of our knowledge, there are two software tools that have been developed for this purpose: scGen [315] and CellOracle [316]. scGen is a package implemented in Python, using TensorFlow variational autoencoders combined with vector arithmetic, to predict gene expression changes in cells. Overall, scGen works like a neural network-empowered regression tool that predicts the changes of gene expression in cells in response to specific perturbations such as disease and drug treatment. scGen requires training data sets from samples before and after being exposed to the same perturbation. CellOracle is a workflow, developed in Python with several R dependencies, that integrates scRNA-seq and single-cell chromatin accessibility data (scATAC-seq) data to infer GRN and predict the changes of gene expression in response to specific perturbations. CellOracle constructs a GRN that accounts for the relationship between transcription factors and their target genes based on sequence motif analysis using the information provided by the scATAC-seq data. After that, the constructed GRN is further refined using regularized Bayesian regression models to remove weak connections and is adjusted to infer the context-dependent GRN using the scRNA-seq data. scTenifoldKnk has a different design compared to either scGen or CellOracle. scTenifoldKnk is specifically designed for performing virtual KO experiments. It requires only the scRNA-seq data from the WT sample: unlike scGen, scTenifoldKnk does not need training data; unlike CellOracle, scTenifoldKnk does not require extra scATAC-seq data to obtain information of transcription factors and their targets. The minimalist design of scTenifoldKnk should allow for adoption widely in most scenarios.

Overall, we have shown evidence that our machine learning workflow represents a powerful and efficient method for conducting virtual KO experiments. The highly efficient implementation of scTenifoldKnk allows systematic deletion of a large number of genes in scRNA-seq data. The prediction power offered by scTenifoldKnk enables accurate prediction of perturbations in regulatory networks caused by the deletion of a gene, revealing its function in a cell type-specific manner. We anticipate that scTenifoldKnk will be adopted and widely applied in the predictive science of single-cell biomedical research.

7. SUMMARY AND CONCLUSION

In this dissertation, I presented five chapters that describe the major work I accomplished during my Ph.D. research to investigate the functional implications of the single-cell gene expression variability in multicellular organization, and its applications in functional genomics.

In Chapter 2, we introduced a single-cell RNA-seq dataset for GM12878 and GM18502 – two LCLs derived from the blood of female donors of European and African ancestry, respectively. The final dataset contained $7,045$ cells from GM12878, $5,189$ from GM18502, and $5,820$ from the mixture, offering valuable information on single-cell gene expression in highly homogenous cell populations. This dataset is a suitable reference for population differentiation in gene expression at the single-cell level. We believe that the data from the mixture provide additional valuable information facilitating the development of statistical methods for data normalization and batch effect correction.

In Chapter 3, we presented the systematic analysis of $5,530,106$ cells reported in $1,349$ annotated datasets available in the PanglaoDB database, and reported that the average mtDNA$\%$ in scRNA-seq data across human tissues is significantly higher than that across mouse tissues. We found this difference is not confounded by single-cell platforms or the technologies used to generate the data. Based on this finding, we propose new reference values of the mtDNA$\%$ threshold for 121 tissues of mouse and 44 tissues of humans. We concluded that in general, for mouse tissues, with very few exceptions (3 of 121 analyzed tissues), the 5% threshold can still be used as a default, which performs well to distinguish between healthy and low-quality cells. We reported that for human tissues, the 5% threshold should be reconsidered as it fails to accurately discriminate between healthy and low-quality cells in 29.5% (13 of 44) tissues analyzed.

In Chapter 4, we analyzed multiple single-cell RNA-seq data sets from lymphoblastoid cell lines (LCLs), lung airway epithelial cells (LAECs), and dermal fibroblasts (DFs) and, for each cell type, selected a group of homogenous cells with highly similar expression profiles. We estimated the single-cell expression variability levels for genes after correcting the mean-variance dependency in that data and identified 465, 466, and 364 highly variable genes (HVGs) in LCLs, LAECs, and DFs, respectively. Functions of these HVGs were found to be enriched with those biological processes precisely relevant to the corresponding cell type's function, from which the scRNA-seq data used to identify HVGs were generated – e.g., cytokine signaling pathways were enriched in HVGs identified in LCLs, collagen formation in LAECs, and keratinization in DFs. We repeated the same analysis with scRNA-seq data from induced pluripotent stem cells (iPSCs) and identified only 79 HVGs with no statistically significant enriched functions; the overall single-cell expression variability in iPSCs was of negligible magnitude. Our results support the 'variation is function' hypothesis, arguing that single-cell expression variability is required for cell type-specific, higher-level system function. Thus, we concluded that quantifying and characterizing single-cell expression variability are of importance for our understating of normal and pathological cellular processes.

In Chapter 5, we introduced a robust and powerful machine learning workflow called 'scTenifoldNet' for comparative gene regulatory network (GRN) analysis of single cells. The scTenifoldNet workflow, consisting of principal component (PC)-regression, low-rank tensor approximation, and manifold alignment, constructs and compares transcriptome-wide single-cell GRNs (scGRNs) from different samples to identify gene expression signatures shifting with cellular activity changes such as those associated with pathophysiological processes and responses to environmental perturbations. We used simulated data to benchmark scTenifoldNet's performance, and then applied scTenifoldNet to several real data sets. In real-data applications, scTenifoldNet identified highly specific changes in gene regulation in response to acute morphine treatment, an antibody anticancer drug, gene knockout, double-stranded RNA stimulus, and amyloid-beta plaques in various types of mouse and human cells. We anticipate that scTenifoldNet can help achieve breakthroughs through constructing and comparing scGRNs in poorly characterized biological systems, by deciphering the full cellular and molecular complexity of the data.

In Chapter 6, we introduced 'scTenifoldKnk', a machine learning workflow for performing virtual KO experiments with data from single-cell RNA-seq. Using existing data sets, we demonstrate that the scTenifoldKnk analysis recapitulates main findings of three real-animal KO experiments and confirms the function of genes underlying three Mendelian diseases. We show the power of scTenifoldKnk as a predictive method to successfully predict the outcomes of two KO experiments we conducted. The two experiments involve intestinal enterocytes in *Ahr*$^{-/-}$ mice and pancreatic islet cells in *Malat1*$^{-/-}$ mice, respectively. Finally, we demonstrate the use of scTenifoldKnk to perform systematic KO analysis. We concluded that scTenifoldKnk is a powerful and efficient virtual KO tool for gene function study, allowing a systematic deletion of a large number of genes in single-cell RNA-seq data to reveal gene functionality in a cell type-specific manner.

REFERENCES

[1] B. Strober, R. Elorbany, K. Rhodes, N. Krishnan, K. Tayeb, A. Battle, and Y. Gilad, "Dynamic genetic regulation of gene expression during cellular differentiation," *Science*, vol. 364, no. 6447, pp. 1287–1290, 2019.

[2] K. Struhl, "Molecular mechanisms of transcriptional regulation in yeast," *Annual review of biochemistry*, vol. 58, no. 1, pp. 1051–1077, 1989.

[3] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[4] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature communications*, vol. 8, no. 1, pp. 1–12, 2017.

[5] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells," *Nature biotechnology*, vol. 33, no. 2, pp. 155–160, 2015.

[6] C. Trapnell, "Defining cell types and states with single-cell genomics," *Genome research*, vol. 25, no. 10, pp. 1491–1498, 2015.

[7] A. Raj and A. Van Oudenaarden, "Nature, nurture, or chance: stochastic gene expression and its consequences," *Cell*, vol. 135, no. 2, pp. 216–226, 2008.

[8] K. E. Willmore, N. M. Young, and J. T. Richtsmeier, "Phenotypic variability: its components, measurement and underlying developmental processes," *Evolutionary Biology*, vol. 34, no. 3, pp. 99–120, 2007.

[9] S. Girirajan and E. E. Eichler, "Phenotypic variability and genetic susceptibility to genomic disorders," *Human molecular genetics*, vol. 19, no. R2, pp. R176–R187, 2010.

[10] E. Simonovsky, R. Schuster, and E. Yeger-Lotem, “Large-scale analysis of human gene expression variability associates highly variable drug targets with lower drug effectiveness and safety,” *Bioinformatics*, vol. 35, no. 17, pp. 3028–3037, 2019.

[11] P. S. Swain, M. B. Elowitz, and E. D. Siggia, “Intrinsic and extrinsic contributions to stochasticity in gene expression,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12795–12800, 2002.

[12] H. Dueck, J. Eberwine, and J. Kim, “Variation is function: are single cell differences functionally important? testing the hypothesis that single cell variation is required for aggregate function,” *Bioessays*, vol. 38, no. 2, pp. 172–180, 2016.

[13] S. Mitchell, K. Roy, T. A. Zangle, and A. Hoffmann, “Nongenetic origins of cell-to-cell variability in b lymphocyte proliferation,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 12, pp. E2888–E2897, 2018.

[14] B. Snijder, R. Sacher, P. Rämö, E.-M. Damm, P. Liberali, and L. Pelkmans, “Population context determines cell-to-cell variability in endocytosis and virus infection,” *Nature*, vol. 461, no. 7263, pp. 520–523, 2009.

[15] A. McDavid, L. Dennis, P. Danaher, G. Finak, M. Krouse, A. Wang, P. Webster, J. Beechem, and R. Gottardo, “Modeling bi-modality improves characterization of cell cycle on gene expression in single cells,” *PLoS Comput Biol*, vol. 10, no. 7, p. e1003696, 2014.

[16] B. Li and L. You, “Predictive power of cell-to-cell variability,” *Quantitative Biology*, vol. 1, no. 2, pp. 131–139, 2013.

[17] D. E. Zak, G. E. Gonye, J. S. Schwaber, and F. J. Doyle, “Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network,” *Genome research*, vol. 13, no. 11, pp. 2396–2405, 2003.

[18] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano, “Reverse engineering of regulatory networks in human b cells,” *Nature genetics*, vol. 37, no. 4, pp. 382–390, 2005.

[19] A. J. Hartemink, “Reverse engineering gene regulatory networks,” *Nature biotechnology*, vol. 23, no. 5, pp. 554–555, 2005.

[20] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke, “Gene regulatory network inference: data integration in dynamic models—a review,” *Biosystems*, vol. 96, no. 1, pp. 86–103, 2009.

[21] A. T. Lun, D. J. McCarthy, and J. C. Marioni, “A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor,” *F1000Research*, vol. 5, 2016.

[22] A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, and T. Murali, “Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data,” *Nature methods*, vol. 17, no. 2, pp. 147–154, 2020.

[23] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, “Single-cell rna-seq denoising using a deep count autoencoder,” *Nature communications*, vol. 10, no. 1, pp. 1–14, 2019.

[24] D. Van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, *et al.*, “Recovering gene interactions from single-cell data using data diffusion,” *Cell*, vol. 174, no. 3, pp. 716–729, 2018.

[25] P. Michalak, “Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes,” *Genomics*, vol. 91, no. 3, pp. 243–248, 2008.

[26] M. Santolini and A.-L. Barabási, “Predicting perturbation patterns from the topology of biological networks,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 27, pp. E6375–E6383, 2018.

[27] D. Li and J. Gao, “Towards perturbation prediction of biological networks using deep learning,” *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[28] F. Markowetz, “How to understand the cell by breaking it: network analysis of gene perturbation screens,” *PLoS Comput Biol*, vol. 6, no. 2, p. e1000655, 2010.

[29] O. Franzén, L.-M. Gan, and J. L. Björkegren, “Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data,” *Database*, vol. 2019, 2019.

[30] H. Dueck, M. Khaladkar, T. K. Kim, J. M. Spaethling, C. Francis, S. Suresh, S. A. Fisher, P. Seale, S. G. Beck, T. Bartfai, *et al.*, “Deep sequencing reveals cell-type-specific patterns of single-cell transcriptome variation,” *Genome biology*, vol. 16, no. 1, pp. 1–17, 2015.

[31] N. Nagy, “Establishment of ebv-infected lymphoblastoid cell lines,” in *Epstein Barr Virus*, pp. 57–64, Springer, 2017.

[32] H. Neitzel, “A routine method for the establishment of permanent growing lymphoblastoid cell lines,” *Human genetics*, vol. 73, no. 4, pp. 320–326, 1986.

[33] A. Mohyuddin, Q. Ayub, S. Siddiqi, D. R. Carvalho-Silva, K. Mazhar, S. Rehman, S. Firasat, A. Dar, C. Tyler-Smith, and S. Q. Mehdi, “Genetic instability in ebv-transformed lymphoblastoid cell lines,” *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1670, no. 1, pp. 81–83, 2004.

[34] 1000 Genomes Project Consortium, “A map of human genome variation from population-scale sequencing,” *Nature*, vol. 467, no. 7319, p. 1061, 2010.

[35] P. C. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, *et al.*, “Genome-wide detection and characterization of positive selection in human populations,” *Nature*, vol. 449, no. 7164, pp. 913–918, 2007.

[36] L. Sie, S. Loong, and E. Tan, “Utility of lymphoblastoid cell lines,” *Journal of neuroscience research*, vol. 87, no. 9, pp. 1953–1959, 2009.

[37] T. Hussain and R. Mulherkar, “Lymphoblastoid cell lines: a continuous in vitro source of cells to study carcinogen sensitivity and dna repair,” *International journal of molecular and cellular medicine*, vol. 1, no. 2, p. 75, 2012.

[38] S. Jiang, L. W. Wang, M. J. Walsh, S. J. Trudeau, C. Gerdt, B. Zhao, and B. E. Gewurz, "Crispr/cas9-mediated genome editing in epstein-barr virus-transformed lymphoblastoid b-cell lines," *Current protocols in molecular biology*, vol. 121, no. 1, pp. 31–12, 2018.

[39] S.-M. Shim, H.-Y. Nam, J.-E. Lee, J.-W. Kim, B.-G. Han, and J.-P. Jeon, "Micrornas in human lymphoblastoid cell lines," *Critical Reviews™ in Eukaryotic Gene Expression*, vol. 22, no. 3, 2012.

[40] H. E. Wheeler and M. E. Dolan, "Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation," *Pharmacogenomics*, vol. 13, no. 1, pp. 55–70, 2012.

[41] D. Gurwitz, "Human ipsc-derived neurons and lymphoblastoid cells for personalized medicine research in neuropsychiatric disorders," *Dialogues in clinical neuroscience*, vol. 18, no. 3, p. 267, 2016.

[42] A. Ansel, J. P. Rosenzweig, P. D. Zisman, M. Melamed, and B. Gesundheit, "Variation in gene expression in autism spectrum disorders: an extensive review of transcriptomic studies," *Frontiers in neuroscience*, vol. 10, p. 601, 2017.

[43] M. Amoli, D. Carthy, H. Platt, and W. Ollier, "Ebv immortalization of human b lymphocytes separated from small volumes of cryo-preserved whole blood," *International journal of epidemiology*, vol. 37, no. suppl_1, pp. i41–i45, 2008.

[44] T. Lappalainen, M. Sammeth, M. R. Friedländer, P. Ac't Hoen, J. Monlong, M. A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, *et al.*, "Transcriptome and genome sequencing uncovers functional variation in humans," *Nature*, vol. 501, no. 7468, pp. 506–511, 2013.

[45] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard, "Understanding mechanisms underlying human gene expression variation with rna sequencing," *Nature*, vol. 464, no. 7289, pp. 768–772, 2010.

[46] A. R. Martin, H. A. Costa, T. Lappalainen, B. M. Henn, J. M. Kidd, M.-C. Yee, F. Grubert, H. M. Cann, M. Snyder, S. B. Montgomery, *et al.*, “Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture,” *PLoS Genet*, vol. 10, no. 8, p. e1004549, 2014.

[47] The ENCODE Project Consortium, “An integrated encyclopedia of dna elements in the human genome,” *Nature*, vol. 489, no. 7414, p. 57, 2012.

[48] A. Sajantila, “Editors’ pick: transcriptomes of 1000 genomes,” *Investigative Genetics*, vol. 4, no. 17, pp. 1–2, 2013.

[49] B. E. Stranger, A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird, C. Beazley, C. E. Ingle, M. Dunning, P. Flicek, D. Koller, *et al.*, “Population genomics of human gene expression,” *Nature genetics*, vol. 39, no. 10, pp. 1217–1224, 2007.

[50] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, *et al.*, “mRNA-seq whole-transcriptome analysis of a single cell,” *Nature methods*, vol. 6, no. 5, pp. 377–382, 2009.

[51] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann, “The technology and biology of single-cell rna sequencing,” *Molecular cell*, vol. 58, no. 4, pp. 610–620, 2015.

[52] A. K. Shalek, R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, *et al.*, “Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells,” *Nature*, vol. 498, no. 7453, pp. 236–240, 2013.

[53] G. K. Marinov, B. A. Williams, K. McCue, G. P. Schroth, J. Gertz, R. M. Myers, and B. J. Wold, “From single-cell to cell-pool transcriptomes: stochasticity in gene expression and rna splicing,” *Genome research*, vol. 24, no. 3, pp. 496–510, 2014.

[54] B. Zhao, L. A. Barrera, I. Ersing, B. Willox, S. C. Schmidt, H. Greenfeld, H. Zhou, S. B. Mollo, T. T. Shi, K. Takasaki, *et al.*, "The nf-κb genomic landscape in lymphoblastoid b cells," *Cell reports*, vol. 8, no. 5, pp. 1595–1606, 2014.

[55] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, "Spatial reconstruction of single-cell gene expression data," *Nature biotechnology*, vol. 33, no. 5, pp. 495–502, 2015.

[56] I. Tirosh, B. Izar, S. M. Prakadan, M. H. Wadsworth, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, *et al.*, "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq," *Science*, vol. 352, no. 6282, pp. 189–196, 2016.

[57] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the national academy of sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[58] R. Gaujoux and C. Seoighe, "A flexible r package for nonnegative matrix factorization," *BMC bioinformatics*, vol. 11, no. 1, pp. 1–9, 2010.

[59] M. Kasowski, F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S. M. Waszak, L. Habegger, J. Rozowsky, M. Shi, A. E. Urban, *et al.*, "Variation in transcription factor binding among humans," *science*, vol. 328, no. 5975, pp. 232–235, 2010.

[60] N. E. Banovich, Y. I. Li, A. Raj, M. C. Ward, P. Greenside, D. Calderon, P. Y. Tung, J. E. Burnett, M. Myrthil, S. M. Thomas, *et al.*, "Impact of regulatory variation across human ipscs and differentiated cells," *Genome research*, vol. 28, no. 1, pp. 122–131, 2018.

[61] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature methods*, vol. 14, no. 4, pp. 417–419, 2017.

[62] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for rna-sequencing and microarray studies," *Nucleic acids research*, vol. 43, no. 7, pp. e47–e47, 2015.

[63] F. Papavasiliou, R. Casellas, H. Suh, X.-F. Qin, E. Besmer, R. Pelanda, D. Nemazee, K. Rajewsky, and M. C. Nussenzweig, "V (d) j recombination in mature b cells: a mechanism for altering antibody responses," *Science*, vol. 278, no. 5336, pp. 298–301, 1997.

[64] S. Tonegawa, "Somatic generation of antibody diversity," *Nature*, vol. 302, no. 5909, pp. 575–581, 1983.

[65] J. L. Ryan, W. K. Kaufmann, N. Raab-Traub, S. E. Oglesbee, L. A. Carey, and M. L. Gulley, "Clonal evolution of lymphoblastoid cell lines," *Laboratory investigation*, vol. 86, no. 11, pp. 1193–1200, 2006.

[66] S. A. MacParland, J. C. Liu, X.-Z. Ma, B. T. Innes, A. M. Bartczak, B. K. Gage, J. Manuel, N. Khuu, J. Echeverri, I. Linares, *et al.*, "Single cell rna sequencing of human liver reveals distinct intrahepatic macrophage populations," *Nature communications*, vol. 9, no. 1, pp. 1–21, 2018.

[67] J. D. Bloom, "Estimating the frequency of multiplets in single-cell rna sequencing from cell-mixing experiments," *PeerJ*, vol. 6, p. e5578, 2018.

[68] C. S. McGinnis, L. M. Murrow, and Z. J. Gartner, "Doubletfinder: doublet detection in single-cell rna sequencing data using artificial nearest neighbors," *Cell systems*, vol. 8, no. 4, pp. 329–337, 2019.

[69] S. L. Wolock, R. Lopez, and A. M. Klein, "Scrublet: computational identification of cell doublets in single-cell transcriptomic data," *Cell systems*, vol. 8, no. 4, pp. 281–291, 2019.

[70] E. A. DePasquale, D. J. Schnell, P.-J. Van Camp, Í. Valiente-Alandí, B. C. Blaxall, H. L. Grimes, H. Singh, and N. Salomonis, "Doubletdecon: deconvoluting doublets from single-cell rna-sequencing data," *Cell reports*, vol. 29, no. 6, pp. 1718–1727, 2019.

[71] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nature biotechnology*, vol. 32, no. 4, p. 381, 2014.

[72] R. Sandberg, “Entering the era of single-cell transcriptomics in biology and medicine,” *Nature methods*, vol. 11, no. 1, pp. 22–24, 2014.

[73] B. Hwang, J. H. Lee, and D. Bang, “Single-cell rna sequencing technologies and bioinformatics pipelines,” *Experimental & molecular medicine*, vol. 50, no. 8, pp. 1–14, 2018.

[74] M. D. Luecken and F. J. Theis, “Current best practices in single-cell rna-seq analysis: a tutorial,” *Molecular systems biology*, vol. 15, no. 6, p. e8746, 2019.

[75] F. Ji and R. I. Sadreyev, “Single-cell rna-seq: Introduction to bioinformatics analysis,” *Current protocols in molecular biology*, vol. 127, no. 1, p. e92, 2019.

[76] R. Guantes, A. Rastrojo, R. Neves, A. Lima, B. Aguado, and F. J. Iborra, “Global variability in gene expression and alternative splicing is modulated by mitochondrial content,” *Genome research*, vol. 25, no. 5, pp. 633–644, 2015.

[77] R. Muir, A. Diot, and J. Poulton, “Mitochondrial content is central to nuclear gene expression: Profound implications for human health,” *Bioessays*, vol. 38, no. 2, pp. 150–156, 2016.

[78] T. Ilicic, J. K. Kim, A. A. Kolodziejczyk, F. O. Bagger, D. J. McCarthy, J. C. Marioni, and S. A. Teichmann, “Classification of low quality cells from single-cell rna-seq data,” *Genome biology*, vol. 17, no. 1, pp. 1–15, 2016.

[79] Q. Zhao, J. Wang, I. V. Levichkin, S. Stasinopoulos, M. T. Ryan, and N. J. Hoogenraad, “A mitochondrial specific stress response in mammalian cells,” *The EMBO journal*, vol. 21, no. 17, pp. 4411–4419, 2002.

[80] A. A. AlJanahi, M. Danielsen, and C. E. Dunbar, “An introduction to the analysis of single-cell rna-sequencing data,” *Molecular Therapy-Methods & Clinical Development*, vol. 10, pp. 189–196, 2018.

[81] A. Ma, Z. Zhu, M. Ye, and F. Wang, “Ensemblekqc: an unsupervised ensemble learning method for quality control of single cell rna-seq sequencing data,” in *International Conference on Intelligent Computing*, pp. 493–504, Springer, 2019.

[82] D. J. McCarthy, K. R. Campbell, A. T. Lun, and Q. F. Wills, "Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r," *Bioinformatics*, vol. 33, no. 8, pp. 1179–1186, 2017.

[83] T. R. Mercer, S. Neph, M. E. Dinger, J. Crawford, M. A. Smith, A.-M. J. Shearwood, E. Haugen, C. P. Bracken, O. Rackham, J. A. Stamatoyannopoulos, *et al.*, "The human mitochondrial transcriptome," *Cell*, vol. 146, no. 4, pp. 645–658, 2011.

[84] S. Lukassen, E. Bosch, A. B. Ekici, and A. Winterpacht, "Single-cell rna sequencing of adult mouse testes," *Scientific data*, vol. 5, no. 1, pp. 1–7, 2018.

[85] V. Svensson, E. da Veiga Beltrame, and L. Pachter, "A curated database reveals trends in single-cell transcriptomics," *Database*, vol. 2020, 2020.

[86] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[87] D. Temple Lang, *XML: Tools for Parsing and Generating XML Within R and S-Plus*, 2020. R package version 3.99-0.5.

[88] H. Wickham, J. Hester, and J. Ooms, *xml2: Parse XML*, 2020. R package version 1.3.2.

[89] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, *et al.*, "Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data," *Genome biology*, vol. 16, no. 1, pp. 1–13, 2015.

[90] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[91] G. Korotkevich, V. Sukhov, and A. Sergushichev, "Fast gene set enrichment analysis," *BioRxiv*, p. 060012, 2019.

[92] D. Ordoñez-Rueda, B. Baying, D. Pavlinic, L. Alessandri, Y. Yeboah, J. J. Landry, R. Calogero, V. Benes, and M. Paulsen, "Apoptotic cell exclusion and bias-free single-cell

selection are important quality control requirements for successful single-cell sequencing applications," *Cytometry Part A*, vol. 97, no. 2, pp. 156–167, 2020.

[93] P.-L. Germain, A. Sonrel, and M. D. Robinson, "pipecomp, a general framework for the evaluation of computational pipelines, reveals performant single cell rna-seq preprocessing tools," *Genome biology*, vol. 21, no. 1, pp. 1–28, 2020.

[94] X. Zhang, T. Li, F. Liu, Y. Chen, J. Yao, Z. Li, Y. Huang, and J. Wang, "Comparative analysis of droplet-based ultra-high-throughput single-cell rna-seq systems," *Molecular cell*, vol. 73, no. 1, pp. 130–142, 2019.

[95] M. S. Ko, "Problems and paradigms: Induction mechanism of a single gene molecule: Stochastic or deterministic?," *Bioessays*, vol. 14, no. 5, pp. 341–346, 1992.

[96] S. Fiering, E. Whitelaw, and D. I. Martin, "To be or not to be active: the stochastic nature of enhancer action," *Bioessays*, vol. 22, no. 4, pp. 381–387, 2000.

[97] A. Eldar and M. B. Elowitz, "Functional roles for noise in genetic circuits," *Nature*, vol. 467, no. 7312, pp. 167–173, 2010.

[98] L. Pelkmans, "Using cell-to-cell variability—a new era in molecular biology," *Science*, vol. 336, no. 6080, pp. 425–426, 2012.

[99] P. Kumar, Y. Tan, and P. Cahan, "Understanding development and stem cells using single cell-based analyses of gene expression," *Development*, vol. 144, no. 1, pp. 17–32, 2017.

[100] M. F. Wernet, E. O. Mazzoni, A. Celik, D. M. Duncan, I. Duncan, and C. Desplan, "Stochastic spineless expression creates the retinal mosaic for colour vision," *Nature*, vol. 440, no. 7081, pp. 174–180, 2006.

[101] H. H. Chang, M. Hemberg, M. Barahona, D. E. Ingber, and S. Huang, "Transcriptome-wide noise controls lineage choice in mammalian progenitor cells," *Nature*, vol. 453, no. 7194, pp. 544–547, 2008.

[102] A. J. Faure, J. M. Schmiedel, and B. Lehner, "Systematic analysis of the determinants of gene expression noise in embryonic stem cells," *Cell systems*, vol. 5, no. 5, pp. 471–484, 2017.

[103] C. P. Martinez-Jimenez, N. Eling, H.-C. Chen, C. A. Vallejos, A. A. Kolodziejczyk, F. Connor, L. Stojic, T. F. Rayner, M. J. Stubbington, S. A. Teichmann, *et al.*, "Aging increases cell-to-cell transcriptional variability upon immune stimulation," *Science*, vol. 355, no. 6332, pp. 1433–1436, 2017.

[104] C. D. Wiley, J. M. Flynn, C. Morrissey, R. Lebofsky, J. Shuga, X. Dong, M. A. Unger, J. Vijg, S. Melov, and J. Campisi, "Analysis of individual cells identifies cell-to-cell variability following induction of cellular senescence," *Aging cell*, vol. 16, no. 5, pp. 1043–1050, 2017.

[105] E. Azizi, A. J. Carr, G. Plitas, A. E. Cornish, C. Konopacki, S. Prabhakaran, J. Nainys, K. Wu, V. Kiseliovas, M. Setty, *et al.*, "Single-cell map of diverse immune phenotypes in the breast tumor microenvironment," *Cell*, vol. 174, no. 5, pp. 1293–1308, 2018.

[106] Å. Segerstolpe, A. Palasantza, P. Eliasson, E.-M. Andersson, A.-C. Andréasson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M. K. Bjursell, *et al.*, "Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes," *Cell metabolism*, vol. 24, no. 4, pp. 593–607, 2016.

[107] S. Tay, J. J. Hughey, T. K. Lee, T. Lipniacki, S. R. Quake, and M. W. Covert, "Single-cell nf-κb dynamics reveal digital activation and analogue information processing," *Nature*, vol. 466, no. 7303, pp. 267–271, 2010.

[108] A. Raj, S. A. Rifkin, E. Andersen, and A. Van Oudenaarden, "Variability in gene expression underlies incomplete penetrance," *Nature*, vol. 463, no. 7283, pp. 913–918, 2010.

[109] E. M. Kernfeld, R. M. Genga, K. Neherin, M. E. Magaletta, P. Xu, and R. Maehr, "A single-cell transcriptomic atlas of thymus organogenesis resolves cell types and developmental maturation," *Immunity*, vol. 48, no. 6, pp. 1258–1270, 2018.

[110] R. J. Miragaia, X. Zhang, T. Gomes, V. Svensson, T. Ilicic, J. Henriksson, G. Kar, and T. Lönnberg, "Single-cell rna-sequencing resolves self-antigen expression during mtec development," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[111] V. Svensson, R. Vento-Tormo, and S. A. Teichmann, "Exponential scaling of single-cell rna-seq in the past decade," *Nature protocols*, vol. 13, no. 4, pp. 599–604, 2018.

[112] K. Geiler-Samerotte, C. Bauer, S. Li, N. Ziv, D. Gresham, and M. Siegal, "The details in the distributions: why and how to study phenotypic variability," *Current opinion in biotechnology*, vol. 24, no. 4, pp. 752–759, 2013.

[113] J. C. Mar, "The rise of the distributions: why non-normality is important for understanding the transcriptome and beyond," *Biophysical reviews*, vol. 11, no. 1, pp. 89–94, 2019.

[114] K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. van den Elzen, M. J. Hirn, R. R. Coifman, *et al.*, "Visualizing structure and transitions in high-dimensional biological data," *Nature biotechnology*, vol. 37, no. 12, pp. 1482–1492, 2019.

[115] D. Osorio, X. Yu, P. Yu, E. Serpedin, and J. J. Cai, "Single-cell rna sequencing of a european and an african lymphoblastoid cell line," *Scientific data*, vol. 6, no. 1, pp. 1–8, 2019.

[116] D. M. Habiel, M. S. Espindola, I. C. Jones, A. L. Coelho, B. Stripp, and C. M. Hogaboam, "Ccr10+ epithelial cells from idiopathic pulmonary fibrosis lungs drive remodeling," *Jci Insight*, vol. 3, no. 16, 2018.

[117] T. Hagai, X. Chen, R. J. Miragaia, R. Rostom, T. Gomes, N. Kunowska, J. Henriksson, J.-E. Park, V. Proserpio, G. Donati, *et al.*, "Gene expression variability across cells and species shapes innate immunity," *Nature*, vol. 563, no. 7730, pp. 197–202, 2018.

[118] C. E. Friedman, Q. Nguyen, S. W. Lukowski, A. Helfer, H. S. Chiu, J. Miklas, S. Levy, S. Suo, J.-D. J. Han, P. Osteil, *et al.*, "Single-cell transcriptomic analysis of cardiac differentiation from human pscs reveals hopx-dependent cardiomyocyte maturation," *Cell stem cell*, vol. 23, no. 4, pp. 586–598, 2018.

[119] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nature biotechnology*, vol. 36, no. 5, pp. 411–420, 2018.

[120] P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, *et al.*, "Accounting for technical noise in single-cell rna-seq experiments," *Nature methods*, vol. 10, no. 11, p. 1093, 2013.

[121] J. J. Cai, "scgeatoolbox: a matlab toolbox for single-cell rna sequencing data analysis," *Bioinformatics*, vol. 36, no. 6, pp. 1948–1949, 2020.

[122] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[123] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, *et al.*, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W90–W97, 2016.

[124] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan, "Enrichr: interactive and collaborative html5 gene list enrichment analysis tool," *BMC bioinformatics*, vol. 14, no. 1, pp. 1–14, 2013.

[125] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, "Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists," *BMC bioinformatics*, vol. 10, no. 1, pp. 1–7, 2009.

[126] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 46, no. D1, pp. D649–D655, 2018.

[127] K. Konganti, G. Wang, E. Yang, and J. J. Cai, "Sbetoolbox: a matlab toolbox for biological network analysis," *Evolutionary Bioinformatics*, vol. 9, pp. EBO–S12012, 2013.

[128] C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano, "Great improves functional interpretation of cis-regulatory regions," *Nature biotechnology*, vol. 28, no. 5, pp. 495–501, 2010.

[129] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk, "Biomart–biological queries made easy," *BMC genomics*, vol. 10, no. 1, pp. 1–12, 2009.

[130] A. Khan, O. Fornes, A. Stigliani, M. Gheorghe, J. A. Castro-Mondragon, R. Van Der Lee, A. Bessy, J. Cheneby, S. R. Kulkarni, G. Tan, *et al.*, "Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework," *Nucleic acids research*, vol. 46, no. D1, pp. D260–D266, 2018.

[131] H.-I. H. Chen, Y. Jin, Y. Huang, and Y. Chen, "Detection of high variability in gene expression from single-cell rna-seq profiling," *BMC genomics*, vol. 17, no. 7, pp. 119–128, 2016.

[132] T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale, "Counting absolute numbers of molecules using unique molecular identifiers," *Nature methods*, vol. 9, no. 1, pp. 72–74, 2012.

[133] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[134] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using umap," *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[135] T. Tabib, C. Morse, T. Wang, W. Chen, and R. Lafyatis, "Sfrp2/dpp4 and fmo1/lsp1 define major fibroblast populations in human skin," *Journal of Investigative Dermatology*, vol. 138, no. 4, pp. 802–810, 2018.

[136] L. E. Tracy, R. A. Minasian, and E. Caterson, "Extracellular matrix and dermal fibroblast function in the healing wound," *Advances in wound care*, vol. 5, no. 3, pp. 119–136, 2016.

[137] P. K. Singhal, S. Sassi, L. Lan, P. Au, S. C. Halvorsen, D. Fukumura, R. K. Jain, and B. Seed, "Mouse embryonic fibroblasts exhibit extensive developmental and phenotypic diversity," *Proceedings of the National Academy of Sciences*, vol. 113, no. 1, pp. 122–127, 2016.

[138] A. Mantsoki, G. Devailly, and A. Joshi, "Gene expression variability in mammalian embryonic stem cells using single cell rna-seq data," *Computational biology and chemistry*, vol. 63, pp. 52–61, 2016.

[139] G. Courtois, A. Smahi, J. Reichenbach, R. Döffinger, C. Cancrini, M. Bonnet, A. Puel, C. Chable-Bessia, S. Yamaoka, J. Feinberg, *et al.*, "A hypermorphic iκbα mutation is associated with autosomal dominant anhidrotic ectodermal dysplasia and t cell immunodeficiency," *The Journal of clinical investigation*, vol. 112, no. 7, pp. 1108–1115, 2003.

[140] E. Lopez-Granados, J. E. Keenan, M. C. Kinney, H. Leo, N. Jain, C. A. Ma, R. Quinones, E. W. Gelfand, and A. Jain, "A novel mutation in nfkbia/ikba results in a degradation-resistant n-truncated protein and is associated with ectodermal dysplasia with immunodeficiency," *Human mutation*, vol. 29, no. 6, pp. 861–868, 2008.

[141] S. L. Nutt, P. D. Hodgkin, D. M. Tarlinton, and L. M. Corcoran, "The generation of antibody-secreting plasma cells," *Nature Reviews Immunology*, vol. 15, no. 3, pp. 160–171, 2015.

[142] R. Sciammas, Y. Li, A. Warmflash, Y. Song, A. R. Dinner, and H. Singh, "An incoherent regulatory network architecture that orchestrates b cell diversification in response to antigen signaling," *Molecular systems biology*, vol. 7, no. 1, p. 495, 2011.

[143] K. Roy, S. Mitchell, Y. Liu, S. Ohta, Y.-s. Lin, M. O. Metzig, S. L. Nutt, and A. Hoffmann, "A regulatory circuit controlling the dynamics of nfκb crel transitions b cells from proliferation to plasma cell differentiation," *Immunity*, vol. 50, no. 3, pp. 616–628, 2019.

[144] Q. H. Nguyen, S. W. Lukowski, H. S. Chiu, A. Senabouth, T. J. Bruxner, A. N. Christ, N. J. Palpant, and J. E. Powell, "Single-cell rna-seq of human induced pluripotent stem cells

reveals cellular heterogeneity and cell state transitions between subpopulations," *Genome Research*, vol. 28, no. 7, pp. 1053–1066, 2018.

[145] M. A. Mandegar, N. Huebsch, E. B. Frolov, E. Shin, A. Truong, M. P. Olvera, A. H. Chan, Y. Miyaoka, K. Holmes, C. I. Spencer, *et al.*, "Crispr interference efficiently induces specific and reversible gene silencing in human ipscs," *Cell stem cell*, vol. 18, no. 4, pp. 541–553, 2016.

[146] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins, "Stochasticity in gene expression: from theories to phenotypes," *Nature Reviews Genetics*, vol. 6, no. 6, pp. 451–464, 2005.

[147] J. M. Raser and E. K. O'shea, "Noise in gene expression: origins, consequences, and control," *Science*, vol. 309, no. 5743, pp. 2010–2013, 2005.

[148] G. Chalancon, C. N. Ravarani, S. Balaji, A. Martinez-Arias, L. Aravind, R. Jothi, and M. M. Babu, "Interplay between gene expression noise and regulatory network architecture," *Trends in genetics*, vol. 28, no. 5, pp. 221–232, 2012.

[149] J. C. Mar, N. A. Matigian, A. Mackay-Sim, G. D. Mellick, C. M. Sue, P. A. Silburn, J. J. McGrath, J. Quackenbush, and C. A. Wells, "Variance of gene expression identifies altered network constraints in neurological disease," *PLoS Genet*, vol. 7, no. 8, p. e1002207, 2011.

[150] R. D. Dar, N. N. Hosmane, M. R. Arkin, R. F. Siliciano, and L. S. Weinberger, "Screening for noise in gene expression identifies drug synergies," *Science*, vol. 344, no. 6190, pp. 1392–1396, 2014.

[151] S. Ecker, V. Pancaldi, A. Valencia, S. Beck, and D. S. Paul, "Epigenetic and transcriptional variability shape phenotypic plasticity," *Bioessays*, vol. 40, no. 2, p. 1700148, 2018.

[152] G. Altan-Bonnet and R. Mukherjee, "Cytokine-mediated communication: a quantitative appraisal of immune complexity," *Nature reviews Immunology*, vol. 19, no. 4, pp. 205–217, 2019.

[153] P. S. Hiemstra, P. B. McCray, and R. Bals, “The innate immune function of airway epithelial cells in inflammatory lung disease,” *European respiratory journal*, vol. 45, no. 4, pp. 1150–1162, 2015.

[154] M. Zhao, J. Zhang, H. Phatnani, S. Scheu, and T. Maniatis, “Stochastic expression of the interferon-β gene,” *PLoS Biol*, vol. 10, no. 1, p. e1001249, 2012.

[155] O. Sacco, M. Silvestri, F. Sabatini, R. Sale, A.-C. Defilippi, and G. A. Rossi, “Epithelial cells and fibroblasts: structural repair and remodelling in the airways,” *Paediatric respiratory reviews*, vol. 5, pp. S35–S40, 2004.

[156] G. Huang, D. Osorio, J. Guan, G. Ji, and J. J. Cai, “Overdispersed gene expression characterizes schizophrenic brains,” *BioRxiv*, p. 441527, 2018.

[157] J. Guan, E. Yang, J. Yang, Y. Zeng, G. Ji, and J. J. Cai, “Exploiting aberrant mrna expression in autism for gene discovery and diagnosis,” *Human genetics*, vol. 135, no. 7, pp. 797–811, 2016.

[158] S. Ecker, V. Pancaldi, D. Rico, and A. Valencia, “Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia,” *Genome medicine*, vol. 7, no. 1, pp. 1–12, 2015.

[159] J. Li, Y. Liu, T. Kim, R. Min, and Z. Zhang, “Gene expression variability within and between human populations and implications toward disease susceptibility,” *PLoS Comput Biol*, vol. 6, no. 8, p. e1000910, 2010.

[160] S. Spreizer, A. Aertsen, and A. Kumar, “From space to time: Spatial inhomogeneities lead to the emergence of spatiotemporal sequences in spiking neuronal networks,” *PLoS computational biology*, vol. 15, no. 10, p. e1007432, 2019.

[161] J. W. Ho, M. Stefani, C. G. Dos Remedios, and M. A. Charleston, “Differential variability analysis of gene expression and its application to human diseases,” *Bioinformatics*, vol. 24, no. 13, pp. i390–i398, 2008.

[162] S. C. Bendall and G. P. Nolan, “From single cells to deep phenotypes in cancer,” *Nature biotechnology*, vol. 30, no. 7, p. 639, 2012.

[163] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, “Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context,” in *BMC bioinformatics*, vol. 7, pp. 1–15, Springer, 2006.

[164] A. Irrthum, L. Wehenkel, P. Geurts, *et al.*, “Inferring regulatory networks from expression data using tree-based methods,” *PloS one*, vol. 5, no. 9, p. e12776, 2010.

[165] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, and G. Stolovitzky, “Wisdom of crowds for robust gene network inference,” *Nature methods*, vol. 9, no. 8, pp. 796–804, 2012.

[166] N. Friedman, M. Linial, I. Nachman, and D. Pe’er, “Using bayesian networks to analyze expression data,” *Journal of computational biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[167] R. Gill, S. Datta, and S. Datta, “A statistical framework for differential network analysis from microarray data,” *BMC bioinformatics*, vol. 11, no. 1, pp. 1–10, 2010.

[168] H. Todorov, R. Cannoodt, W. Saelens, and Y. Saeys, “Network inference from single-cell transcriptomic data,” in *Gene regulatory networks*, pp. 235–249, Springer, 2019.

[169] S. Aibar, C. B. González-Blas, T. Moerman, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, *et al.*, “Scenic: single-cell regulatory network inference and clustering,” *Nature methods*, vol. 14, no. 11, pp. 1083–1086, 2017.

[170] M. W. Fiers, L. Minnoye, S. Aibar, C. Bravo González-Blas, Z. Kalender Atak, and S. Aerts, “Mapping gene regulatory networks from single-cell omics data,” *Briefings in functional genomics*, vol. 17, no. 4, pp. 246–254, 2018.

[171] S. Chen and J. C. Mar, “Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data,” *BMC bioinformatics*, vol. 19, no. 1, pp. 1–21, 2018.

[172] S. Rabanser, O. Shchur, and S. Günnemann, “Introduction to tensor decompositions and their applications in machine learning,” *arXiv preprint arXiv:1711.10781*, 2017.

[173] R. Roscher, F. Schindler, and W. Förstner, “High dimensional correspondences from low dimensional manifolds–an empirical comparison of graph-based dimensionality reduction algorithms,” in *Asian Conference on Computer Vision*, pp. 334–343, Springer, 2010.

[174] Y. Chen, S. Z. Khong, and T. T. Georgiou, “On the definiteness of graph laplacians with negative weights: Geometrical and passivity-based approaches,” in *2016 American Control Conference (ACC)*, pp. 2488–2493, IEEE, 2016.

[175] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[176] U. Von Luxburg, “A tutorial on spectral clustering,” *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[177] A. Lachmann, H. Xu, J. Krishnan, S. I. Berger, A. R. Mazloom, and A. Ma’ayan, “Chea: transcription factor regulation inferred from integrating genome-wide chip-x experiments,” *Bioinformatics*, vol. 26, no. 19, pp. 2438–2444, 2010.

[178] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545–15550, 2005.

[179] O. Ríos, S. Frias, A. Rodríguez, S. Kofman, H. Merchant, L. Torres, and L. Mendoza, “A boolean network model of human gonadal sex determination,” *Theoretical Biology and Medical Modelling*, vol. 12, no. 1, pp. 1–18, 2015.

[180] T. E. Chan, M. P. Stumpf, and A. C. Babtie, “Gene regulatory network inference from single-cell data using multivariate information measures,” *Cell systems*, vol. 5, no. 3, pp. 251–267, 2017.

[181] S. Kim, “ppcor: an r package for a fast calculation to semi-partial correlation coefficients,” *Communications for statistical applications and methods*, vol. 22, no. 6, p. 665, 2015.

[182] A. T. Specht and J. Li, “Leap: constructing gene co-expression networks for single-cell rna-sequencing data using pseudotime ordering,” *Bioinformatics*, vol. 33, no. 5, pp. 764–766, 2017.

[183] T. Moerman, S. Aibar Santos, C. Bravo González-Blas, J. Simm, Y. Moreau, J. Aerts, and S. Aerts, “Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks,” *Bioinformatics*, vol. 35, no. 12, pp. 2159–2161, 2019.

[184] A. Deshpande, L.-F. Chu, R. Stewart, and A. Gitter, “Network inference with granger causality ensembles on single-cell transcriptomic data,” *BioRxiv*, p. 534834, 2019.

[185] N. Papili Gao, S. M. Ud-Dean, O. Gandrillon, and R. Gunawan, “Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles,” *Bioinformatics*, vol. 34, no. 2, pp. 258–266, 2018.

[186] P.-C. Aubin-Frankowski and J.-P. Vert, “Gene regulation inference from single-cell rna-seq data with linear differential equations and velocity inference,” *Bioinformatics*, vol. 36, no. 18, pp. 4774–4780, 2020.

[187] H. Matsumoto, H. Kiryu, C. Furusawa, M. S. Ko, S. B. Ko, N. Gouda, T. Hayashi, and I. Nikaido, “Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation,” *Bioinformatics*, vol. 33, no. 15, pp. 2314–2321, 2017.

[188] M. Sanchez-Castillo, D. Blanco, I. M. Tienda-Luna, M. Carrion, and Y. Huang, “A bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data,” *Bioinformatics*, vol. 34, no. 6, pp. 964–970, 2018.

[189] S. Woodhouse, N. Piterman, C. M. Wintersteiger, B. Göttgens, and J. Fisher, “Scns: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data,” *BMC systems biology*, vol. 12, no. 1, pp. 1–7, 2018.

[190] X. Qiu, A. Rahimzamani, L. Wang, B. Ren, Q. Mao, T. Durham, J. L. McFaline-Figueroa, L. Saunders, C. Trapnell, and S. Kannan, “Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe,” *Cell systems*, vol. 10, no. 3, pp. 265–274, 2020.

[191] P. Dibaeinia and S. Sinha, “Sergio: a single-cell expression simulator guided by gene regulatory networks,” *Cell Systems*, vol. 11, no. 3, pp. 252–271, 2020.

[192] D. Osorio, X. Yu, Y. Zhong, G. Li, E. Serpedin, J. Z. Huang, and J. J. Cai, “Single-cell expression variability implies cell function,” *Cells*, vol. 9, no. 1, p. 14, 2020.

[193] W. H. Beasley and J. L. Rodgers, “Resampling methods,” *The Sage handbook of quantitative methods in psychology*, pp. 362–386, 2009.

[194] M. G. Kendall, “A course in multivariate analysis,” tech. rep., 1957.

[195] M. Baburaj and S. N. George, “Reweighted low-rank tensor decomposition based on t-svd and its applications in video denoising,” *arXiv preprint arXiv:1611.05963*, 2016.

[196] L. Yuan, Q. Zhao, L. Gui, and J. Cao, “High-order tensor completion via gradient-based optimization under tensor train format,” *Signal Processing: Image Communication*, vol. 73, pp. 53–61, 2019.

[197] C. Battaglino, G. Ballard, and T. G. Kolda, “A practical randomized cp tensor decomposition,” *SIAM Journal on Matrix Analysis and Applications*, vol. 39, no. 2, pp. 876–901, 2018.

[198] K. R. Moon, J. S. Stanley III, D. Burkhardt, D. van Dijk, G. Wolf, and S. Krishnaswamy, “Manifold learning-based methods for analyzing single-cell rna-sequencing data,” *Current Opinion in Systems Biology*, vol. 7, pp. 36–46, 2018.

[199] H. Vu, C. Carey, and S. Mahadevan, “Manifold warping: Manifold alignment over time,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, 2012.

[200] C. Wang and S. Mahadevan, “A general framework for manifold alignment.,” in *AAAI fall symposium: manifold learning and its applications*, pp. 79–86, 2009.

[201] N. D. Nguyen, I. K. Blaby, and D. Wang, “Maninetcluster: a novel manifold learning approach to reveal the functional links between gene networks,” *BMC genomics*, vol. 20, no. 12, pp. 1–14, 2019.

[202] F. Diaz and D. Metzler, “Pseudo-aligned multilingual corpora.,” in *IJCAI*, pp. 2727–2732, 2007.

[203] C. Wang and S. Mahadevan, “Manifold alignment preserving global geometry.,” in *IJCAI*, pp. 1743–1749, Citeseer, 2013.

[204] R. C. Wilson and P. Zhu, “A study of graph spectra for comparing graphs and trees,” *Pattern Recognition*, vol. 41, no. 9, pp. 2833–2841, 2008.

[205] M. A. Skinnider, J. W. Squair, C. Kathe, M. A. Anderson, M. Gautier, K. J. Matson, M. Milano, T. H. Hutson, Q. Barraud, A. A. Phillips, *et al.*, “Cell type prioritization in single-cell data,” *Nature Biotechnology*, vol. 39, no. 1, pp. 30–34, 2021.

[206] J. L. Rodgers, “The bootstrap, the jackknife, and the randomization test: A sampling taxonomy,” *Multivariate Behavioral Research*, vol. 34, no. 4, pp. 441–456, 1999.

[207] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edger: a bioconductor package for differential expression analysis of digital gene expression data,” *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[208] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, “Bayesian approach to single-cell differential expression analysis,” *Nature methods*, vol. 11, no. 7, pp. 740–742, 2014.

[209] C. Soneson and M. D. Robinson, “Bias, robustness and scalability in single-cell differential expression analysis,” *Nature methods*, vol. 15, no. 4, p. 255, 2018.

[210] D. Avey, S. Sankararaman, A. K. Yim, R. Barve, J. Milbrandt, and R. D. Mitra, "Single-cell rna-seq uncovers a robust transcriptional response to morphine by glia," *Cell reports*, vol. 24, no. 13, pp. 3619–3629, 2018.

[211] L. T. Kagohara, F. Zamuner, E. F. Davis-Marcisak, G. Sharma, M. Considine, J. Allen, S. Yegnasubramanian, D. A. Gaykalova, and E. J. Fertig, "Integrated single-cell and bulk gene expression and atac-seq reveals heterogeneity and early changes in pathways associated with resistance to cetuximab in hnscc-sensitive cell lines," *British journal of cancer*, vol. 123, no. 1, pp. 101–113, 2020.

[212] D. R. Little, K. N. Gerner-Mauro, P. Flodby, E. D. Crandall, Z. Borok, H. Akiyama, S. Kimura, E. J. Ostrin, and J. Chen, "Transcriptional control of lung alveolar type 1 cell development and maintenance by nk homeobox 2-1," *Proceedings of the National Academy of Sciences*, vol. 116, no. 41, pp. 20545–20555, 2019.

[213] Y. Zhou, W. M. Song, P. S. Andhey, A. Swain, T. Levy, K. R. Miller, P. L. Poliani, M. Cominelli, S. Grover, S. Gilfillan, *et al.*, "Human and mouse single-nucleus transcriptomics reveal trem2-dependent and trem2-independent cellular responses in alzheimer's disease," *Nature medicine*, vol. 26, no. 1, pp. 131–142, 2020.

[214] D. S. Goodsell, "The molecular perspective: morphine," *The oncologist*, vol. 9, no. 6, pp. 717–718, 2004.

[215] P. H. Tso and Y. H. Wong, "Molecular basis of opioid dependence: role of signal regulation by g-proteins.," *Clinical and experimental pharmacology & physiology*, vol. 30, no. 5-6, pp. 307–316, 2003.

[216] M. Jalabert, R. Bourdy, J. Courtin, P. Veinante, O. J. Manzoni, M. Barrot, and F. Georges, "Neuronal circuits underlying acute morphine action on dopamine neurons," *Proceedings of the national academy of sciences*, vol. 108, no. 39, pp. 16446–16450, 2011.

[217] W. Krz, N. Ghyselinck, T. A. Samad, V. Dupé, P. Kastner, E. Borrelli, P. Chambon, *et al.*, "Impaired locomotion and dopamine signaling in retinoid receptor mutant mice," *Science*,

vol. 279, no. 5352, pp. 863–867, 1998.

[218] M. Tafti and N. B. Ghyselinck, “Functional implication of the vitamin a signaling pathway in the brain,” *Archives of neurology*, vol. 64, no. 12, pp. 1706–1711, 2007.

[219] H. Morikawa and C. A. Paladini, “Dynamic regulation of midbrain dopamine neuron activity: intrinsic, synaptic, and plasticity mechanisms,” *Neuroscience*, vol. 198, pp. 95–111, 2011.

[220] S. Johnson and R. North, “Opioids excite dopamine neurons by hyperpolarization of local interneurons,” *Journal of neuroscience*, vol. 12, no. 2, pp. 483–488, 1992.

[221] A. Laakso, A. R. Mohn, R. R. Gainetdinov, and M. G. Caron, “Experimental genetic approaches to addiction,” *Neuron*, vol. 36, no. 2, pp. 213–228, 2002.

[222] K.-S. Kim, K.-W. Lee, K.-W. Lee, J.-Y. Im, J. Y. Yoo, S.-W. Kim, J.-K. Lee, E. J. Nestler, and P.-L. Han, “Adenylyl cyclase type 5 (ac5) is an essential mediator of morphine action,” *Proceedings of the National Academy of Sciences*, vol. 103, no. 10, pp. 3908–3913, 2006.

[223] M. Korostynski, M. Piechota, D. Kaminska, W. Solecki, and R. Przewlocki, “Morphine effects on striatal transcriptome in mice,” *Genome biology*, vol. 8, no. 6, pp. 1–17, 2007.

[224] S. K. Blick and L. J. Scott, “Cetuximab,” *Drugs*, vol. 67, no. 17, pp. 2585–2607, 2007.

[225] J. Harding and B. Burtness, “An epidermal growth factor receptor chimeric human-murine monoclonal antibody,” *Drugs Today (Barc)*, vol. 41, pp. 107–127, 2005.

[226] B. Vincenzi, A. Zoccoli, F. Pantano, O. Venditti, and S. Galluzzo, “Cetuximab: from bench to bedside,” *Current cancer drug targets*, vol. 10, no. 1, p. 80, 2010.

[227] R. S. Herbst and D. M. Shin, “Monoclonal antibodies to target epidermal growth factor receptor–positive tumors: a new paradigm for cancer therapy,” *Cancer*, vol. 94, no. 5, pp. 1593–1611, 2002.

[228] B. Burtness, “The role of cetuximab in the treatment of squamous cell cancer of the head and neck,” *Expert opinion on biological therapy*, vol. 5, no. 8, pp. 1085–1093, 2005.

[229] T. J. Desai, D. G. Brownfield, and M. A. Krasnow, "Alveolar progenitor and stem cells in lung development, renewal and cancer," *Nature*, vol. 507, no. 7491, pp. 190–194, 2014.

[230] D. H. Tompkins, V. Besnard, A. W. Lange, A. R. Keiser, S. E. Wert, M. D. Bruno, and J. A. Whitsett, "Sox2 activates cell proliferation and differentiation in the respiratory epithelium," *American journal of respiratory cell and molecular biology*, vol. 45, no. 1, pp. 101–110, 2011.

[231] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri, "Fast, sensitive and accurate integration of single-cell data with harmony," *Nature methods*, vol. 16, no. 12, pp. 1289–1296, 2019.

[232] M. Li, W. Shillinglaw, W. J. Henzel, and A. A. Beg, "The rela (p65) subunit of nf-κb is essential for inhibiting double-stranded rna-induced cytotoxicity," *Journal of Biological Chemistry*, vol. 276, no. 2, pp. 1185–1194, 2001.

[233] N. Kopitar-Jerala, "The role of interferons in inflammation and inflammasome activation," *Frontiers in immunology*, vol. 8, p. 873, 2017.

[234] M. P. Gantier and B. R. Williams, "The response of mammalian cells to double-stranded rna," *Cytokine & growth factor reviews*, vol. 18, no. 5-6, pp. 363–371, 2007.

[235] H. B. Levy, L. W. Law, and A. S. Rabson, "Inhibition of tumor growth by polyinosinic-polycytidylic acid," *Proceedings of the National Academy of Sciences*, vol. 62, no. 2, pp. 357–361, 1969.

[236] H. Oakley, S. L. Cole, S. Logan, E. Maus, P. Shao, J. Craft, A. Guillozet-Bongaarts, M. Ohno, J. Disterhoft, L. Van Eldik, *et al.*, "Intraneuronal β-amyloid aggregates, neurodegeneration, and neuron loss in transgenic mice with five familial alzheimer's disease mutations: potential factors in amyloid plaque formation," *Journal of Neuroscience*, vol. 26, no. 40, pp. 10129–10140, 2006.

[237] M.-S. Tan, J.-T. Yu, and L. Tan, "Bridging integrator 1 (bin1): form, function, and alzheimer's disease," *Trends in molecular medicine*, vol. 19, no. 10, pp. 594–603, 2013.

[238] C. J. Holler, P. R. Davis, T. L. Beckett, T. L. Platt, R. L. Webb, E. Head, and M. P. Murphy, "Bridging integrator 1 (bin1) protein expression increases in the alzheimer's disease brain and correlates with neurofibrillary tangle pathology," *Journal of Alzheimer's Disease*, vol. 42, no. 4, pp. 1221–1227, 2014.

[239] M. Ximerakis, S. L. Lipnick, B. T. Innes, S. K. Simmons, X. Adiconis, D. Dionne, B. A. Mayweather, L. Nguyen, Z. Niziolek, C. Ozek, *et al.*, "Single-cell transcriptomic profiling of the aging mouse brain," *Nature neuroscience*, vol. 22, no. 10, pp. 1696–1708, 2019.

[240] L. Kester and A. van Oudenaarden, "Single-cell transcriptomics meets lineage tracing," *Cell stem cell*, vol. 23, no. 2, pp. 166–179, 2018.

[241] X. Zheng, Y. Huang, and X. Zou, "scpadgrn: A preconditioned admm approach for reconstructing dynamic gene regulatory network using single-cell rna sequencing data," *PLoS computational biology*, vol. 16, no. 7, p. e1007471, 2020.

[242] X. Ma, P. Sun, and G. Qin, "Identifying condition-specific modules by clustering multiple networks," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 5, pp. 1636–1648, 2017.

[243] X. Ma, D. Dong, and Q. Wang, "Community detection in multi-layer networks using joint nonnegative matrix factorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 2, pp. 273–286, 2018.

[244] D. Osorio, Y. Zhong, G. Li, J. Z. Huang, and J. J. Cai, "sctenifoldnet: a machine learning workflow for constructing and comparing transcriptome-wide gene regulatory networks from single-cell data," *Patterns*, vol. 1, no. 9, p. 100139, 2020.

[245] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, *et al.*, "String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic acids research*, vol. 47, no. D1, pp. D607–D613, 2019.

[246] H. Van Hove, L. Martens, I. Scheyltjens, K. De Vlaminck, A. R. P. Antunes, S. De Prijck, N. Vandamme, S. De Schepper, G. Van Isterdael, C. L. Scott, *et al.*, "A single-cell atlas of mouse brain macrophages reveals unique transcriptional identities shaped by ontogeny and tissue environment," *Nature neuroscience*, vol. 22, no. 6, pp. 1021–1035, 2019.

[247] H. Han, L. A. Davidson, Y.-Y. Fan, J. S. Goldsby, G. Yoon, U.-H. Jin, G. A. Wright, K. K. Landrock, B. R. Weeks, R. C. Wright, *et al.*, "Loss of aryl hydrocarbon receptor potentiates foxm1 signaling to enhance self-renewal of colonic stem and progenitor cells," *The EMBO Journal*, vol. 39, no. 19, p. e104319, 2020.

[248] G. L. Szot, P. Koudria, and J. A. Bluestone, "Murine pancreatic islet isolation," *Journal of visualized experiments: JoVE*, no. 7, 2007.

[249] J. Chen, L. Zhong, J. Wu, S. Ke, B. Morpurgo, A. Golovko, N. Ouyang, Y. Sun, S. Guo, and Y. Tian, "A murine pancreatic islet cell-based screening for diabetogenic environmental chemicals," *JoVE (Journal of Visualized Experiments)*, no. 136, p. e57327, 2018.

[250] B. Marquina-Sanchez, N. Fortelny, M. Farlik, A. Vieira, P. Collombat, C. Bock, and S. Kubicek, "Single-cell rna-seq with spike-in cells enables accurate quantification of cell-specific drug effects in pancreatic islets," *Genome biology*, vol. 21, pp. 1–22, 2020.

[251] A. A. Nugent, K. Lin, B. Van Lengerich, S. Lianoglou, L. Przybyla, S. S. Davis, C. Llapashtica, J. Wang, D. Xia, A. Lucas, *et al.*, "Trem2 regulates microglial cholesterol metabolism upon chronic phagocytic challenge," *Neuron*, vol. 105, no. 5, pp. 837–854, 2020.

[252] L. Chen, N. H. Toke, S. Luo, R. P. Vasoya, R. L. Fullem, A. Parthasarathy, A. O. Perekatt, and M. P. Verzi, "A reinforcing hnf4–smad4 feed-forward module stabilizes enterocyte identity," *Nature genetics*, vol. 51, no. 5, pp. 777–785, 2019.

[253] D. B. Frank, I. J. Penkala, J. A. Zepp, A. Sivakumar, R. Linares-Saldana, W. J. Zacharias, K. G. Stolz, J. Pankin, M. Lu, Q. Wang, *et al.*, "Early lineage specification defines alveolar

epithelial ontogeny in the murine lung," *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4362–4371, 2019.

[254] A. B. Rubenstein, G. R. Smith, U. Raue, G. Begue, K. Minchev, F. Ruf-Zamojski, V. D. Nair, X. Wang, L. Zhou, E. Zaslavsky, *et al.*, "Single-cell transcriptional profiles in human skeletal muscle," *Scientific reports*, vol. 10, no. 1, pp. 1–15, 2020.

[255] A. Zeisel, H. Hochgerner, P. Lönnerberg, A. Johnsson, F. Memic, J. Van Der Zwan, M. Häring, E. Braun, L. E. Borm, G. La Manno, *et al.*, "Molecular architecture of the mouse nervous system," *Cell*, vol. 174, no. 4, pp. 999–1014, 2018.

[256] S. J. Attarian, S. L. Leibel, P. Yang, D. N. Alfano, B. P. Hackett, F. S. Cole, and A. Hamvas, "Mutations in the thyroid transcription factor gene nkx2-1 result in decreased expression of sftpb and sftpc," *Pediatric research*, vol. 84, no. 3, pp. 419–425, 2018.

[257] Y. Shi and D. M. Holtzman, "Interplay between innate immunity and alzheimer disease: Apoe and trem2 in the spotlight," *Nature Reviews Immunology*, vol. 18, no. 12, pp. 759–772, 2018.

[258] R. Guerreiro, A. Wojtas, J. Bras, M. Carrasquillo, E. Rogaeva, E. Majounie, C. Cruchaga, C. Sassi, J. S. Kauwe, S. Younkin, *et al.*, "Trem2 variants in alzheimer's disease," *New England Journal of Medicine*, vol. 368, no. 2, pp. 117–127, 2013.

[259] Y. Atagi, C.-C. Liu, M. M. Painter, X.-F. Chen, C. Verbeeck, H. Zheng, X. Li, R. Rademakers, S. S. Kang, H. Xu, *et al.*, "Apolipoprotein e is a ligand for triggering receptor expressed on myeloid cells 2 (trem2)," *Journal of Biological Chemistry*, vol. 290, no. 43, pp. 26043–26050, 2015.

[260] P. L. Poliani, Y. Wang, E. Fontana, M. L. Robinette, Y. Yamanishi, S. Gilfillan, M. Colonna, *et al.*, "Trem2 sustains microglial expansion during aging and response to demyelination," *The Journal of clinical investigation*, vol. 125, no. 5, pp. 2161–2170, 2015.

[261] D. A. Jaitin, L. Adlung, C. A. Thaiss, A. Weiner, B. Li, H. Descamps, P. Lundgren, C. Bleriot, Z. Liu, A. Deczkowska, *et al.*, "Lipid-associated macrophages control metabolic homeostasis in a trem2-dependent manner," *Cell*, vol. 178, no. 3, pp. 686–698, 2019.

[262] J.-P. Babeu, M. Darsigny, C. R. Lussier, and F. Boudreau, "Hepatocyte nuclear factor 4α contributes to an intestinal epithelial phenotype in vitro and plays a partial role in mouse intestinal epithelium differentiation," *American Journal of Physiology-Gastrointestinal and Liver Physiology*, vol. 297, no. 1, pp. G124–G134, 2009.

[263] C. Carter, "Monogenic disorders.," *Journal of medical genetics*, vol. 14, no. 5, p. 316, 1977.

[264] D. H. Andersen and R. G. HODGES, "Celiac syndrome: V. genetics of cystic fibrosis of the pancreas with a consideration of etiology," *American journal of diseases of children*, vol. 72, no. 1, pp. 62–80, 1946.

[265] F. S. Collins, "Cystic fibrosis: molecular biology and therapeutic implications," *Science*, vol. 256, no. 5058, pp. 774–779, 1992.

[266] F. Liu, Z. Zhang, L. Csanády, D. C. Gadsby, and J. Chen, "Molecular structure of the human cftr ion channel," *Cell*, vol. 169, no. 1, pp. 85–95, 2017.

[267] K. Yoshimura, H. Nakamura, B. C. Trapnell, C.-S. Chu, W. Dakemans, A. Pavirani, J.-P. Lecocq, and R. G. Crystal, "Expression of the cystic fibrosis transmembrane conductance regulator gene in cells of non-epithelial origin," *Nucleic acids research*, vol. 19, no. 19, pp. 5417–5423, 1991.

[268] D. A. Stoltz, D. K. Meyerholz, and M. J. Welsh, "Origins of cystic fibrosis lung disease," *New England Journal of Medicine*, vol. 372, no. 4, pp. 351–362, 2015.

[269] Z. Lin, N. Thorenoor, R. Wu, S. L. DiAngelo, M. Ye, N. J. Thomas, X. Liao, T. R. Lin, S. Warren, and J. Floros, "Genetic association of pulmonary surfactant protein genes, sftpa1, sftpa2, sftpb, sftpc, and sftpd with cystic fibrosis," *Frontiers in immunology*, vol. 9, p. 2256, 2018.

[270] C. Von Bredow, A. Wiesener, and M. Griese, "Proteolysis of surfactant protein d by cystic fibrosis relevant proteases," *Lung*, vol. 181, no. 2, pp. 79–88, 2003.

[271] F. Bühling, M. Kouadio, C. E. Chwieralski, U. Kern, J. M. Hohlfeld, N. Klemm, N. Friedrichs, W. Roth, J. M. Deussing, C. Peters, *et al.*, "Gene targeting of the cysteine peptidase cathepsin h impairs lung surfactant in mice," *PloS one*, vol. 6, no. 10, p. e26247, 2011.

[272] R. D. Cohn and K. P. Campbell, "Molecular basis of muscular dystrophies," *Muscle & nerve*, vol. 23, no. 10, pp. 1456–1471, 2000.

[273] C. Brinkmeyer-Langford, C. Chu, C. Balog-Alvarez, X. Yu, J. J. Cai, M. Nabity, and J. N. Kornegay, "Expression profiling of disease progression in canine model of duchenne muscular dystrophy," *PLoS One*, vol. 13, no. 3, p. e0194485, 2018.

[274] K. P. Campbell, "Three muscular dystrophies: loss of cytoskeleton-extracellular matrix linkage," *Cell*, vol. 80, no. 5, pp. 675–679, 1995.

[275] E. P. Hoffman, R. H. Brown Jr, and L. M. Kunkel, "Dystrophin: the protein product of the duchenne muscular dystrophy locus," *Cell*, vol. 51, no. 6, pp. 919–928, 1987.

[276] J. L. Neul, W. E. Kaufmann, D. G. Glaze, J. Christodoulou, A. J. Clarke, N. Bahi-Buisson, H. Leonard, M. E. Bailey, N. C. Schanen, M. Zappella, *et al.*, "Rett syndrome: revised diagnostic criteria and nomenclature," *Annals of neurology*, vol. 68, no. 6, pp. 944–950, 2010.

[277] R. E. Amir, I. B. Van den Veyver, M. Wan, C. Q. Tran, U. Francke, and H. Y. Zoghbi, "Rett syndrome is caused by mutations in x-linked mecp2, encoding methyl-cpg-binding protein 2," *Nature genetics*, vol. 23, no. 2, pp. 185–188, 1999.

[278] X. Nan, F. J. Campoy, and A. Bird, "Mecp2 is a transcriptional repressor with abundant binding sites in genomic chromatin," *Cell*, vol. 88, no. 4, pp. 471–481, 1997.

[279] M. J. Lyst and A. Bird, "Rett syndrome: a complex disorder with simple roots," *Nature Reviews Genetics*, vol. 16, no. 5, pp. 261–275, 2015.

[280] L. Abuhatzira, K. Makedonski, Y. Kaufman, A. Razin, and R. Shemer, “Mecp2 deficiency in the brain decreases bdnf levels by rest/corest-mediated repression and increases trkb production,” *Epigenetics*, vol. 2, no. 4, pp. 214–222, 2007.

[281] M. V. Johnston, O.-H. Jeon, J. Pevsner, M. E. Blue, and S. Naidu, “Neurobiology of rett syndrome: a genetic disorder of synapse development,” *Brain and Development*, vol. 23, pp. S206–S213, 2001.

[282] L. Medrihan, E. Tantalaki, G. Aramuni, V. Sargsyan, I. Dudanova, M. Missler, and W. Zhang, “Early defects of gabaergic synapses in the brain stem of a mecp2 mouse model of rett syndrome,” *Journal of neurophysiology*, vol. 99, no. 1, pp. 112–121, 2008.

[283] H.-T. Chao, H. Chen, R. C. Samaco, M. Xue, M. Chahrour, J. Yoo, J. L. Neul, S. Gong, H.-C. Lu, N. Heintz, *et al.*, “Dysfunction in gaba signalling mediates autism-like stereotypies and rett syndrome phenotypes,” *Nature*, vol. 468, no. 7321, pp. 263–269, 2010.

[284] R. Romaniello, F. Saettini, E. Panzeri, F. Arrigoni, M. T. Bassi, and R. Borgatti, “A denovo stxbp1 gene mutation in a patient showing the rett syndrome phenotype,” *Neuroreport*, vol. 26, no. 5, pp. 254–257, 2015.

[285] A. Metidji, S. Omenetti, S. Crotta, Y. Li, E. Nye, E. Ross, V. Li, M. R. Maradana, C. Schiering, and B. Stockinger, “The environmental sensor ahr protects from inflammatory damage by maintaining intestinal stem cell homeostasis and barrier integrity,” *Immunity*, vol. 49, no. 2, pp. 353–362, 2018.

[286] K. Kawajiri and Y. Fujii-Kuriyama, “The aryl hydrocarbon receptor: a multifunctional chemical sensor for host defense and homeostatic maintenance,” *Experimental animals*, vol. 66, no. 2, pp. 75–89, 2017.

[287] B. Stockinger, P. D. Meglio, M. Gialitakis, and J. H. Duarte, “The aryl hydrocarbon receptor: multitasking in the immune system,” *Annual review of immunology*, vol. 32, pp. 403–432, 2014.

[288] P. Ramadoss, C. Marcus, and G. H. Perdew, "Role of the aryl hydrocarbon receptor in drug metabolism," *Expert opinion on drug metabolism & toxicology*, vol. 1, no. 1, pp. 9–21, 2005.

[289] C.-I. Ko, Q. Wang, Y. Fan, Y. Xia, and A. Puga, "Pluripotency factors and polycomb group proteins repress aryl hydrocarbon receptor expression in murine embryonic stem cells," *Stem cell research*, vol. 12, no. 1, pp. 296–308, 2014.

[290] D. R. Hipfner and S. M. Cohen, "Connecting proliferation and apoptosis in development and disease," *Nature Reviews Molecular Cell Biology*, vol. 5, no. 10, pp. 805–815, 2004.

[291] H. Samimi, V. Haghpanah, S. Irani, E. Arefian, A. N. Sohi, P. Fallah, and M. Soleimani, "Transcript-level regulation of malat1-mediated cell cycle and apoptosis genes using dual mek/aurora kinase inhibitor "bi-847325" on anaplastic thyroid carcinoma," *DARU Journal of Pharmaceutical Sciences*, vol. 27, no. 1, pp. 1–7, 2019.

[292] Y. Sun and L. Ma, "New insights into long non-coding rna malat1 in cancer and metastasis," *Cancers*, vol. 11, no. 2, p. 216, 2019.

[293] Q. Wang, G. Lu, and Z. Chen, "Malat1 promoted cell proliferation and migration via malat1/mir-155/mef2a pathway in hypoxia of cardiac stem cells," *Journal of cellular biochemistry*, vol. 120, no. 4, pp. 6384–6394, 2019.

[294] J. Chen, S. Ke, L. Zhong, J. Wu, A. Tseng, B. Morpurgo, A. Golovko, G. Wang, J. J. Cai, X. Ma, *et al.*, "Long noncoding rna malat1 regulates generation of reactive oxygen species and the insulin responses in male mice," *Biochemical pharmacology*, vol. 152, pp. 94–103, 2018.

[295] P. Puthanveetil, S. Chen, B. Feng, A. Gautam, and S. Chakrabarti, "Long non-coding rna malat 1 regulates hyperglycaemia induced inflammatory process in the endothelial cells," *Journal of cellular and molecular medicine*, vol. 19, no. 6, pp. 1418–1425, 2015.

[296] H. Zhang, W. Li, W. Gu, Y. Yan, X. Yao, and J. Zheng, “Malat1 accelerates the development and progression of renal cell carcinoma by decreasing the expression of mir-203 and promoting the expression of birc5,” *Cell proliferation*, vol. 52, no. 5, p. e12640, 2019.

[297] P. Malakar, A. Shilo, A. Mogilevsky, I. Stein, E. Pikarsky, Y. Nevo, H. Benyamini, S. Elgavish, X. Zong, K. V. Prasanth, *et al.*, “Long noncoding rna malat1 promotes hepatocellular carcinoma development by srsf1 upregulation and mtor activation,” *Cancer research*, vol. 77, no. 5, pp. 1155–1167, 2017.

[298] P. Malakar, I. Stein, A. Saragovi, R. Winkler, N. Stern-Ginossar, M. Berger, E. Pikarsky, and R. Karni, “Long noncoding rna malat1 regulates cancer glucose metabolism by enhancing mtor-mediated translation of tcf7l2,” *Cancer research*, vol. 79, no. 10, pp. 2480–2493, 2019.

[299] C. Yan, J. Chen, and N. Chen, “Long noncoding rna malat1 promotes hepatic steatosis and insulin resistance by increasing nuclear srebp-1c protein stability,” *Scientific reports*, vol. 6, no. 1, pp. 1–11, 2016.

[300] H. Liu, Z. Zhang, W. Xiong, L. Zhang, Y. Du, Y. Liu, and X. Xiong, “Long non-coding rna malat 1 mediates hypoxia-induced pro-survival autophagy of endometrial stromal cells in endometriosis,” *Journal of cellular and molecular medicine*, vol. 23, no. 1, pp. 439–452, 2019.

[301] R. Sun, C. Qin, B. Jiang, S. Fang, X. Pan, L. Peng, Z. Liu, W. Li, Y. Li, and G. Li, “Down-regulation of malat1 inhibits cervical cancer cell invasion and metastasis by inhibition of epithelial–mesenchymal transition,” *Molecular BioSystems*, vol. 12, no. 3, pp. 952–962, 2016.

[302] X. Cai, Y. Liu, W. Yang, Y. Xia, C. Yang, S. Yang, and X. Liu, “Long noncoding rna malat1 as a potential therapeutic target in osteosarcoma,” *Journal of Orthopaedic Research*, vol. 34, no. 6, pp. 932–941, 2016.

[303] M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields, and C. Trapnell, “Dimensionality reduction by umap to visualize physical and genetic interactions,” *Nature communications*,

vol. 11, no. 1, pp. 1–6, 2020.

[304] H. Lund, M. Pieber, R. Parsa, J. Han, D. Grommisch, E. Ewing, L. Kular, M. Needhamsen, A. Espinosa, E. Nilsson, *et al.*, “Competitive repopulation of an empty microglial niche yields functionally distinct subsets of microglia-like cells,” *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.

[305] Q. Li, Z. Cheng, L. Zhou, S. Darmanis, N. F. Neff, J. Okamoto, G. Gulati, M. L. Bennett, L. O. Sun, L. E. Clarke, *et al.*, “Developmental heterogeneity of microglia and brain myeloid cells revealed by deep single-cell rna sequencing,” *Neuron*, vol. 101, no. 2, pp. 207–223, 2019.

[306] T. R. Hammond, C. Dufort, L. Dissing-Olesen, S. Giera, A. Young, A. Wysoker, A. J. Walker, F. Gergits, M. Segel, J. Nemesh, *et al.*, “Single-cell rna sequencing of microglia throughout the mouse lifespan and in the injured brain reveals complex cell-state changes,” *Immunity*, vol. 50, no. 1, pp. 253–271, 2019.

[307] E. F. Willis, K. P. MacDonald, Q. H. Nguyen, A. L. Garrido, E. R. Gillespie, S. B. Harley, P. F. Bartlett, W. A. Schroder, A. G. Yates, D. C. Anthony, *et al.*, “Repopulating microglia promote brain repair in an il-6-dependent manner,” *Cell*, vol. 180, no. 5, pp. 833–846, 2020.

[308] V. Bortnov, D. S. Annis, F. J. Fogerty, K. T. Barretto, K. B. Turton, and D. F. Mosher, “Myeloid-derived growth factor is a resident endoplasmic reticulum protein,” *Journal of Biological Chemistry*, vol. 293, no. 34, pp. 13166–13175, 2018.

[309] V. Bortnov, M. Tonelli, W. Lee, Z. Lin, D. S. Annis, O. N. Demerdash, A. Bateman, J. C. Mitchell, Y. Ge, J. L. Markley, *et al.*, “Solution structure of human myeloid-derived growth factor suggests a conserved function in the endoplasmic reticulum,” *Nature communications*, vol. 10, no. 1, pp. 1–14, 2019.

[310] M. Korf-Klingebiel, M. R. Reboll, S. Klede, T. Brod, A. Pich, F. Polten, L. C. Napp, J. Bauersachs, A. Ganser, E. Brinkmann, *et al.*, “Myeloid-derived growth factor (c19orf10)

mediates cardiac repair following myocardial infarction," *Nature medicine*, vol. 21, no. 2, pp. 140–149, 2015.

[311] Y. Wang, Y. Li, J. Feng, W. Liu, Y. Li, J. Liu, Q. Yin, H. Lian, L. Liu, and Y. Nie, "Mydgf promotes cardiomyocyte proliferation and neonatal heart regeneration," *Theranostics*, vol. 10, no. 20, p. 9100, 2020.

[312] R. Huang, I. Grishagin, Y. Wang, T. Zhao, J. Greene, J. C. Obenauer, D. Ngan, D.-T. Nguyen, R. Guha, A. Jadhav, *et al.*, "The ncats bioplanet–an integrated platform for exploring the universe of cellular signaling pathways for toxicology, systems biology, and chemical genomics," *Frontiers in pharmacology*, vol. 10, p. 445, 2019.

[313] H. Sunagozaka, M. Honda, T. Yamashita, R. Nishino, H. Takatori, K. Arai, T. Yamashita, Y. Sakai, and S. Kaneko, "Identification of a secretory protein c19orf10 activated in hepatocellular carcinoma," *International journal of cancer*, vol. 129, no. 7, pp. 1576–1585, 2011.

[314] B. P. Busby, E. Niktab, C. A. Roberts, J. P. Sheridan, N. V. Coorey, D. S. Senanayake, L. M. Connor, A. B. Munkacsi, and P. H. Atkinson, "Genetic interaction networks mediate individual statin drug response in saccharomyces cerevisiae," *NPJ systems biology and applications*, vol. 5, no. 1, pp. 1–13, 2019.

[315] M. Lotfollahi, F. A. Wolf, and F. J. Theis, "scgen predicts single-cell perturbation responses," *Nature methods*, vol. 16, no. 8, pp. 715–721, 2019.

[316] K. Kamimoto, C. M. Hoffmann, and S. A. Morris, "Celloracle: Dissecting cell identity via network inference and in silico gene perturbation," *BioRxiv*, 2020.