Novel Statistics and Machine Learning Analysis in Bioinformatics and Big Data

by

Mengrui Zhang

(Under the Direction of Ping Ma)

Abstract

With the rapid development of science and technology, large and complex data have been generated in many biological science areas, such as single-cell RNA-Seq, neuroscience and human dynamics. However, the task of analyzing big data itself poses significant challenges. On the one hand, the ultra-large size of datasets renders the application of many statistical methods computationally impossible. On the other hand, with the system being studied getting more complicated, the model setup for some popular off-the-shelf methods may not be applicable anymore. Developing new theoretically justifiable and computationally efficient methods for tackling big data problems from a computational and modeling perspective is the primary motivation for my research. The proposed methods can be widely applied to various scientific disciplines and greatly help scientific development.

Index words: Big Data Analysis, Bioinformatics, Human Dynamics, Single Cell

Novel Statistics and Machine Learning Analysis in Bioinformatics and Big Data

by

Mengrui Zhang

B.S., University of California, Santa Barbara, 2015

M.A., University of California, Santa Barbara, 2016

A Dissertation Submitted to the Graduate Faculty of the

University of Georgia in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2022

©2022

Mengrui Zhang

All Rights Reserved

Novel Statistics and Machine Learning Analysis in Bioinformatics and Big Data

by

Mengrui Zhang

Major Professor:

Ping Ma

Committee:

Wenxuan Zhong

Magdy Alabady

Yuan Ke

Electronic Version Approved:

Ron Walcott

Dean of the Graduate School

The University of Georgia

May 2022

Acknowledgments

This work was made possible through the support of many people. I cannot thank my advisor Prof. Ping Ma enough, who has generously provided me with assistance, guidance, and support throughout my Ph.D. study. I have been greatly impressed and influenced by his insights, wisdom and passion for statistics. I would like to express my deepest gratitude to him. Meanwhile, I also want to express my sincere gratitude to my committee members: Prof. Wenxuan Zhong, Prof. Magdy Alabady and Prof. Yuan Ke, for their advice, help, and encouragement on my dissertation research. It is my honor to have them all on my committee.

I want to express my special thanks to Prof. Magdy Alabady at Georgia Genomics and Bioinformatics Core (GGBC). I wouldn't be able to be skilled in bioinformatics without your help. In addition, I want to appreciate all my collaborators and co-authors in my journal papers. This work would not be possible without their excellent collaboration and guidance.

I want to express my thanks to all labmates of Big Data Analytics Lab (BDAL) for their supports: Xiaoxiao Sun, Yiwen Liu, Xin Xing, Xinlian Zhang, Wei Xu, Rui Xie, Cheng Meng, Jingyi Zhang, Ye Wang, Huimin Cheng, Jun Yu, Yongkai Chen, Shushan Wu, Zhen Wang, Haoran Lu, Jiazhang Cai, Luyang Fang and many others whose names are not listed. It was fantastic to have the opportunity to work with these young researchers in the past five years. I would also like to express my genuine gratitude to department head Prof. T.N. Sriram, Prof. Liang Liu, Prof. Pengsheng Ji, Prof. Jaxk Reeves and all members in the Department of Statistics and at the University of Georgia. They have truly influenced me through their extraordinary research experiences. Additionally, I benefited tremendously from intellectual discussions with colleagues from Dr. Walt Lorenz, Dr. Zengyan Wang and all members at the Georgia

Genomics and Bioinformatics Core (GGBC). Without your help, I wouldn't be able to learn so many things in bioinformatics.

Part of the dissertation research is supported by U.S. National Science Foundation grants DMS-1903226, DMS-1925066, DMS-2124493 and NIH grants R01GM1222080a to Prof. Ma and Pro. Zhong.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the rapid development of science and technology, large and complex data have been generated in many areas, such as genomics, neuroscience, and social science. The extraordinary amount of data land for artificial intelligence and revolutionize our conventional decision making system. This new phenomenon posts great challenges on current statistical research. For example, the ultra-large size of dataset renders the application of many statistical methods computationally infeasible. Developing new theoretically justifiable and computationally efficient methods for tackling big data problems from a computational and modeling perspective is the primary goal for my research. To achieve my goals in this thesis,

1. In Chapter 2, I reviewed existing methods on cell lineage tracing and trajectory inference and discuss some machine learning applications in single-cell RNA-Seq dataset.
2. In Chapter 3, I developed an innovative machine learning method called Cell Smoothing Transformation (CellST) to elucidate dynamic individual cell behaviors and dynamic gene expression patterns in cell progression processes.
3. In Chapter 4, I developed a few statistical/machine learning analysis pipelines with application in bioinformatics and human dynamics.

A review of modern single-cell dynamic lineage analysis methods with integration in machine learning

Understanding the heterogeneous and stochastic nature of multicellular tissues is crucial for studies in developmental biology. Researchers in recent years are particularly interested in dynamic cell lineage and fates analysis, which includes processes such as cell reprogramming, differentiation, and morphological development. The rising of single-cell sequencing technology offers unprecedented insights into the functionality and development of complex individual cell behaviors. Since scRNA-seq permits comparison of the transcriptomes of individual cells, effective use of scRNA-seq has been used to assess transcriptional similarities and differences within a population of cells, with early reports revealing previously unappreciated levels of heterogeneity of individual cells, for example, in embryonic and immune cells. Thus, the heterogeneity cell lineage analysis become a core reason for conducting scRNA-seq studies. In this chapter, I conduct a detailed review of existing cell lineage methods and cell fate prediction. In addition, I discuss some challenges and a possible way to observe individual cells' dynamic behaviors through optimal transport.

Elucidation of cell progression and dynamic gene networks in time course single-cell RNA-seq data

Over time, the heterogeneity in individual cell progression becomes a core reason for conducting single-cell RNA-Seq (scRNA-seq) studies. Cell lineage tracing has been widely used in cell progression processes such as cell reprogramming and differentiation. Researchers construct cell trajectories by ordering cells chronologically to represent trace cell lineages in time-course experiments. However, cells in time-course experiments are sacrificed and sequenced at each time point. Thus there are no cell correspondences between time points, which creates a significant challenge to elucidate cell lineages. Additionally, we only observe discrete cell information at each time point to construct the cell lineages, which is a continuous process. Therefore, methods that can reconstruct the cell continuation are highly desirable.

In this chapter, I develop an innovative machine learning method called Cell Smoothing Transformation (CellST) to elucidate dynamic cell lineage behaviors and dynamic gene expression patterns in cell progression. I provide extensive simulations and real scRNA-seq studies on the process including cell reprogramming and differentiation. The proposed method provides massive numbers of cell trajectories in parallel for individual cells. Those trajectories are highly accurate in reflecting dynamic cell progression behaviors. Moreover, based on the individual cell trajectories, the constructed dynamic gene networks accurately model gene-gene relationships and discover critical genes for cell lineage tracing. For cells with various cell types, unlike the bulk cell trajectory, individual cell trajectories construct unique trajectories and observe individual cell lineages even for cells with less frequent cell types.

Novel Analysis Pipelines with Applications in Bioinformatics, Human Dynamic Big Data and Biological Behaviors Dataset

With the rapid development of science and technology, large and complex data have been generated in many science areas especially with biological science and human health care. In this chapter, I propose a series of novel analysis pipelines that focus to elucidate the complex large scale data generated from wearable device that measures human dynamic as well as modern biological dataset including locomotor behavior data and Next-Generation Sequencing (NGS) dataset.

Locomotor Behavior Data The locomotor behavior data of zebrafish is of high-throughput, time-related and involves both experimental and biological variables. Its systematic studies have provided new insights into neurobiology, pharmacology, and toxicology. I have established a coherent statistical analysis framework for analyzing such data. In this section, I compared the time-related behavior profiles of zebrafish in several commonly-used scenarios.

Human Dynamic Data The human dynamic trajectories data collected through wearable devices contain continuously precise GPS and physical activities. Such data can be used to study the dynamic patterns of human behavior. We propose a two-layer graph convolutional network (GCN) framework for graph classification on the data and achieve $> 85\%$ on the testing accuracy. Moreover, I utilize the

smoothing spline models to explore the activity differences between office-based worker and non-office based worker.

Chapter 2

Modern Single Cell Dynamic Lineage Analysis with Machine Learning

2.1 Introduction on single cell dynamics

Understanding the heterogeneous and stochastic nature of multicellular tissues is crucial for studies in developmental biology. Researchers in recent years are particularly interested in dynamic cell lineage and fates analysis, which includes processes such as cell reprogramming, differentiation, and morphological development(Burrows et al., 2020; Guo et al., 2017). One interesting question in this research field is to elucidate the behaviors of stem cells differentiate into the myriad diverse cell types that ultimately form the multicellular tissues (Moris et al., 2016; Plass et al., 2018; Sipp et al., 2018). However, unfortunately, studying this process at the population level masks rare or transient intermediates. In recent years, single cell sequencing technology's rising offers unprecedented insights into the functionality and development of complex individual cell behaviors. Since scRNA-seq permits comparison of the transcriptomes of individual cells, effective use of scRNA-seq has been used to assess transcriptional similarities and differences

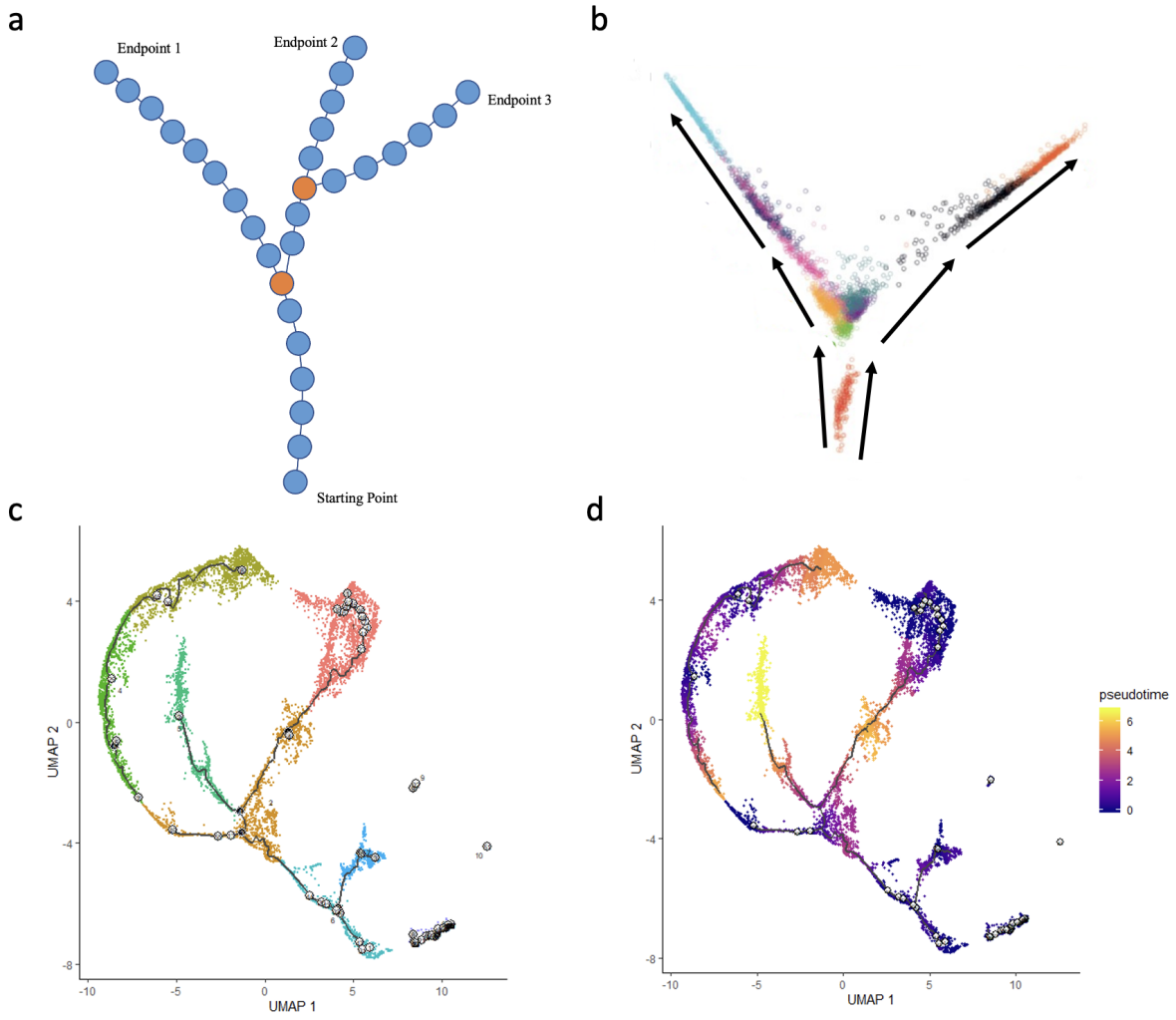

Figure 2.1: a: A multi-potent stem cell can give rise to multiple cell fate endpoints. **b:** Example stem cell progression process with simple 2 branches. **c & d:** Example of Cell lineage tracing and prediction on a cat stem cell therapy.

within a population of cells, with early reports revealing previously unappreciated levels of heterogeneity of individual cells, for example, in embryonic and immune cells (Zhou et al., 2020). Thus, the heterogeneity cell lineage analysis become a core reason for conducting scRNA-seq studies.

To appropriate conduct the cell lineage analysis, researchers must define how cells change through time and map the paths they take during differentiation. For example, cells in cell differentiation change state by undergoing gradual transcriptional changes, with progress being driven by an underlying temporal

variable or pseudotime. To observe such process, lineage tracing (Wagner & Klein, 2020), the technique of following a cell or group of cells and observing their descendants, is an important tool for defining the fate potential of cells and detecting the outcomes of differentiation. The single cell lineage tracing can be modeled computationally using trajectory inference methods, which order cells along a trajectory based on similarities in their gene expression patterns. Nowadays, lineage tracing has become a powerful approach for detecting individual cellular variation within a complex population, and the analysis results can be applied to identify a broad range of cell types and states in cells growth (Davis et al., 2016; Simeonov et al., 2021). For example, (Rodriguez-Fraticelli et al., 2020) uses single cell lineage tracing to investigate the life-long regenerative capacity of hematopoietic stem cells (HSCs). Their framework elucidates clone-intrinsic molecular programs associated with functional stem cell heterogeneity and identifies a mechanism for maintaining the self-renewing HSC state. (Ludwig et al., 2019) in their paper also showed that somatic mutations in mtDNA can be tracked by single cell RNA or assay for transposase accessible chromatin (ATAC) sequencing.

The single cell RNA-Seq experiments can be classified into static single cell experiments and time-course single cell experiments. In a static single cell experiment, all cells were observed simultaneously. In other words, the static single cell experiment is equivalent to taking a snapshot of all cells and their gene expressions at one time point (Hrvatin et al., 2018; Lawson et al., 2015), whereas the time course scRNA-seq experiments takes snapshots at multiple time points. There are already many existing analysis methods published in recent years. In general, those lineage prediction tools order cells along the axis of differentiation based on progressive changes in gene expression. In some cases, branch points in the lineage trajectories can also be predicted (Figure 2.1). Panel **a** and **b** in figure 2.1 are illustrations of stem cells development that a stem cell can possibly develop into multiple cell types in the end (panel **a**). A simple case to study would be the stem cell development process with only two branches (panel **b**). In real-world single cell experiments, panel **c** and **d** in figure 2.1 provides an examples of cat stem cell therapy process. This example is an static single cell experiment and the cell lineage trajectories were predicted using Monocle3 package in R (Trapnell et al., 2014). Panel **c** illustrates multiple cell clusters labeled with different

colors along the trajectories, and panel **d** shows the cells labeled with predicted pseudo cell development time. The clusters can potentially represents different cell developing stages over pseudo-time.

The cell lineage tracing is conducted differently in static and time-course single cell experiments. In static scRNA-seq experiments, one natural approach to predict cell lineage is to order cells into a continuous cell trajectory by constructing a pseudo time to order cells chronologically (Cannoodt, Saelens & Saeys, 2016; Chen et al., 2019; Ji & Ji, 2016; Liu et al., 2017; Qiu et al., 2017; Trapnell, 2015; Trapnell et al., 2014). However, there are some limitations when applying those methods to predict the cell fates (Tritschler et al., 2019). The time-course scRNA-seq experiments contains intrinsically much more informative than the static scRNA-seq data, particularly for the inference of cellular dynamic development patterns (An et al., 2019; Ko et al., 2020; Sun et al., 2021; Torii et al., 2020; Yuan & Bar-Joseph, 2021). However, cells are sacrificed and sequenced at each time point. Thus there is no connecting information for cells between two time points, which creates a significant challenge to elucidate the dynamic behaviors of cell progression. Moreover, the sequenced cells cannot be carried to the next time point, and the cell-cell variation is too large to be ignored. It is also challenging to align and register different cells sequenced in two adjacent time points since cell variation affects gene expression drastically affected by cell variation (Alonge et al., 2020; Ren et al., 2017). Without controlling the cell variation, gene expression analysis can be significantly biased.

In this review paper, we conduct a detailed review of existing cell lineage methods and cell fate prediction. We also discuss some challenges and a possible way to observe individual cells' dynamic behaviors through optimal transport. The rest of the paper is organized as follows: we start in section 2 by reviewing the up-to-date analysis of static single cell RNA-Seq experiments, which constructs the cell progression and order cell by pseudotime. In Section 3, we review the up-to-date analysis of time-course single cell RNA-Seq experiments. The cell progression fates can be constructed using cells' experimental time points. Section 4 introduces a new concept on elucidating cell lineage progress using optimal transport in the machine learning field. In Section 5, we show several applications of machine learning methods in real-world single cell experiments in developmental biology.

2.2 Methods on static single cell RNA-Seq experiments

In this section, we review some major existing cell lineage analysis methods and the methods development path over the past few years. After the deconvolution of complex tissues and cluster/group cells into different and multiple cell types based on scRNA-seq, cells in experiments can be analyzed using cell lineage prediction or trajectory inference tools. Those tools order cells along a pseudonym axis, which could potentially represent dynamic cell lineage over processes such as cell differentiation. In general, most of the existing methods are based on progressive changes in gene expression. These methods usually take reduced dimension gene expression data as an input (e.g., after using principal component analysis) or nearest-neighbor graph representations and attempt to infer the branching lineage trajectory structure and order cells along the trajectories. In some cases, cells could develop into multiple branch points, and some of the existing methods can predict such branched lineage trajectories as well (Fletcher et al., 2017; Gadye et al., 2017; Nowakowski et al., 2017). We use the publicly available single cell RNA-Seq dataset (PBMC dataset) from the 10X genomics database. The PBMC dataset is Peripheral blood mononuclear cells (PBMCs) from a healthy donor. This dataset contains 3,000 cells with different cell types such as Naive CD4+T, CD8+T and FCGR3A+Mono. Different cell types have different marker genes, which are critical in the cell clustering process.

The first iterations of trajectory prediction algorithms were capable of ordering cells along a single trajectory. Still, they were largely unable to accommodate branching lineage trajectories where a stem cell gives rise to more than one lineage or cell type (figure 2.2) (Stévant et al., 2018; Treutlein et al., 2016). Scorpius (Cannoodt, Saelens, Sichien et al., 2016) is a popular method to predict cell lineage with no branches. This method assumes that the given dataset contains the gene expression profiles of hundreds to thousands of cells, which were uniformly sampled from a cell linear dynamic process. The Scorpius construct a single cell trajectory by clustering the data with k-means clustering and finding the shortest path through the cluster centers. Then this initial trajectory is subsequently refined in an iterative way using a principal curves algorithm. The individual cells can then be ordered by projecting the n-dimensional

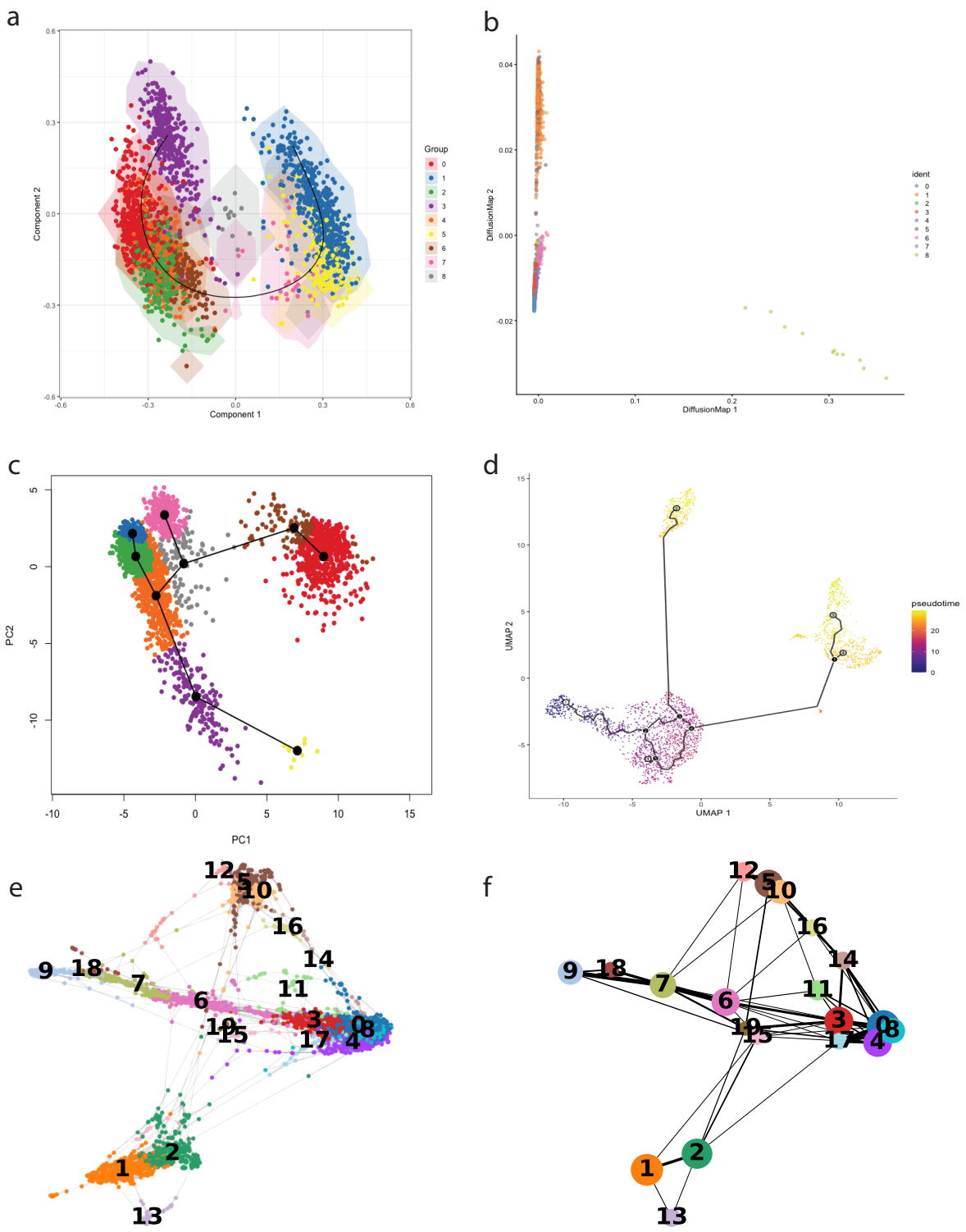

Figure 2.2: Major methods of cell lineage prediction on static single cell expression data (PBMC 3k data). **a:** Linear prediction method (SCORPIUS) on the dataset. **b:** Bifurcation cell lineages were predicted using diffusion map. **c & d:** We use tree methods Slingshot and Monocle3 to prediction cell lineages. Those tree type method provide significantly more cell dynamic information than linear and bifurcation type. **e & f:** The last type is graph method (PAGA). Cell clusters were connected with trajectories.10

points onto the trajectory. In conclusion, the scorpius method assumes a linear transformation between cell at different time (Figure 2.2 **a**).

Subsequent methods attempted to predict where the branch points in trajectories occur. Bifurcating trajectory is another type of cell lineage prediction method. Bifurcating trajectory analysis uses diffusion map (Haghverdi et al., 2015; Haghverdi et al., 2016) for dimension reduction process and construct cell lineage trajectory based on diffusion pseudotime (Figure 2.2 **b**). As there are many existing dimension reduction tools available, existing trajectory inference methods become a variety, which is not limited to typically fixed the topology algorithmic. More trajectory method types are arriving, such as tree-based and graph-based methods. Slingshot and Monocle are two major trajectory inference methods for tree-based analysis. Those methods can accommodate to predict cells with more than two developing branches. Slingshot takes a normalized expression matrix as an input. Based on the number of cell clusters or states, i.e., disjoint subsets of cells, which are typically obtained by clustering the cells based on their gene expression measures, slingshot defines a lineage as an ordered set of clusters and output the total number of lineages (Figure 2.2 **c**). Specifically, slingshot identifies lineages by treating clusters of cells as nodes in a graph and drawing a minimum spanning tree (MST) between the nodes. Similarly, the Monocle method uses DDRTree, a scalable RGE algorithm, to learn a principal tree on a population of single cells. This tree was built based on global gene expression changes inside a cell development through the biological process. Both Slingshot and Monocle are recommended to build based on the reduced dimension spaces.

However, predicting dynamic cell lineage using the previously mentioned trajectory method faces the problem that experimental data do not conform with a connected manifold. Therefore, the linear, bifurcation and tree structures have little meaning in predicting the biological cell lineage. Moreover, those methods might be making the invalid assumption that clusters conform with a connected tree-like topology and rely on feature-space based inter-cluster distances, like the euclidean distance of cluster means. Such distance measures quantify the biological similarity of cells only at a local scale and are fraught with problems when used for larger-scale objects like clusters. Partition-based graph abstraction (PAGA) resolves these fundamental problems by generating graph-like maps of cells that preserve both continuous

and disconnected structures in data at multiple resolutions. PAGA method constructs a symmetrized kNN-like graph using the approximate nearest neighbor search within UMAP (Becht et al., 2019; McInnes et al., 2020) dimension reduction. Moreover, PAGA can reconstruct branching gene expression changes across different datasets.

In conclusion, despite the effectiveness, these methods may fail in the following circumstances (Tritschler et al., 2019). First of all, most existing trajectory inference methods construct a bulk cell trajectory, i.e., the mean trajectory of the population cells across time rather than that of individual cells. However, some individual cells' behaviors may oscillate up and down around the cell mean expressions or severely deviate from it. Cell progression behaviors are dominated by cells with major cell types, and patterns with less frequent might be hidden in the dataset. Second, individual cell developing trajectories may follow different complex topologies, including loops or alternative paths during the development. Those methods may introduce a significant bias and are hard to validate, as cells are ordered based only on the selected reduced dimensions. Moreover, regardless of the specific approach, the existing methods rely upon the assumption that cells that are more similar in gene expression are closer together on a lineage trajectory. While this is a reasonable assumption, there are situations where cell fate transitions represent more saltatory changes in gene expression rather than subtle changes along a continuum (Kester & van Oudenaarden, 2018). Lastly, they also rely upon a second assumption that the paths are unidirectional, which presents difficulties in modeling stem cell self-renewal.

2.3 Methods on time course single cell RNA-Seq experiment

Time-course scRNA-seq experiments contain intrinsically much more informative than the static scRNA-seq data, particularly for the prediction of dynamic cell lineage (An et al., 2019; Ko et al., 2020; Sun et al., 2021; Torii et al., 2020; Yuan & Bar-Joseph, 2021). Different than static single cell expression data that cells' development time was computed as pseudo-time, the time-course single cell expression data have experimental time for individual cells. Using time-course single cell data has become increasingly popular in cell lineage analysis. Researchers are interested in the development of cells during a specific period.

While many single cell RNA-seq trajectory inference methods for static experiments exist, few have been designed to consider time-series information for analysis time-course single cell data (Ko et al., 2020; Tran & Bader, 2020; Yuan & Bar-Joseph, 2021). Waddington-OT is a trajectory inference method that explicitly incorporates temporal information. This method models cells' movement through dynamic processes using the optimal transport framework, and CSHMM, which uses a continuous-state hidden Markov model to assign cells to developmental paths (Lin & Bar-Joseph, 2019; Schiebinger et al., 2019). Tempora is another cell trajectory inference method that orders cells using time information from time-series scRNA-seq data. This method orders cells at different time points based on the established assumption that cells with similar gene expression profiles are closer in the underlying cell lineage. Figure 2.3 illustrate a format of time-course single cell experiment data and a time-course trajectory computed by Tempora (Figure 2.3 **c**). We use the Murine cerebral cortex dataset (MouseCortex) that contains approximately 6,000 neural cells collected at embryonic days 11.5 (E11.5), E13.5, E15.5, and E17.5. Cells were sequenced using DropSeq, and these cells cover a wide spectrum of neuronal development, from the early precursors (apical and radial precursors) to intermediate progenitors and differentiated cortical neurons. We observed that the Tempora method constructs a coarse cell trajectory that only connects the cell clusters at each developing stage. Moreover, we also compare the Tempora method with the static method Monocle, and we validate that the time-course trajectory has a more accurate prediction in modeling dynamic cell lineage. The trajectories constructed by methods on static single cell expression are generally based on the reduced cell space. Utilizing the time-course information provides a direction that enhances this cell space's trajectory estimation.

In conclusion, the behavior and change of cell-level behaviors along experimental time points have not been extensively studied yet since the time-course cell trajectory only connects at the cell cluster level. A major challenge for using such data to infer gene relationships is the fact that the dynamic gene expression over multiple time points cannot be tracked. Thus, it is not clear which cell in the next time point is a descendent (or closely related) to a specific cell in the previous time point, making it hard to determine exact trajectories for genes. Most of the existing analyses only focus on cell distribution at each time point

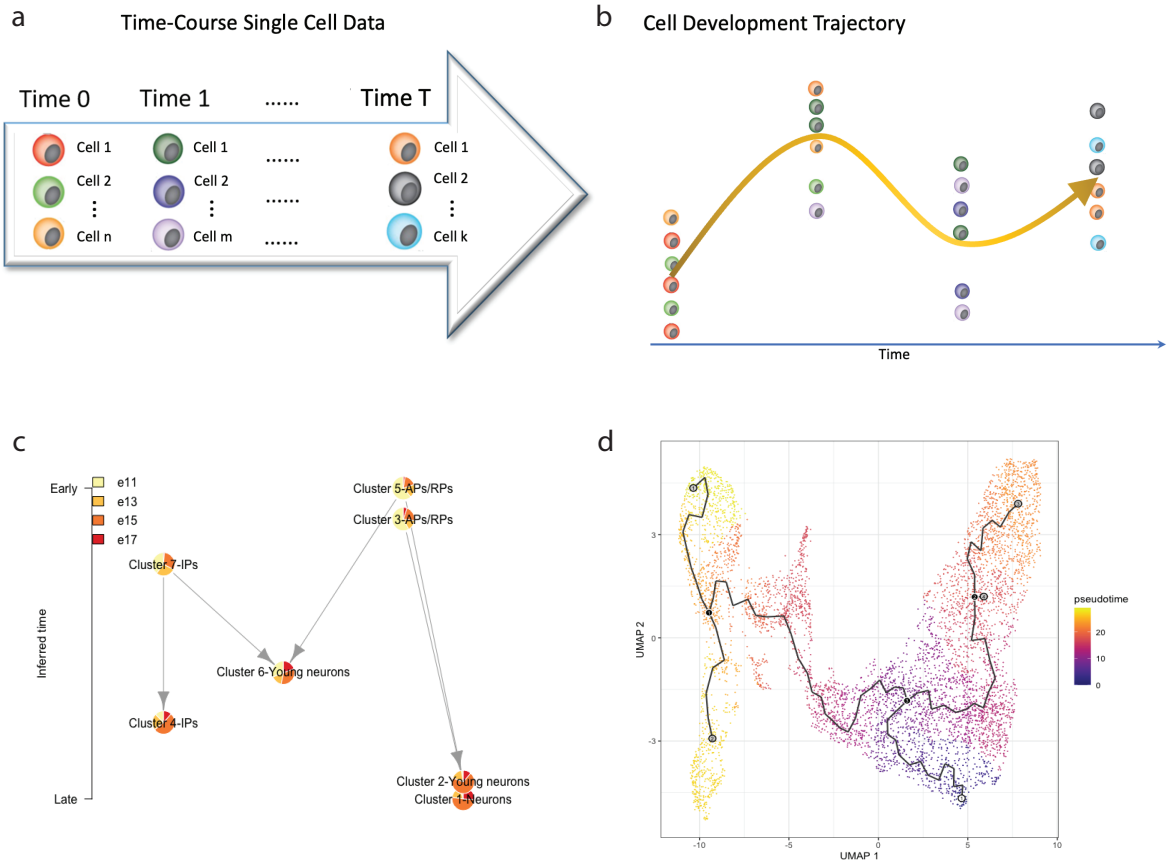

Figure 2.3: a: An illustration of time-course scRNA-seq data. Cells contain no information between time points. **b:** Sample cells development trajectory along the time points. **b:** Time-course trajectory built by Tempora method. **d:** Implementation of Monocle method on time-course single cell dataset.

individually. There is a lack of analysis methods that mainly focus on the effect of time on cell development and gene expression. Additionally, there is a lack of analyses in cell-level developments for a longer period. Another challenge arises from the large number of cells being profiled at each time point. Finally, the fact that cells in a time point may be from several different types and may not be fully synchronized makes it harder to establish a specific pattern for temporal analysis. In the later section, we will propose ideas and directions for constructing individual cell trajectories in a time-course single cell RNA-Seq dataset.

2.4 Optimal transport in machine learning

The optimal transport map (OTM) recently drew great attention in machine learning and statistics because OT theory can be used for computing distances between probability distributions. Nowadays, there also are many different ways to calculate the optimal transport map, even in high-dimensional space (Meng et al., 2019; Paty et al., 2020). The Waddington-OT (Schiebinger et al., 2019) provides a similar way of modeling cell dynamics at multiple time points. However, utilizing the optimal transport in observing cell dynamics has different approaches. From our point of view, the optimal transport technique can be used to fill the gap between any single cell samples. For example, the optimal transport can construct individual cell-to-cell connecting in multiple cell batches, replicates, or time points. In the next section, we will show application examples using optimal transport to observe dynamic cell behaviours.

In this section, we first introduce the essential background of optimal transport. As a powerful tool to transform one probability measure to another, optimal transport methods recently find extensive applications in machine learning (Alvarez-Melis et al., 2018; Arjovsky et al., 2017; Canas & Rosasco, 2012; Courty et al., 2016; Flamary et al., 2016; Flamary et al., 2018; Meng et al., 2019; Paty & Cuturi, 2019; Peyré, Cuturi et al., 2019; Redko et al., 2019; Wang, Zhang et al., 2020; Wang, Zhou et al., 2020; Zhao, Wang, Wu et al., 2020; Zhao, Wang, Zhang et al., 2020), statistics (Cazelles et al., 2018; Courty et al., 2017; Cuturi et al., 2019; Del Barrio et al., 2019; Flamary et al., 2019; Meng et al., 2020; Panaretos & Zemel, 2019; Paty et al., 2020; Zhang et al., 2020), computer vision (Deshpande et al., 2019; Ferradans et al., 2014; Peyré, Cuturi et al., 2019; Rabin et al., 2014; Su et al., 2015; Wu et al., 2019), among others.

Let $\mathscr{P}(\mathbb{R}^p)$ be the set of Borel probability measures in $\mathbb{R}^p$, and let

$$\mathscr{P}_2(\mathbb{R}^p) = \left\{\mu \in \mathscr{P}(\mathbb{R}^p) \middle| \int ||x||^2 d\mu(x) < \infty\right\}.$$

Let $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^p)$ be the probability measures and $\#$ be the pullback operator, such that $\phi_{\#}(\mu)(\Omega) = \mu(\phi^{-1}(\Omega))$ for any measurable set $\Omega \subset \mathbb{R}^p$. Denote Φ as the set of all the measure-preserving map $\phi : \mathbb{R}^p \to \mathbb{R}^p$, such that $\phi_{\#}(\mu) = \nu$ and $\phi_{\#}^{-1}(\nu) = \mu$. Among all the maps in Φ, the optimal one

respecting to the L_2 transport cost is defined as

$$\phi^* = \underset{\phi \in \Phi}{\operatorname{arginf}} \int_{\mathbb{R}^p} \|x - \phi(x)\|^2 \mathrm{d}\mu(x), \tag{2.1}$$

where $\|\cdot\|$ is the Euclidean norm. The minimizer of Equation (2.1) is usually called the optimal transport map, or the Monge map. One limitation for the Monge map is that, it may be infeasible in some extreme cases, say, when μ is a Dirac measure but ν is not. Kantorovich overcame such an limitation by considering the following set of "couplings" (Kantorovich, 2006; Kantorovitch, 1958),

$$\begin{aligned}\Pi(\mu, \nu) = \{\pi \in \mathscr{P}(\mathbb{R}^p \times \mathbb{R}^p) \quad s.t. \quad \forall \quad \text{Borel set} \quad A, B \subset \mathbb{R}^p, \\ \pi(A \times \mathbb{R}^p) = \mu(A), \quad \pi(\mathbb{R}^p \times B) = \nu(B)\}.\end{aligned}$$

Kantorovich formulated the optimal transport problem as finding the optimal joint probability measure π from $\Pi(\mu, \nu)$,

$$\pi^* = \underset{\pi \in \Pi(\mu, \nu)}{\operatorname{arginf}} \int \|x - y\|^2 \mathrm{d}\pi(x, y). \tag{2.2}$$

The minimizer of Equation (2.2) is called the optimal transport plan or the optimal coupling. Consider the cases when both μ and ν are continuous probability measures defined on a compact set, and both have continuous densities respecting the Lebesgue measure. In such cases, the well-known Brenier theorem (Brenier, 1991) guarantees the existence of the Monge map ϕ^*, and shows ϕ^* is equivalent to the optimal transport plan π^*.

2.5 Observing dynamic individual cell lineages using optimal transport

In living tissues, there are a large number of cell types with the assumption that each cell type has a distinct lineage and function. However, recent evidence from studies of single cells reveals that this assumption is incorrect. Cells may be morphologically and genetically identical but are actually heterogeneous, made up of individual cells that differ dramatically. These differences can have important consequences for the health and function of the entire population. The single cell analysis allows the study of cell-to-cell variation within a cell population (organ, tissue, and cell culture). In order to study diseases and drug development, in-depth analysis of stem cell differentiation, cancer, physiological functions in embryos and adults can only be accomplished with single cell analysis. Moreover, its discriminatory ability allows researchers to identify rare cell types in a larger population that would be obscured in bulk level analyses. As we reviewed, cell trajectory inference is an appropriate tool for estimating cell lineage/fates for stem cell differentiation, cancer, or physiological functions. However, most of the existing trajectory inference methods construct a bulk cell trajectory, i.e., the average trajectory respecting the population of cells across the timeline rather than individual ones. New methods such as RNA velocity and molecular recording are beginning to address this limitation by estimating dynamics from information obtained from the single cell measurement.

We define a cell distribution as the normalized gene expression matrix at each time points, denote as $X^{n\times d}$, where n indicates the number of cells and d indicates the number of genes in a single cell dataset. Genes are selected based on the analysis focus either using dimension reduction algorithms or known gene groups such as cell cycles and pluripotency gene groups. For each time point t_i, we have a cell distribution X related to it. Let two cell distributions at time t and time $t+1$ as X^t. The two cell distributions $X^t, X^{t+1} \in \mathbb{R}^d$, where d is the number of genes/features for cells. The optimal transport is define as $T : \mathbb{R}^d \to \mathbb{R}^d$ and the T is the optimal transport map. we calculate T to minimize a transportation cost

$C(\mathbf{T})$:

$$C(\mathbf{T}) = \int_{\mathbb{R}^d} c(\mathbf{x}, \mathbf{T}(\mathbf{x}))d\mu(\mathbf{x}) \tag{2.3}$$

where the $C(\mathbf{T})$ can be interpreted as the energy required to move a probability mass $\mu(\mathbf{x})$ from x to $\mathbf{T}(\mathbf{x})$. Let us define Π as the set of all probabilistic couplings $\in \mathcal{P}\left(\mathbb{R}^d \times \mathbb{R}^d\right)$ with marginals μ_t and μ_{t+1}. The Kantorovitch problem seeks for a general coupling $\gamma \in \Pi$ between $\mathbb{R}^d$ and $\mathbb{R}^d$:

$$\gamma_0 = \underset{\gamma \in \Pi}{\operatorname{argmin}} \int_{\mathbb{R}^d \times \mathbb{R}^d} c\left(\mathbf{x}_t, \mathbf{x}_{t+1}\right) d\gamma\left(\mathbf{x}_t, \mathbf{x}_{t+1}\right) \tag{2.4}$$

where γ_0 is known as transportation plan and coupling matrix. γ can be understood as a joint probability measure with marginals μ_t and μ_{t+1}. c is the cost function over distance between each cell coupling. Then the Kantorovich or Wasserstein distance for two cell distribution is written as:

$$W_p\left(\mu_t, \mu_{t+1}\right) \stackrel{\text{def}}{=} \left(\inf_{\gamma \in \Pi} \int_{\mathbb{R}^d \times \mathbb{R}^d} d\left(\mathbf{x}_t, \mathbf{x}_{t+1}\right)^p d\gamma\left(\mathbf{x}_t, \mathbf{x}_{t+1}\right)\right)^{\frac{1}{p}} \tag{2.5}$$

where $p \geq 1$. We solve the problem using regularized optimal transport method (Courty et al., 2016). We first denote $\mathcal{B}$ the set of probabilistic couplings between the two empirical distributions defined as:

$$\mathcal{B} = \left\{\gamma \in \left(\mathbb{R}^+\right)^{\mathbf{n}_t \times \mathbf{n}_{t+1}} | \gamma \mathbf{1}_{\mathbf{n}_t} = \mu_{\mathbf{t}}, \gamma^{\mathbf{T}} \mathbf{1}_{\mathbf{n}_s} = \mu_{\mathbf{t+1}}\right\} \tag{2.6}$$

where $\mathbf{1}_d$ is a d-dimensional vector of ones. Then in our case the Kantorovitch problem of optimal transport is:

$$\gamma_0 = \underset{\gamma \in \mathcal{B}}{\operatorname{argmin}} \langle \gamma, \mathbf{C} \rangle_F + \lambda \Omega_s(\gamma) \tag{2.7}$$

where the $\Omega_s(\gamma) = \sum_{i,j} \gamma(i,j) \log \gamma(i,j)$ interpreted as a Kullback-Leibler divergence($KL\left(\gamma \| \gamma_u\right)$). $< \ldots, \ldots >_F$ is the Frobenius inner product which equals to the sum of element-wise product for two matrices and $\mathbf{C} \geq 0$ is the cost function matrix, this cost was chosen as the squared Euclidean distance between the two points $C(i,j) = \|\mathbf{x}_i - \mathbf{y}_j\|_2^2$.

The constructed cell-to-cell connections between cells in different conditions or time points can potentially measure the similarity between cells. In a longitudinal dataset, these connections could predict cell lineages and cell behaviors. However, lineage inference for complex heterogeneous data still remains challenging. Compounding this is the lack of data for which the ground truth is known. In many cases, even the cell states are not known. This makes the validation of lineage relationships infeasible.

2.6 Individual cell lineage analysis on different scRNA-Seq dataset

2.6.1 Optimal transport in single cell batch effect

Large-scale single cell RNA sequencing (scRNA-seq) data sets that are produced in different laboratories and at different times contain batch effects that may compromise the integration and interpretation of the data. Batch effects can be highly nonlinear, making it difficult to align different datasets while preserving key biological variations correctly. More than ten analysis methods focus on the single cell batch effect correction. One existing batch effect correction method (Haghverdi et al., 2018) called "MNNs," which identifies mutual nearest neighbors to establish connections between two datasets. However, this approach is computationally demanding in terms of CPU time and memory because of the high dimension of genes. Existing faster analysis methods are only based on a low dimensional space to make connections between two datasets by applying a dimension reduction method beforehand (Butler et al., 2018; Hie et al., 2019; Korsunsky et al., 2019; Polański et al., 2020; Stuart et al., 2019).

In some examples of dimension reduction methods, they use techniques such as principal component analysis(PCA) (Jolliffe & Cadima, 2016), canonical correlation analysis (CCA)(Hardoon et al., 2004) and some more. However, there is a lack of methods that comprehensively include all the genes' features and achieve a good computation efficiency. Here, we present a strategy for batch effect correction based on the optimal transport map method. The optimal transport method helps us align the cells in a different

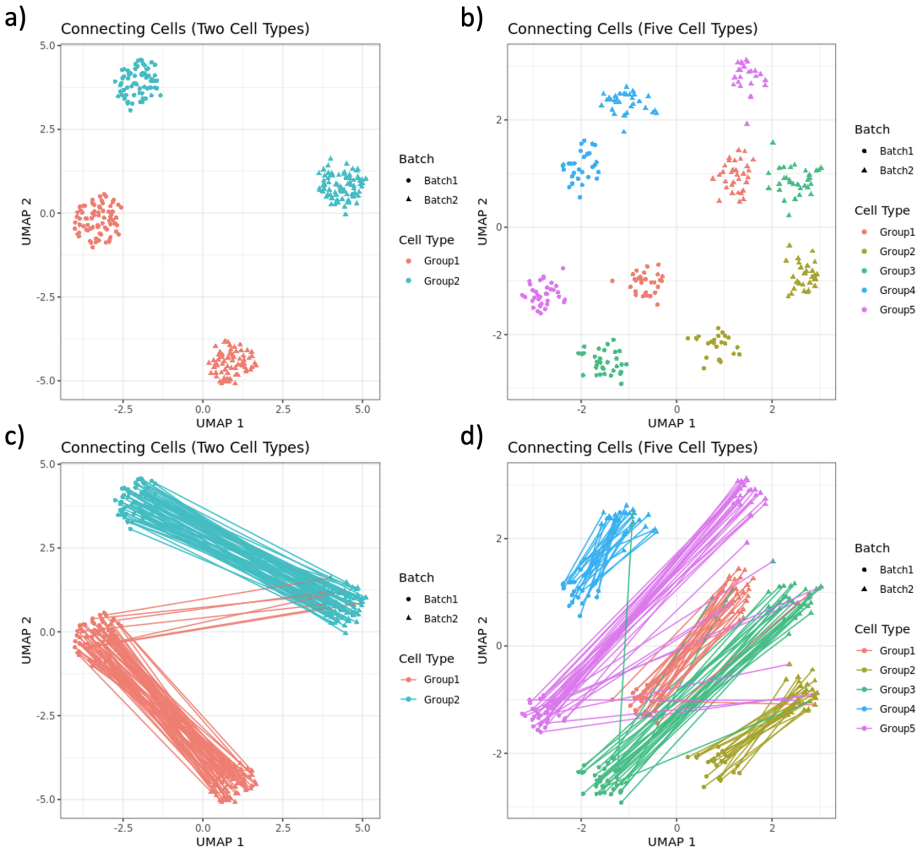

Figure 2.4: a)-b) Simulated single cell data with two different batches. a) Simulated two cell types in one batch. b) Simulated five cell types in one batch. c) - b) Built cell-to-cell paths to connect cells between two batches. c) Two cell types in one batch. d) Five cell types in one batch.

dataset with a batch effect. Our framework can comprehensively include all the genes' features in the original dimension space of the two datasets with a good computation time. In addition to it, we can also extend the resolution to a cell-to-cell coupling in two datasets. We aim to predict the one-to-one paired cells between two batches.

We present a simulated example of connecting cells between different batches. We test the accuracy of connecting cells for the same cell type between two batches. Correctly connecting cells with the same cell type indicates that the CellOT method can capture individual cells' gene and pathway expression features. This is crucial when cells start to develop into different types in a time-course dataset. The simulation

Table 2.1: Two cell types setting: The results shows the accuracy of connecting cell for two different batches with two cell types in each batch.

Two Cell Type					
	100 Cells	200 Cells	300 Cells	400 Cells	600 Cells
10 Genes	0.92	0.74	0.89	0.82	0.49
20 Genes	0.76	0.86	**0.95**	**0.98**	**0.99**
50 Genes	0.88	**0.96**	0.92	0.93	**0.99**
100 Genes	**1**	0.8	0.76	0.91	**0.99**
200 Genes	0.88	0.94	0.92	**0.95**	**0.99**
500 Genes	**0.96**	**0.96**	**0.95**	**0.95**	**0.99**

Table 2.2: Five cell types setting: The results shows the accuracy of connecting cell for two different batches with five cell types in each batch.

Five Cell Type					
	100 Cells	200 Cells	300 Cells	400 Cells	600 Cells
10 Genes	0.38	0.41	0.55	0.43	0.72
20 Genes	0.6	0.44	**0.87**	0.79	0.68
50 Genes	0.56	0.77	0.53	0.72	**0.85**
100 Genes	0.62	0.5	0.73	0.75	**0.89**
200 Genes	0.7	0.74	0.78	**0.83**	**0.87**
500 Genes	**0.86**	**0.8**	**0.9**	**0.86**	**0.83**

contains three different settings: two cell types, four cell types, and five cell types. We test cell number ranging from 100 cells to 600 cells, and the number of genes in each cell range from 10 to 500 genes for each of the three settings.

Table 2.1 and 2.2 presents the accuracy of connecting cells between two batches. The CellOT method achieves relatively high accuracy for all different cell type settings with 50 genes or more cells. The accuracy has an increasing trend as we put more genes in cells for the simulated data. Since more genes contain more information about each cell, the optimal transport algorithm will get more information to correct cell-to-cell connection paths. Therefore, The accuracy shows an increase from the tables. In addition to the above, there is also a slightly increasing trend as we simulate more cells for each cell type. Those cells can be treated as information replicates, in which more cells can enhance the gene features for a certain

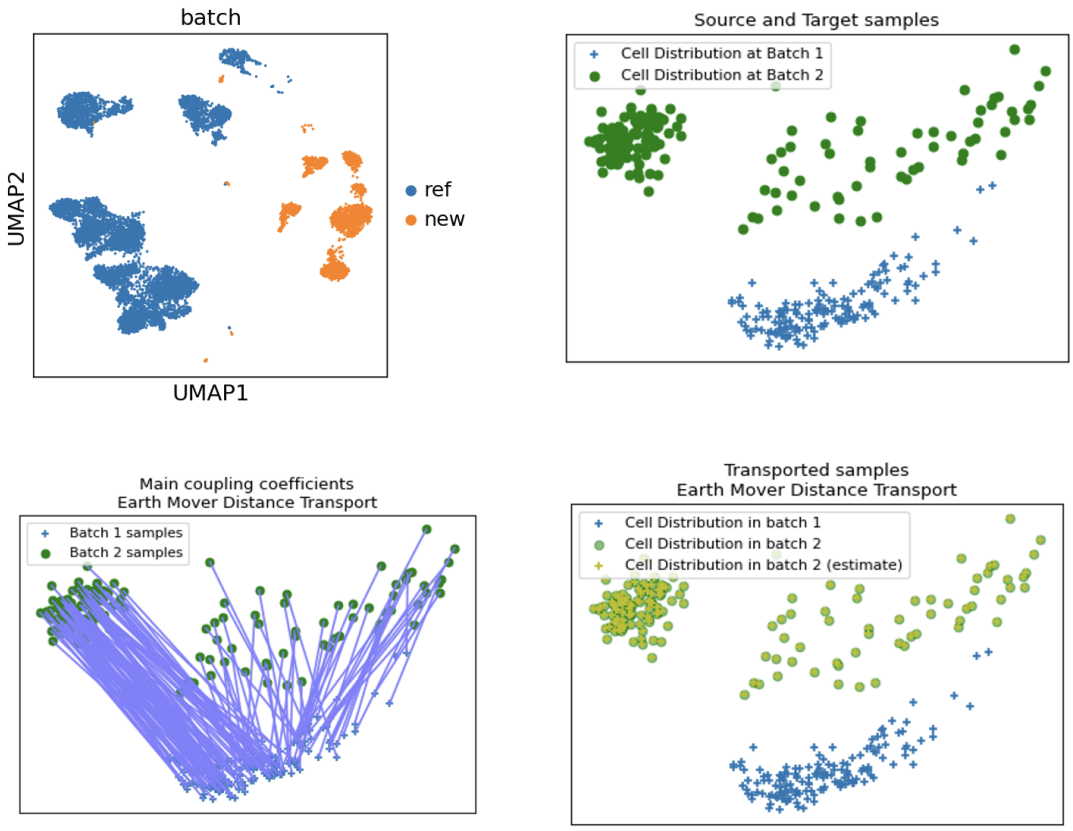

Figure 2.5: Single cell Optimal Transport on Batch Effect: a)-b) Cell type and cluster visualization for each batch. c) Batch effect visualization. d) Cell type visualization combining the two batches

cell type. Our proposed cell-to-cell connection can achieve 99% accuracy when there are only two cell types and $\sim 86\%$ when there are five cell types in each batch.

We present another real single cell expression dataset using optimal transport to construct cell-to-cell connections. The dataset we use here is a single cell RNA-seq dataset (cell = 14,693) on the pancreas and has four different batches (Lotfollahi et al., 2019). The data were downloaded from (ftp://ngs.sanger.ac.uk). The dataset in each batch can be represented as a cell distribution, including gene expression matrix and batch labels. We take the two batches with the most number of cells for our calculation. Figure 2.5 **a** and **b** are the two batches we use here to perform the batch effect alignment. These two figures show the cluster and cell distributions in the data. We notice that the number of cell type are different in the two batches.

We subset the datasets for batches and take an intersection of the two datasets. In this case, we can ensure our labels are consistent in the two batches and better show our cells' coupling accuracy. Figure 2.5 **c** and **d** are the cell coupling visualization using UMAP for the dataset. We subset only two clusters in each batch for easy visualization. The blue "+" points indicate the cell distribution in one batch, and the green "o" points indicate the cell distribution in other batches. Each cluster in the dataset represents one cell type. The figure 2.5 **d** indicates a good cell coupling accuracy that most of our cells in one batch are correctly paired with cells in another batch with the correct cell type or clusters.

2.6.2 Individual cell lineage tracing in COVID-19 single cell expression data

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has caused more than 98 million infections and more than 2.1 million deaths according to the statistics of World Health Organization (WHO) as of January 24, 2021. Although many COVID-19 patients are asymptomatic or experience only mild or moderate symptoms, some patients progress to severe disease or even death. It is thus important to understand the disease mechanisms to control the pandemic. The single cell RNA-seq sequencing technology (Grün & van Oudenaarden, 2015; Nawy, 2013; Shapiro et al., 2013) enable us to observe each cell separately and measure genes expressions simultaneously. The scRNA-seq is a powerful tool at dissecting the immune responses and has been applied to COVID-19 studies (Cao et al., 2020; Chua et al., 2020; Fan et al., 2020; Su et al., 2020; Wen et al., 2020; Xie et al., 2020; Zhang et al., 2020a, 2020b). However, existing technologies require destroying cell's transcription-continuity in the course of sequencing their gene expression profiles. Such static single cell RNA-seq (sc-RNA-Seq) experiments are insufficient to reveal the dynamics of cellular and gene dynamic changes (Hrvatin et al., 2018; Lawson et al., 2015; Spiller et al., 2010; Weinreb et al., 2018). There is a lack of approach on the changes between patients at different covid-19 levels.

This example aims to use the optimal transport method to coupling cells in different groups (healthy control, mild, severe, recovered). Then we propose to perform differential gene tests to help us understand

the genes or cell changes among these groups of patients. We are looking to find the abnormal caused by COVID-19. Through literature review, we obtain one single cell RNA-seq dataset for COVID-19 patients.

This dataset has single cell RNA sequencing (scRNA-seq) on Bronchoalveolar lavage fluid (BALF) cells from three patients with moderate COVID-19 (M1–M3), six patients with severe/critical infection (S1–S6), three healthy controls (HC1–HC3) and a publicly available BALF (HC4)4 sample. Some preliminary data processing results are shown in Figure 2.6. Comparing the severe patients with healthy controls, the cells in severe patients are not clearly clustered together as the cells in healthy controls. This indicates the cell types for severe patients have changed ambiguously, and genes in severe patients are mutated. We might be able to understand the changes by preform gene tests on the coupled cells.

This project framework first couples the cells respecting two different patients from different groups (healthy, mild and severe) using the optimal transport technique, which is a powerful tool to transform one probability measure to another (Peyré, Cuturi et al., 2019; Villani, 2008). The resulting couplings then are linearly interpolated to construct the coarse individual cell trajectories for each cell across the three patients groups.

Optimal transport for cell coupling

In order to link cells using optimal transport, we need to decide where our mass will lie and use the digital gene expression matrix as an experimental distribution as the input. A cell distribution can be described as the gene expression of all cells in a certain space. The input of optimal transport problem needs a source distribution and target distribution. We choose a space that we call ambient space X where we will have all our cells in different conditions we choose to put in this space. Next, we need to choose a cost function $c : \chi \times \chi \to \mathbb{R} \cup \{\infty\}$, which tell us what the effort is of moving one cell from one patient condition to another patient condition in this space. The cost function can be infinite if one cell has no possible ways to transport to another cell at two different conditions.

The single cell optimal transport happens between two cell distributions, so that we need to choose true probability measures μ and v. These cell distributions are called probability measures because these

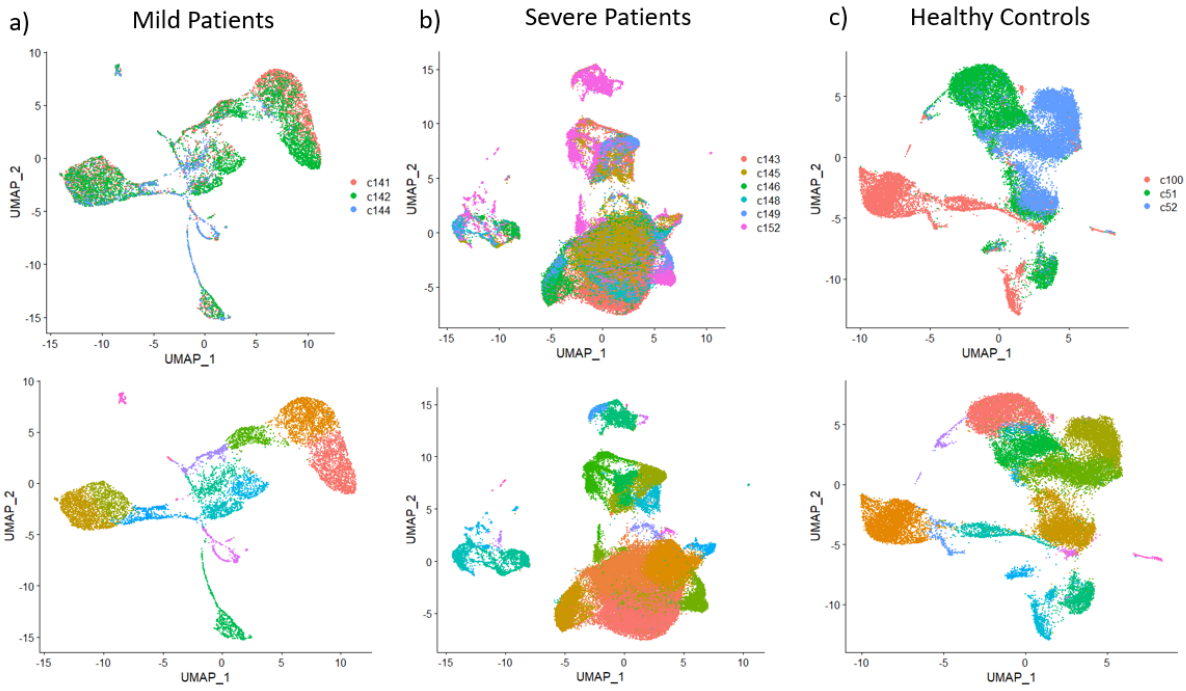

Figure 2.6: Single cell dataset for different types of COVID-19 patients: a) COVID-19 mild patients IDs (top) louvain cluster (bottom) b) COVID-19 severe patients IDs (top) louvain cluster (bottom). c) COVID-19 healthy controls IDs (top) louvain cluster (bottom)

are distributions of gene expressions in cells. Because if we want to transport cell from one to the another. Of course, the gene expressions need to be preserved during transportation. Now, we will have to decide to describe a transport plan between two measures using a symbol and call it γ. This γ tells us how much gene expression values in my transport are moved from one cell in distribution μ to one cell in distribution υ.

With this transportation plan γ, we can describe a way to move all the cells from the distribution μ to the cells in distribution υ. The cost associated with this whole transport displacement can be computed by the integral of all cell pairs on the space X. The formula contains the cost of cell pairs times the amount of gene expressions that we will transport between two different conditions. Therefore the optimal transport problem is just to minimize over all possible cell pairs transportation plans.

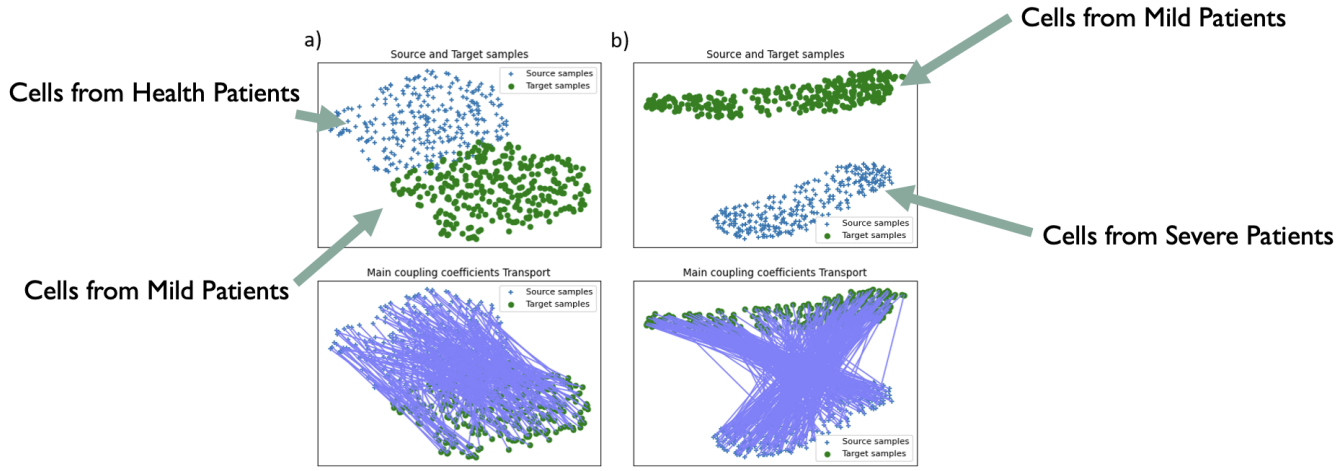

Figure 2.7: Individual cell connection between different types of patients. **a:** Individual cell Individual cell connections constructed between health patients and mild patients. **b:** Individual cell Individual cell connections constructed between mild patients and mild patients.

$$\min \int_{x \times x} c(x, y)\gamma(dx, dy) \text{ subject to } \gamma \in \text{coupling}(\mu, v)$$

There are constraints on this transportation plan γ, and we need the γ to be something that describes a good plan. We call this γ a cell coupling plan. The coupling is a distribution of gene expression on the pairs of cells. Therefore, this object γ actually describes a way to transport the mass from μ to the cell distribution v.

Figure 2.7 a)- b) shows an example of optimal transport at the different condition of COVID-19 patients. The cell distributions of health patients in figure 2.7 a) show in blue points, and the green points indicate the cell distribution of patients with mild symptoms. Similar to figure 2.7 b), the cell distributions of mild symptom patients in figure 2.7 a) are showing in blue points, and the green points indicate the cell distribution of patients with severe symptoms. The top figures show sample cell distributions of different patient types. The bottom three figures indicate the cell-to-cell coupling lines two different cell distributions. We first construct the cell-to-cell coupling lines for each consecutive condition. They are then connecting lines for each cell through all patient types to construct the individual cell trajectories.

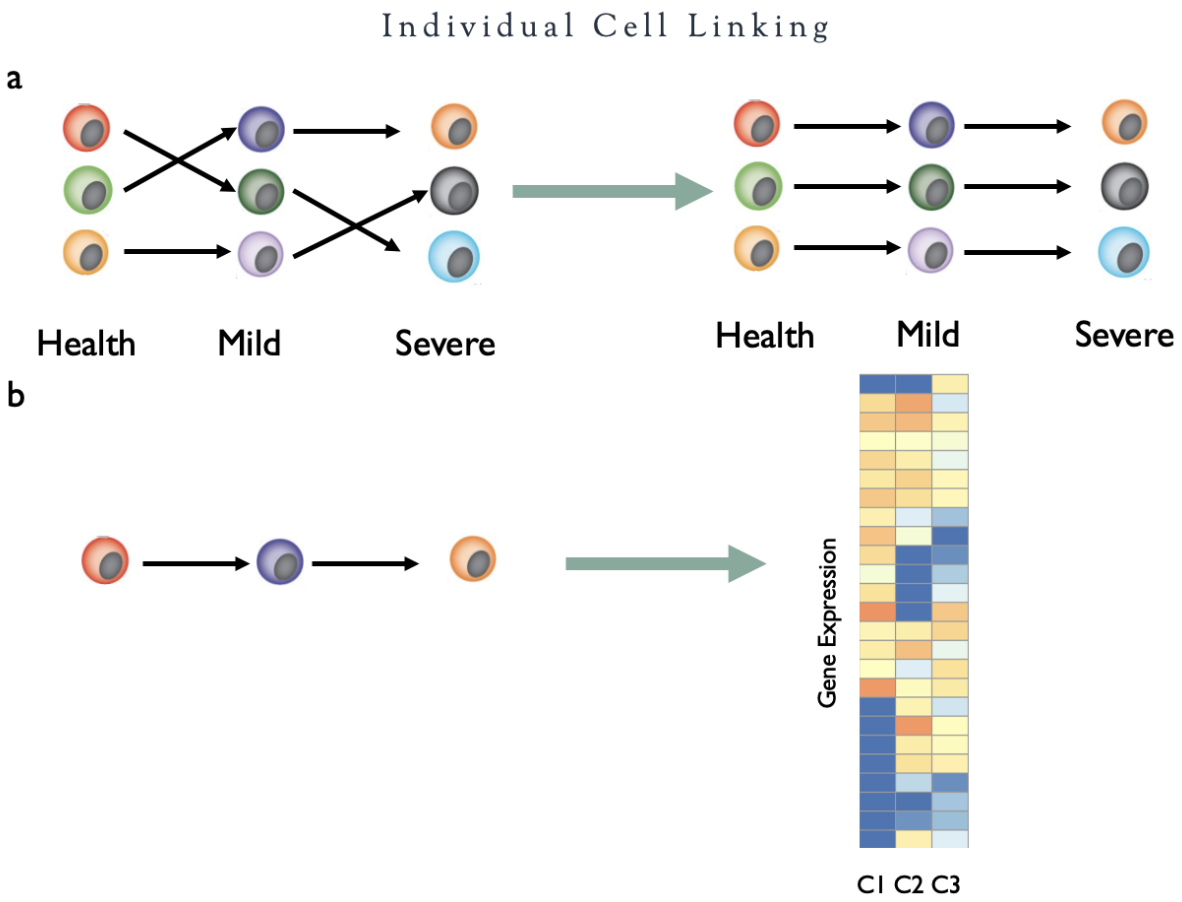

Figure 2.8: Illustration of individual cell linking between different types of patients. **a:** Individual cell can potentially represents the cell lineages/developments from health patient to severe patients through mild patients. **b:** We can observe the gene/pathway expression patterns with in one individual cell connection.

We know that optimal transport can be used to coupling points from one distribution to points in another distribution. In the analysis result section, I will implement this method to coupling cells in different conditions. This method can actually solve a big problem we are facing in the single cell sequencing data analysis. For example, if we want to analyze the dynamic changes for a certain number of cells. We have to sequence the cells at different time points separately. By doing so, we lose the transportation information for cells over time. This optimal transport can help us reconstruct the connection for cells at different time points or conditions.

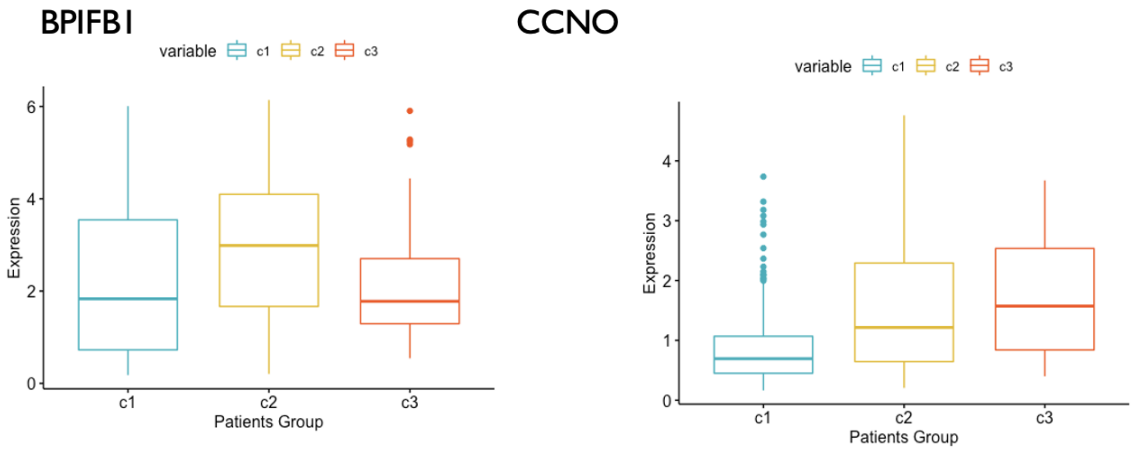

Figure 2.9: Differentially expressed genes estimated from the individual cell connections: We illustrate two example genes that are differently expressed in different types of patients.

Gene testing and gene ontology analysis We perform a statistical test for individual genes in all three patient types. The differentially expressed genes might be a regulator of COVID-19 virus development. We use the following one-way ANOVA test formula for individual genes:

$$H_0 : \mu_{i1} = \mu_{i2} = \mu_{i3} \quad \text{where } \mu \text{ is the expression for gene(i) in group(I).}$$

The alternative hypothesis (H_a): at least two groups' mean expressions are statistically significantly different from each other. The Gene Ontology analysis for all differentially expressed genes is performed using The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 with default settings for significant genes Bronchoalveolar lavage fluid(BALF) cells.

Analysis results

We first subset the data based on the different conditions to ensure the same number of cells in each patient type. Then we perform the optimal transport coupling method on health patients to mild to severe patients. We transform the cell dynamic development problem into the optimal transport problem. We

have two conditions with all cells and their gene expressions. As I described before, the optimal transport problem calculates the transport plan from the source cell distribution to the target cell distribution. Here we take the health patients cell distribution as the source distribution and the latter mild cell distribution as the target distribution. The optimal transport method is implemented here to find the cell-to-cell coupling from the two cell distributions. Then we are coupling the cell for all three different conditions. In other words, in all three different patients conditions, a cell in the health patient group will be coupling a cell in the mild patient group, and then the cell will connect a cell in the severe patient group. For each cell in the beginning health condition, we build a cell trajectory across all time points (Figure 2.8).

Table 2.3: The function cluster of differentially expressed genes

Term	Count	Enrichment	Pvalue
Antigen binding	28	43.2925444	3.11E-35
Immunoglobulin V region	18	64.8560924	6.45E-25
Extracellular space	48	5.74696967	1.30E-22
Immunoglobulin domain	31	9.81948164	1.19E-19
Disulfide bond	63	3.17292473	1.29E-17
Immunoglobulin-like domain	33	6.96801829	3.01E-16
Blood microparticle	19	20.159292	5.30E-17
Immunoglobulin receptor binding	12	73.5021771	9.68E-17
Regulation of immune response	20	16.9976718	3.32E-16
Immunoglobulin-like fold	34	5.82033686	7.92E-15
Adaptive immunity	18	17.8913358	5.02E-15

We apply the optimal transport method to coupling the cells at each time point (Figure 2.8 a)). Then we construct individual trajectories as many as possible in the dataset (Figure 2.8 a)). For one linked individual cell trajectory, there are approximately 2000 highly variable gene expression patterns. We cluster the genes using the K-mean clustering method (Figure 2.8 b)). Some clusters of genes have high expression values in healthy patients but low expression values in the patients with COVID-19 virus from the heat-map figure. Some clusters of genes are showing the opposite patterns. For the next step, we will perform the deferentially expressed gene test to analyze those genes further.

As we described in the method section, we perform the one-way ANOVA test on each gene. Figure 2.9 indicates two top differentially expressed genes by the ANOVA test. The box shows the actual expression

values in three types of patients. The expression value of the BPIFBI gene is high in mild patients and relatively low in healthy and severe patients. The CCNO gene shows an increased expression pattern as the patient type move from health to mild to severe. Those genes might be COVID-19 related genes because the gene expression values are significantly different in one of the patient types. We perform the ANOVA test on all genes, and 564 genes are significant in the test results.

For the last step, the gene ontology analysis was performed using DAVID. Gene Ontology functional clusters (table 2.3) indicate the major function of those genes is for the human body's immune system. Regulation genes can be found by testing individual cell trajectories in different patient types.

References

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D. et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, *182*(1), 145–161.

Alvarez-Melis, D., Jaakkola, T. & Jegelka, S. (2018). Structured optimal transport. *International Conference on Artificial Intelligence and Statistics*, 1771–1780.

An, S., Ma, L. & Wan, L. (2019). Tsee: An elastic embedding method to visualize the dynamic gene expression patterns of time series single-cell rna sequencing data. *BMC genomics*, *20*(2), 224.

Arjovsky, M., Chintala, S. & Bottou, L. (2017). Wasserstein generative adversarial networks. *International conference on machine learning*, 214–223.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F. & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, *37*(1), 38.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, *44*(4), 375–417.

Burrows, N., Bashford-Rogers, R. J., Bhute, V. J., Peñalver, A., Ferdinand, J. R., Stewart, B. J., Smith, J. E., Deobagkar-Lele, M., Giudice, G., Connor, T. M. et al. (2020). Dynamic regulation of hypoxia-

inducible factor-1α activity is essential for normal b cell development. *Nature Immunology*, *21*(11), 1408–1420.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, *36*(5), 411–420.

Canas, G. & Rosasco, L. (2012). Learning probability measures with respect to optimal transport metrics. *Advances in Neural Information Processing Systems*, 2492–2500.

Cannoodt, R., Saelens, W. & Saeys, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *European journal of immunology*, *46*(11), 2496–2506.

Cannoodt, R., Saelens, W., Sichien, D., Tavernier, S., Janssens, S., Guilliams, M., Lambrecht, B., De Preter, K. & Saeys, Y. (2016). Scorpius improves trajectory inference and identifies novel modules in dendritic cell development. *Biorxiv*, 079509.

Cazelles, E., Seguy, V., Bigot, J., Cuturi, M. & Papadakis, N. (2018). Geodesic PCA versus log-PCA of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing*, *40*(2), B429–B456.

Chen, H., Albergante, L., Hsu, J. Y., Lareau, C. A., Bosco, G. L., Guan, J., Zhou, S., Gorban, A. N., Bauer, D. E., Aryee, M. J. et al. (2019). Single-cell trajectories reconstruction, exploration and mapping of omics data with stream. *Nature communications*, *10*(1), 1–14.

Courty, N., Flamary, R., Habrard, A. & Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 3730–3739.

Courty, N., Flamary, R., Tuia, D. & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, *39*(9), 1853–1865.

Cuturi, M., Teboul, O. & Vert, J.-P. (2019). Differentiable ranking and sorting using optimal transport. *Advances in Neural Information Processing Systems*, 6861–6871.

Davis, F. M., Lloyd-Lewis, B., Harris, O. B., Kozar, S., Winton, D. J., Muresan, L. & Watson, C. J. (2016). Single-cell lineage tracing in the mammary gland reveals stochastic clonal dispersion of stem/progenitor cell progeny. *Nature communications*, *7*(1), 1–13.

Del Barrio, E., Gordaliza, P., Lescornel, H. & Loubes, J.-M. (2019). Central limit theorem and bootstrap procedure for Wasserstein's variations with an application to structural relationships between distributions. *Journal of Multivariate Analysis*, *169*, 341–362.

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D. & Schwing, A. G. (2019). Max-sliced wasserstein distance and its use for GANs. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 10648–10656.

Ferradans, S., Papadakis, N., Peyré, G. & Aujol, J.-F. (2014). Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, *7*(3), 1853–1882.

Flamary, R., Courty, N., Tuia, D. & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*

Flamary, R., Cuturi, M., Courty, N. & Rakotomamonjy, A. (2018). Wasserstein discriminant analysis. *Machine Learning*, *107*(12), 1923–1945.

Flamary, R., Lounici, K. & Ferrari, A. (2019). Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation. *arXiv preprint arXiv:1905.10155*.

Fletcher, R. B., Das, D., Gadye, L., Street, K. N., Baudhuin, A., Wagner, A., Cole, M. B., Flores, Q., Choi, Y. G., Yosef, N. et al. (2017). Deconstructing olfactory stem cell trajectories at single-cell resolution. *Cell stem cell*, *20*(6), 817–830.

Gadye, L., Das, D., Sanchez, M. A., Street, K., Baudhuin, A., Wagner, A., Cole, M. B., Choi, Y. G., Yosef, N., Purdom, E. et al. (2017). Injury activates transient olfactory stem cell states with diverse lineage capacities. *Cell stem cell*, *21*(6), 775–790.

Grün, D. & van Oudenaarden, A. (2015). Design and analysis of single-cell sequencing experiments. *Cell*, *163*(4), 799–810.

Guo, J., Grow, E. J., Yi, C., Mlcochova, H., Maher, G. J., Lindskog, C., Murphy, P. J., Wike, C. L., Carrell, D. T., Goriely, A. et al. (2017). Chromatin and single-cell rna-seq profiling reveal dynamic signaling and metabolic transitions during human spermatogonial stem cell development. *Cell Stem Cell*, *21*(4), 533–546.

Haghverdi, L., Buettner, F. & Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, *31*(18), 2989–2998.

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, *13*(10), 845–848.

Haghverdi, L., Lun, A. T., Morgan, M. D. & Marioni, J. C. (2018). Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, *36*(5), 421–427.

Hardoon, D. R., Szedmak, S. & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, *16*(12), 2639–2664.

Hie, B., Bryson, B. & Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature biotechnology*, *37*(6), 685–691.

Hrvatin, S., Hochbaum, D. R., Nagy, M. A., Cicconet, M., Robertson, K., Cheadle, L., Zilionis, R., Ratner, A., Borges-Monroy, R., Klein, A. M. et al. (2018). Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nature neuroscience*, *21*(1), 120–129.

Ji, Z. & Ji, H. (2016). Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, *44*(13), e117–e117.

Jolliffe, I. T. & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

Kantorovich, L. V. (2006). On a problem of Monge. *Journal of Mathematical Sciences*, *133*(4), 1383–1383.

Kantorovitch, L. (1958). On the translocation of masses. *Management Science*, *5*(1), 1–4.

Kester, L. & van Oudenaarden, A. (2018). Single-cell transcriptomics meets lineage tracing. *Cell stem cell*, *23*(2), 166–179.

Ko, M. E., Williams, C. M., Fread, K. I., Goggin, S. M., Rustagi, R. S., Fragiadakis, G. K., Nolan, G. P. & Zunder, E. R. (2020). Flow-map: A graph-based, force-directed layout algorithm for trajectory mapping in single-cell time course datasets. *Nature protocols*, *15*(2), 398–420.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r. & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 1–8.

Lawson, D. A., Bhakta, N. R., Kessenbrock, K., Prummel, K. D., Yu, Y., Takai, K., Zhou, A., Eyob, H., Balakrishnan, S., Wang, C.-Y. et al. (2015). Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature*, *526*(7571), 131–135.

Lin, C. & Bar-Joseph, Z. (2019). Continuous-state hmms for modeling time-series single-cell rna-seq data. *Bioinformatics*, *35*(22), 4707–4715.

Liu, Z., Lou, H., Xie, K., Wang, H., Chen, N., Aparicio, O. M., Zhang, M. Q., Jiang, R. & Chen, T. (2017). Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nature communications*, *8*(1), 1–9.

Lotfollahi, M., Wolf, F. A. & Theis, F. J. (2019). Scgen predicts single-cell perturbation responses. *Nature methods*, *16*(8), 715–721.

Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., Pelka, K., Ge, W., Oren, Y., Brack, A. et al. (2019). Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell*, *176*(6), 1325–1339.

McInnes, L., Healy, J. & Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction.

Meng, C., Ke, Y., Zhang, J., Zhang, M., Zhong, W. & Ma, P. (2019). Large-scale optimal transport map estimation using projection pursuit. *Advances in Neural Information Processing Systems*, 8116–8127.

Meng, C., Zhang, X., Zhang, J., Zhong, W. & Ma, P. (2020). More efficient approximation of smoothing splines via space-filling basis selection. *Biometrika*, *107*, 723–735.

Moris, N., Pina, C. & Arias, A. M. (2016). Transition states and cell fate decisions in epigenetic landscapes. *Nature Reviews Genetics*, *17*(11), 693–703.

Nawy, T. (2013). Single-cell sequencing. *Nature methods*, *11*(1), 18.

Nowakowski, T. J., Bhaduri, A., Pollen, A. A., Alvarado, B., Mostajo-Radji, M. A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S. J., Velmeshev, D. et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science*, *358*(6368), 1318–1323.

Panaretos, V. M. & Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual review of statistics and its application*, *6*, 405–431.

Paty, F.-P. & Cuturi, M. (2019). Subspace robust wasserstein distances. *arXiv preprint arXiv:1901.08949*.

Paty, F.-P., d'Aspremont, A. & Cuturi, M. (2020). Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport. *International Conference on Artificial Intelligence and Statistics*, 1222–1232.

Peyré, G., Cuturi, M. et al. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, *11*(5-6), 355–607.

Plass, M., Solana, J., Wolf, F. A., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F. J., Kocks, C. & Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, *360*(6391), eaaq1723.

Polański, K., Young, M. D., Miao, Z., Meyer, K. B., Teichmann, S. A. & Park, J.-E. (2020). Bbknn: Fast batch alignment of single cell transcriptomes. *Bioinformatics*, *36*(3), 964–965.

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A. & Trapnell, C. (2017). Single-cell mrna quantification and differential analysis with census. *Nature methods*, *14*(3), 309.

Rabin, J., Ferradans, S. & Papadakis, N. (2014). Adaptive color transfer with relaxed optimal transport. *2014 IEEE International Conference on Image Processing (ICIP)*, 4852–4856.

Redko, I., Courty, N., Flamary, R. & Tuia, D. (2019). Optimal transport for multi-source domain adaptation under target shift. *The 22nd International Conference on Artificial Intelligence and Statistics*, 849–858.

Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., Larson, D. R. & Zhao, K. (2017). Ctcf-mediated enhancer-promoter interaction is a critical regulator of cell-to-cell variation of gene expression. *Molecular Cell*, *67*(6), 1049–1058.

Rodriguez-Fraticelli, A. E., Weinreb, C., Wang, S.-W., Migueles, R. P., Jankovic, M., Usart, M., Klein, A. M., Lowell, S. & Camargo, F. D. (2020). Single-cell lineage tracing unveils a role for tcf15 in haematopoiesis. *Nature*, *583*(7817), 585–589.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P. et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, *176*(4), 928–943.

Shapiro, E., Biezuner, T. & Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, *14*(9), 618–630.

Simeonov, K. P., Byrns, C. N., Clark, M. L., Norgard, R. J., Martin, B., Stanger, B. Z., Shendure, J., McKenna, A. & Lengner, C. J. (2021). Single-cell lineage tracing of metastatic cancer reveals selection of hybrid emt states. *Cancer Cell*, *39*(8), 1150–1162.

Sipp, D., Robey, P. G. & Turner, L. (2018). Clear up this stem-cell mess.

Spiller, D. G., Wood, C. D., Rand, D. A. & White, M. R. (2010). Measurement of single-cell dynamics. *Nature*, *465*(7299), 736–745.

Stévant, I., Neirijnck, Y., Borel, C., Escoffier, J., Smith, L. B., Antonarakis, S. E., Dermitzakis, E. T. & Nef, S. (2018). Deciphering cell lineage specification during male sex determination with single-cell rna sequencing. *Cell reports*, *22*(6), 1589–1599.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, *177*(7), 1888–1902.

Su, Z., Wang, Y., Shi, R., Zeng, W., Sun, J., Luo, F. & Gu, X. (2015). Optimal mass transport for shape matching and comparison. *IEEE transactions on pattern analysis and machine intelligence*, *37*(11), 2246–2259.

Sun, C., Wang, H., Ma, Q., Chen, C., Yue, J., Li, B. & Zhang, X. (2021). Time-course single-cell rna sequencing reveals transcriptional dynamics and heterogeneity of limbal stem cells derived from human pluripotent stem cells. *Cell & Bioscience*, *11*(1), 1–12.

Torii, K., Kubota, A., Araki, T. & Endo, M. (2020). Time-series single-cell rna-seq data reveal auxin fluctuation during endocycle. *Plant and Cell Physiology*, *61*(2), 243–254.

Tran, T. N. & Bader, G. D. (2020). Tempora: Cell trajectory inference using time-series single-cell rna sequencing data. *PLoS computational biology*, *16*(9), e1008205.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome research*, *25*(10), 1491–1498.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, *32*(4), 381.

Treutlein, B., Lee, Q. Y., Camp, J. G., Mall, M., Koh, W., Shariati, S. A. M., Sim, S., Neff, N. F., Skotheim, J. M., Wernig, M. et al. (2016). Dissecting direct reprogramming from fibroblast to neuron using single-cell rna-seq. *Nature*, *534*(7607), 391–395.

Tritschler, S., Büttner, M., Fischer, D. S., Lange, M., Bergen, V., Lickert, H. & Theis, F. J. (2019). Concepts and limitations for learning developmental trajectories from single cell genomics. *Development*, *146*(12).

Villani, C. (2008). *Optimal transport: Old and new*. Springer Science & Business Media.

Wagner, D. E. & Klein, A. M. (2020). Lineage tracing meets single-cell omics: Opportunities and challenges. *Nature Reviews Genetics*, *21*(7), 410–427.

Wang, Z., Zhang, Y. & Wu, H. (2020). Structural-aware sentence similarity with recursive optimal transport. *arXiv preprint arXiv:2002.00745*.

Wang, Z., Zhou, D., Yang, M., Zhang, Y., Rao, C. & Wu, H. (2020). Robust document distance with wasserstein-fisher-rao metric. *Asian Conference on Machine Learning*.

Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, *115*(10), E2467–E2476.

Wu, J., Huang, Z., Acharya, D., Li, W., Thoma, J., Paudel, D. P. & Gool, L. V. (2019). Sliced wasserstein generative models. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3713–3722.

Yuan, Y. & Bar-Joseph, Z. (2021). Deep learning of gene relationships from single cell time-course expression data. *Briefings in Bioinformatics*, *22*(5), bbab142.

Zhang, J., Zhong, W. & Ma, P. (2020). A review on modern computational optimal transport methods with applications in biomedical research. *arXiv preprint arXiv:2008.02995*.

Zhao, X., Wang, Z., Wu, H. & Zhang, Y. (2020). Semi-supervised bilingual lexicon induction by two-way interaction. *Empirical Method in Natural Language Processing*.

Zhao, X., Wang, Z., Zhang, Y. & Wu, H. (2020). A relaxed matching procedure for unsupervised BLI. In D. Jurafsky, J. Chai, N. Schluter & J. R. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, online, july 5-10, 2020* (pp. 3036–3041). Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.274/

Zhou, Y., Yang, D., Yang, Q., Lv, X., Huang, W., Zhou, Z., Wang, Y., Zhang, Z., Yuan, T., Ding, X. et al. (2020). Single-cell rna landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. *Nature communications*, *11*(1), 1–17.

Chapter 3

Elucidation of Cell Lineages and Dynamic Gene Networks in Time-course Single Cell Expression Data

3.1 Background in cell trajectories and lineage tracing

A comprehensive understanding of any complex biological process such as tissue development and regeneration requires the investigation of cell progression behaviors across a wide range of samples and experimental time points (Spiller et al., 2010). Cell progression, including cell reprogramming, differentiation, and morphological development, is a dynamic and continuous process (Burrows et al., 2020; Guo et al., 2017). Cell progression processes includes rapid changes in metabolism, gene expressions and cell types over time. To profile such cell progression behaviors, single cell RNA-seq sequencing (scRNA-seq) technology has been developed rapidly (Grün & van Oudenaarden, 2015; Nawy, 2013; Shapiro et al., 2013; Tanay & Regev, 2017). In particular, scRNA-seq enables researchers to observe the gene expressions of all

cells simultaneously. Such single cell sequencing techniques are usually conducted in static or time-course experiments. The static scRNA-seq experiment takes a snapshot of all cells and their gene expressions at one time point (Hrvatin et al., 2018; Lawson et al., 2015), whereas the time course scRNA-seq experiments takes snapshots at multiple time points. Despite its importance, quantifying the dynamic cellular changes of cell development is still challenging due to some limitations (Stegle et al., 2015). In time course scRNA-seq experiments, cells are sacrificed and sequenced at each time point. Thus there is no cell correspondence information for cells between two time points, which creates a significant challenge to elucidate the dynamic behaviors of cell progression. Moreover, the cell-cell variation is too large to be ignored. It is also challenging to align and register different cells sequenced in two adjacent time points since the gene expression is drastically affected by cell variation (Alonge et al., 2020; Ren et al., 2017). Without controlling the cell variation, gene expression analysis can be significantly biased.

For static scRNA-seq experiments, one natural approach to surmounting the challenges is to order cells into a continuous cell trajectory. Many methods have been proposed to achieve this goal. In these methods, researchers construct a pseudo time to order cells chronologically (Cannoodt et al., 2016; Chen et al., 2019; Ji & Ji, 2016; Liu et al., 2017; Qiu et al., 2017; Trapnell, 2015; Trapnell et al., 2014). Despite their effectiveness, such methods may fail in the following circumstances (Tritschler et al., 2019). First of all, most existing trajectory inference methods construct a bulk cell trajectory, i.e., the mean trajectory of the population cells across time rather than that of individual cells. However, some individual cells' behaviors may oscillate up and down around the cell mean expressions or severely deviate from it. Cell progression behaviors are dominated by cells with major cell types, and patterns with less frequent might be hidden in the dataset. Second, individual cell developing trajectories may follow different complex topologies, including loops or alternative paths during the development. For example, analysis approaches in (Moon et al., 2018) and (Dai et al., 2020) used dimension reduction methods to identify a low-dimensional space of the gene expression space before construct cell trajectory (Saelens et al., 2019; Wagner et al., 2018). Those methods may introduce a significant bias and are hard to validate, as cells are ordered based only on the selected reduced dimensions. Finally, the cells may not be synchronized at the same developing time points.

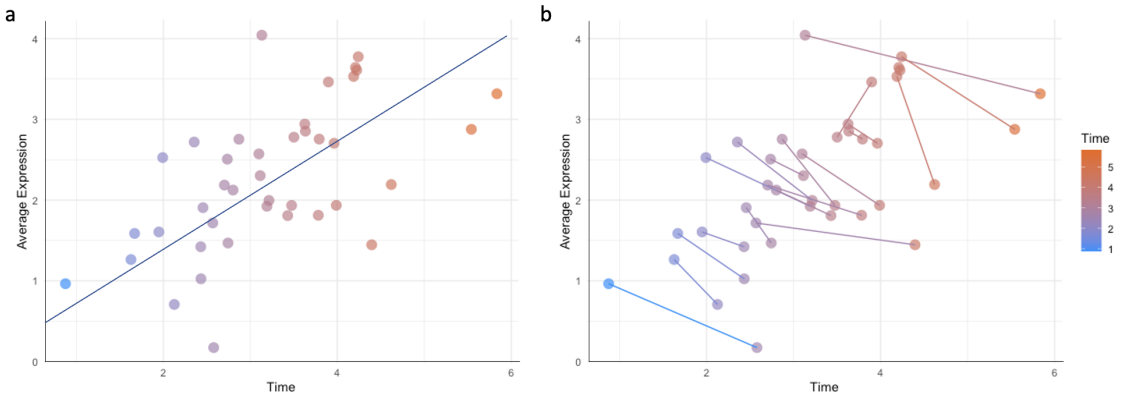

Figure 3.1: Example advantage of individual cell linking: **a**: Cell progression over time (x-axis) reflects a increasing trend in average cell expressions (y-axis). **b**: The individual cell correspondences at different time points reflect a decreasing trend in average cell expression (y-axis).

Cells within the same time point can be expressed at different developing stages. Under this situation, the bulk cell trajectory that takes the average pattern of cells at different stages might result in unreliable scientific discovery.

Time-course scRNA-seq experiments contain intrinsically much more informative than the static scRNA-seq data, particularly for the inference of cellular dynamic development patterns (An et al., 2019; Ko et al., 2020; Sun et al., 2021; Torii et al., 2020; Yuan & Bar-Joseph, 2021). However, the challenges of cell correspondence and cell variation remain unsolved.

The scRNA-seq data are still static at a time point, and the cell correspondence information through multiple time points is still missing. Moreover, the existing cell trajectory inference methods may neglect some hidden expression patterns in the cell development process. We illustrate this problem by using a simulated time-course cell dataset as a toy example (Figure 3.1). Under some cell development and differentiation circumstances, cells show an increasing pattern if we only construct one average cell trajectory to order cells (Figure 3.1a). However, when individual cells are linked at different time points, the individual cell trajectories reflect unique decreasing patterns, which are in contrast to the average expression (Figure 3.1b). Those cell development patterns can be easily misled by the average cell trajectory and thus reflecting spurious cell progression behaviors.

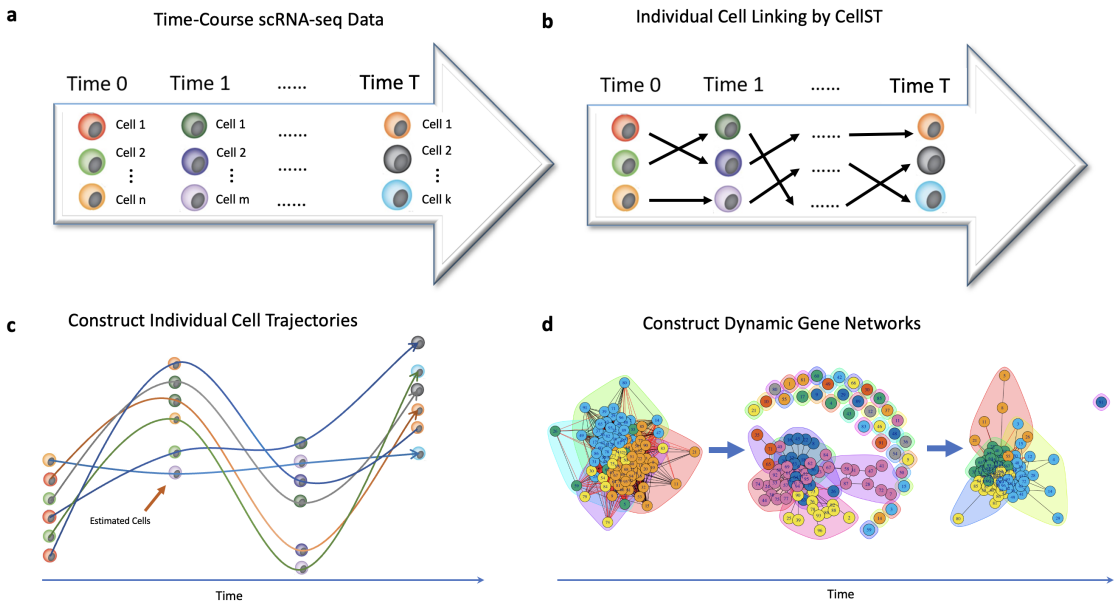

Figure 3.2: Cell smoothing transformation (CellST) overview: **a**: An illustration of time-course scRNA-seq data. Cells contain no information between time points. **b**: CellST construct cell correspondence information between time points using optimal transport technique. **c**: CellST utilizing smoothing spline technique on gene expression in each cell to finish construct the estimated cell-to-cell trajectories. **d**: CellST construct dynamic gene networks based on the calculated dynamic relationship between genes.

In this paper, we propose a novel analysis framework named Cell Smoothing Transformation (CellST) to overcome the aforementioned limitations. The CellST framework constructs individual cell trajectories and dynamic gene co-expression networks for time course scRNA-seq data (Figure 3.2a). In the CellST framework, we propose a cell linking method, which pairwise couples individual cells and construct cell correspondences between adjacent two time points. Those cell couplings can potentially represent individual cell lineages, tracing cell progression behaviors by constructing an individual cell trajectory. The cell linking method couples individual cells using the optimal transport technique (Meng et al., 2019; J. Zhang et al., 2020), which is a powerful tools that can be used to model cell dynamics (Schiebinger et al., 2019; Tong et al., 2020). The resulting cell couplings are then linked by a straight line to construct cell correspondences and cell-cell alignment across the time (Figure 3.2b). Next, we utilize the smoothing spline models to construct the smoothing trajectories to reduce both gene-gene and cell-cell variance. The smoothing spline method models the gene expression patterns in the cells correspondence constructed

from the previous step and builds the estimated smoothing individual cell trajectories. Lastly, we narrow down our focus to utilize the gene expression patterns from those individual cell trajectories to construct dynamic gene networks (Figure 3.2d). The dynamic gene networks is constructed by estimating the dynamic relationship of pairwise gene expression patterns using functional concurrent models (Wang et al., 2016) and smoothing spline models (Gu & Ma, 2005). Furthermore, the CellST dynamic network can be used to find critical genes by profiling genes that have a significantly different patterns with other genes.

Our major contribution is to develop the first analysis framework (CellST) to construct individual cell-level trajectories, which can help researchers trace individual cell progression behaviors over time. The promise of the single cell sequencing technology enables researchers to observe individual cells' behaviors instead of observing the bulk behaviors of cells. However, existing methods are still estimating the bulk trajectory in scRNA-seq datasets. Those analysis methods may overlook the hidden patterns in the cell progression process and thus create a spurious cell progression trajectory. Furthermore, we propose the dynamic gene (co-expression) network based on the individual cell trajectories to estimate the dynamic gene-gene relationship over time. The empirical performance of the proposed framework is evaluated by several simulated and real data studies.

3.2 Observing dynamic cell lineages in simulation and real data

3.2.1 Simulation results

We evaluated the performance of the CellST framework on constructing individual cell trajectories by analyzing simulated scRNA-seq datasets. The details of generating simulation data can be found in the method section. The simulation analysis was conducted in two scenarios: In the first scenario, we simulated scRNA-seq datasets with cells at only two time points to investigate the accuracy of constructing cell correspondence between time points. In the second scenario, we simulated a time course scRNA-seq

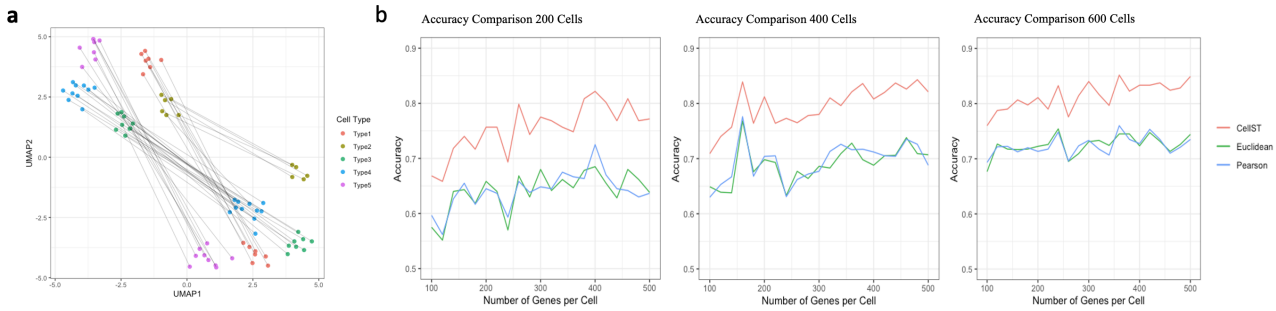

Figure 3.3: Simulation example of cell linking process at two time points. **a**: Cell linking process with five (right) cell types in both time points. **b**: Accuracy comparison of the cell linking process (red) with other gene similarity measurements (Pearson correlation (blue) and Euclidean distance (green)).

dataset with multiple time points to examine individual cell progression patterns in the individual cell trajectories.

Scenario one: CellST construct cell correspondences in high accuracy

To investigate the accuracy of the CellST cell linking method, we simulated scRNA-seq experiments with only two different time points. The simulated datasets, which contain the same number of cells and cell types, were generated independently for each time point.

These simulation datasets contain five same cell types in both time points. In the simulation setting, the number of cells ranges from 200 to 600, and the number of genes in one cell ranges from 100 to 500. The cell alignment and one to one correspondences were constructed using the CellST cell linking method based only on the gene expression information of cells at each time points and no information on the benchmark labels of cell types. Specifically, we estimated an empirical transportation cost for the individual cell correspondence between two time points using the gene expressions in cells. We linked cells by selecting pairs with the smallest transportation cost. Since cell dynamics is a gradually development process and cells within the same cell type tend to have similar gene expression profiles, the cell linking accuracy can be validated by counting the number of linked cell pairs with the same cell type (Figure 3.3b).

We noticed that the accuracy of the cell linking method has an increasing trend as we add more genes in cells for the simulated data. This observation is due to the fact that the CellST gets more information to learn the patterns of genes when more genes are simulated in each cell. Similarly, increasing cell numbers will also increase the linking accuracy since cells can be treated as information replicates to enhance the accuracy. We also compared the accuracy of coupled cells with the Euclidean distance and Pearson's correlation methods. Those two methods are the most commonly used distances and similarity measures for gene expression analysis. (Angermueller et al., 2016; Klimovskaia et al., 2020; Skinnider et al., 2019). The accuracy comparison results (Figure 3.3b) shows the CellST method achieves the best cell linking accuracy in the simulation settings. In summary, the cell linking method achieves high accuracy and captures the significant gene expressions when linking cells and construct individual cell correspondences at two different time points. This is crucial for the down-streaming individual cell trajectories construction.

Scenario two: CellST provides unique individual cell progression behaviors

To investigate the effectiveness in constructing the individual cell trajectories, we simulated a time course scRNA-seq data contains 160 cells at each time point and 13 experimental time points. This simulation dataset has two pathways with different development expression patterns, and each pathway contains 100 genes. The first pathway is created using the contact inhibition genes that keep cells growing into only a layer one cell thick (monolayer) (Mendonsa et al., 2018; Pavel et al., 2018). The growth of cells' average expression in this simulated pathway is diminishing and approaching an equilibrium expression over time. We simulated the second pathway according to the cellular division process, which is more active in cells under mitosis and less active in cells in interphase (Tomasetti et al., 2017). At each time point, eighty cells contain only the contact inhibition pathway, and eighty cells contain only the cellular division pathway.

To observe the dynamic of cell progression in this data, we utilized cells' experimental time information and built individual cell trajectories using CellST to reconstruct the individual cell development patterns. Cells' correspondence information between adjacent time points were constructed using cell linking method in CellST (Figure 3.4a). Those linked curves are smoothed using smoothing spline tech-

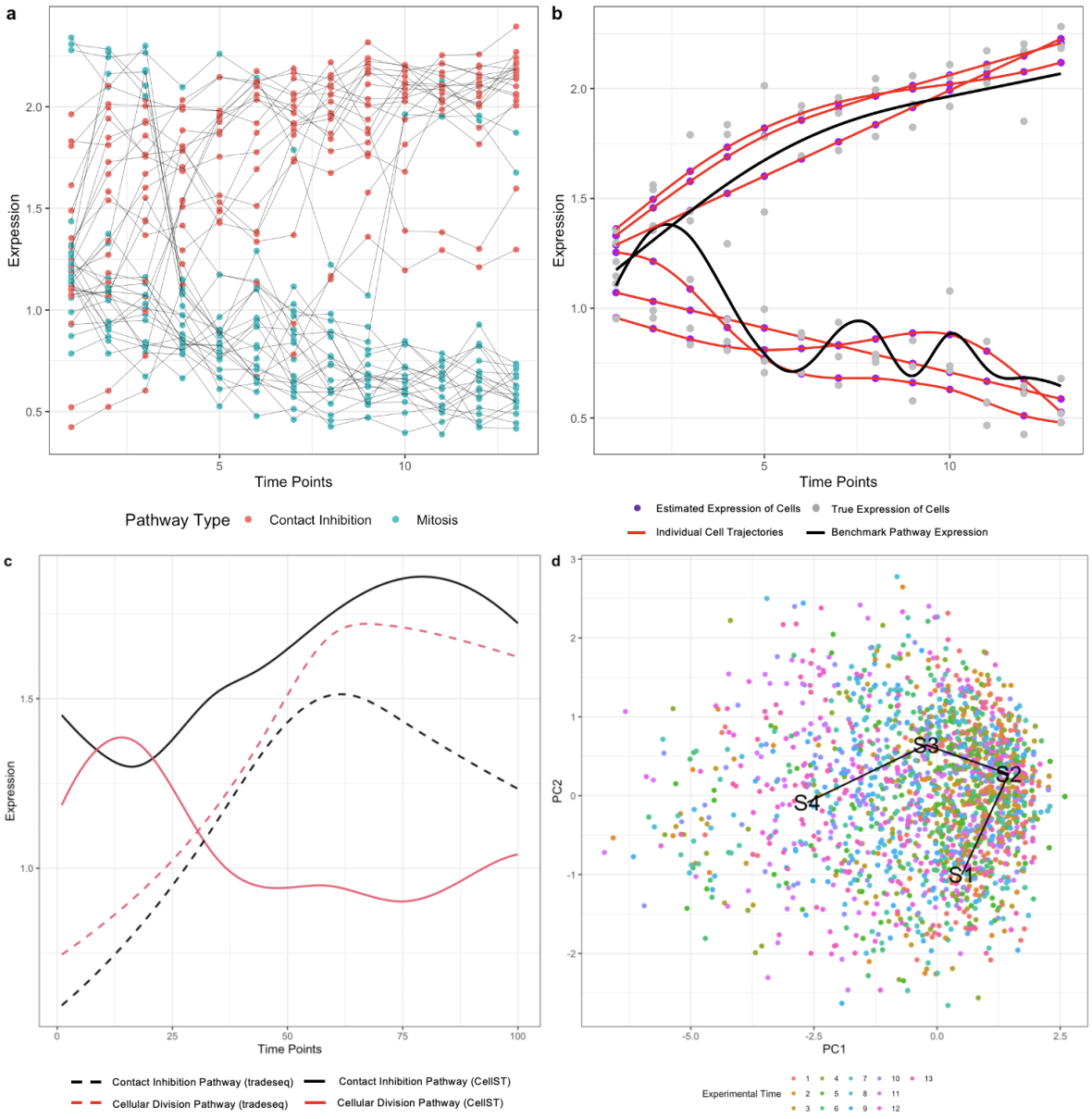

Figure 3.4: **a**: Cell couplings through all time points constructed by CellST method. The cells are classified by the pathway they contained. **b**: The individual cell trajectories (red curves) built by CellST and Benchmark average expression cell trajectory(black curve). **c**: Development expression patterns for a simulated gene (m_55). The red and black curves estimated by CellST method indicate the gene expression in two different pathways. The dotted two curves is constructed by tradeseq method. **d**: The pseudo time bulk trajectory constructed by TSCAN method. Cells are colored with experimental time and S1 to S4 indicates the four developing stages predicted by the TSCAN method.

nique in CellST to estimate the individual cell trajectories. Figure 3.4b illustrates the estimated individual cell trajectories (red curves). The two types of cells, simulated by the two distinct pathways, are well separated by CellST method. The expressions of individual cell trajectories were compared with the benchmark pathway expression patterns (black curves). The expression of individual cell trajectories illustrates a consistent patterns with the benchmark expression of the two simulated pathways over time. In addition to the consistency, we observed from the individual cell trajectories that cells have unique behaviors over time. Some cells grow slower and have lower expression values, while others grow faster and have higher expression values than the simulated average development patterns. In summary, the individual cell trajectories estimates the unique cell development patterns by constructing individual cell trajectories.

Next, we performed a comparative analysis of CellST with the existing trajectories analysis method "tradeseq" (Van den Berge et al., 2020). The tradeseq is a trajectory based method to estimate the dynamic expressions of differentially expressed genes. Comparing between the dynamic gene expression patterns constructed by CellST and tradeseq (Figure 3.4c), the tradeseq method constructed two similar expression patterns for a simulated gene expression (m_55, dotted curves), which the CellST method built two distinct expression patterns (Figure 3.4c black and red curves). Those constructed dynamic gene expression curves by CellST are also consistent with the simulated benchmark expression by showing the distinct expression patterns. When constructing the individual cell trajectories, the CellST method can automatically classify cells that contain different pathway expressions and construct cell correspondences within the same pathway.

Additionally, to compare the performance of the existing trajectory inference method based on the pseudo-time construction, we built a pseudo-time bulk trajectory by using the TSCAN method (Ji & Ji, 2016) (Figure 3.4d). In figure 3.4d, the four pseudo-time cell development stages from the TSCAN method were marked as S1 to S4 and the bulk cell trajectory was constructed by connecting the pseudo-time cell stages. Each point is a cell in the figure and the cells' color representing the cells' experimental time. The pseudo-time cell bulk trajectory constructed by the TSCAN does not match with cells' actually experimental times to reflect the correct cell development patterns, which are regulated by the two distinct

pathways. Therefore, the constructed pseudo-time bulk trajectory might be spurious in observing the cell progression.

3.2.2 Case study one: individual cell trajectories on cell reprogramming process

To evaluate the performance of CellST on real data, we use a mouse cell reprogramming scRNA-seq dataset (Schiebinger et al., 2019). The experiment in this dataset uses single cell RNA sequencing to reconstruct the landscape of reprogramming from induced pluripotent stem cells (iPSCs). The dataset contains 149,155 cells collected at 17 time points across 8 days, with samples taken every 12 hours. Cells were collected from established iPSCs cell lines reprogrammed from the same MEFs and maintained in serum medium. This scRNA-seq data were generated at each time point using the 10X Genomics Chromium Controller Instrument and ChromiumTM single cell 30 Reagent Kits. The details of data preprocessing procedures are included in the method section. All genes in this dataset are classified into different pathways, including cell cycle pathway, pluripotency pathway, epithelial pathway, etc. The pathways were determined based on co-expression signatures correlated with the gene of interest. For instance, the cell cycle pathway includes 97 co-expression genes regulating the cell cycles process, and the Astrocytes pathway has 224 genes involved in regulating the center nervous system in mouses. The individual cell trajectories were built based on the gene expressions in each pathway using CellST (Figure 3.5a and 3.5b).

Since the biological cell cycle concept is already well established for cell cycle pathway genes, the expression of the cell cycle pathway will periodically reach a peak in the mitosis phase and stay quiet in other phases. The individual cell trajectories for cell cycle genes (Figure 3.5b red curves) reflect similar patterns consistent with the average developing patterns of the cell cycle pathway. Moreover, the individual cell trajectories reveal the unique individual cell progression behaviors. Those individual cell behaviors illustrate different cell growth rates and times when cells reach expression peaks. In addition to the cell cycle pathway analysis, we also constructed individual cell trajectories for Astrocytes, Placental, Pluripotency, Trophoblast, and Epithelial pathways (Figures in Supplementary file). The individual cell trajectories return consistent expression patterns compared with the Waddington-OT method (Fig. 3.5 blue line).

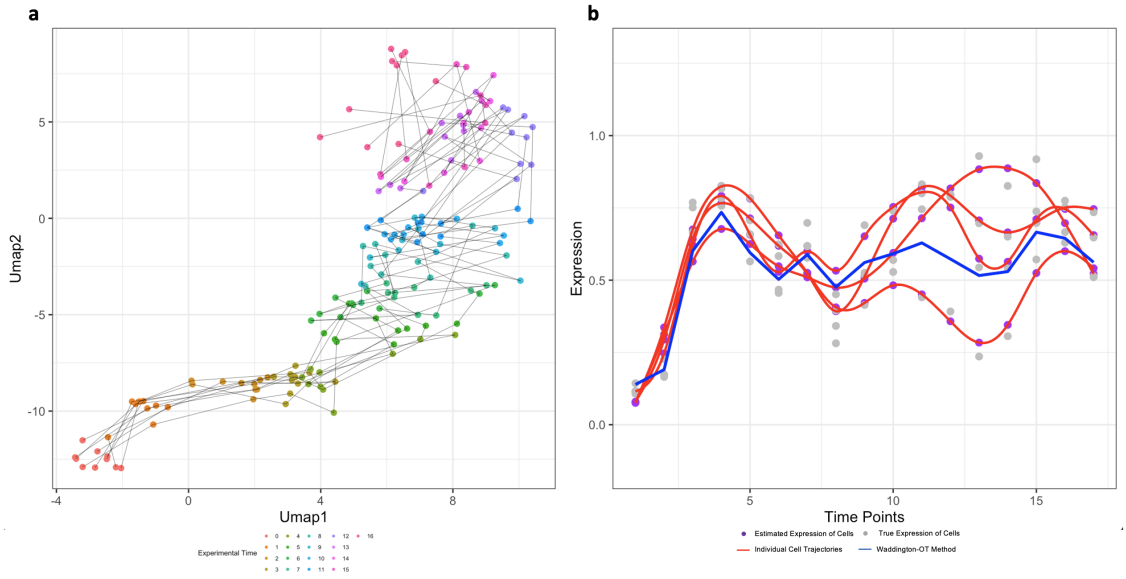

Figure 3.5: Individual cell trajectories constructed by CellST. **a**: The Umap visualization of the individual cell trajectories constructed using genes in the Cell Cycle pathway by the CellST cell linking process. **b**: Individual cell trajectories estimated by CellST for cell cycle genes.

After constructing the cell cycle individual cell trajectories and their gene expression patterns, we further built the dynamic gene network by estimating the dynamic relationship of pairwise genes (Figure 3.6). The gene-gene dynamic relationships were estimated using a functional concurrent model, and the dynamic gene network was constructed based on the measurement of the gene-gene relationships. Figure 3.6c illustrates the dynamic network constructed based on the cell cycle pathway genes. The nodes are genes from the cell cycle pathway, and the edges indicate the relationship from one gene to another gene. We further calculated co-expression gene communities using a network community detection algorithm (Louvain) (De Meo et al., 2011; Traag et al., 2019) for those cell cycle genes. The dynamic co-expression gene community was constructed by connecting the gene communities at each time point based on the number of common genes in the community. The expression patterns of the dynamic gene community indicate consistent periodical expression patterns compared with cell cycle pathway patterns (Figure 3.6b). We observed that the gene-gene relationship in the dynamic network also shows a periodical pattern. The dynamic gene network illustrates a strong relationship between genes in the mitosis phase and a weak

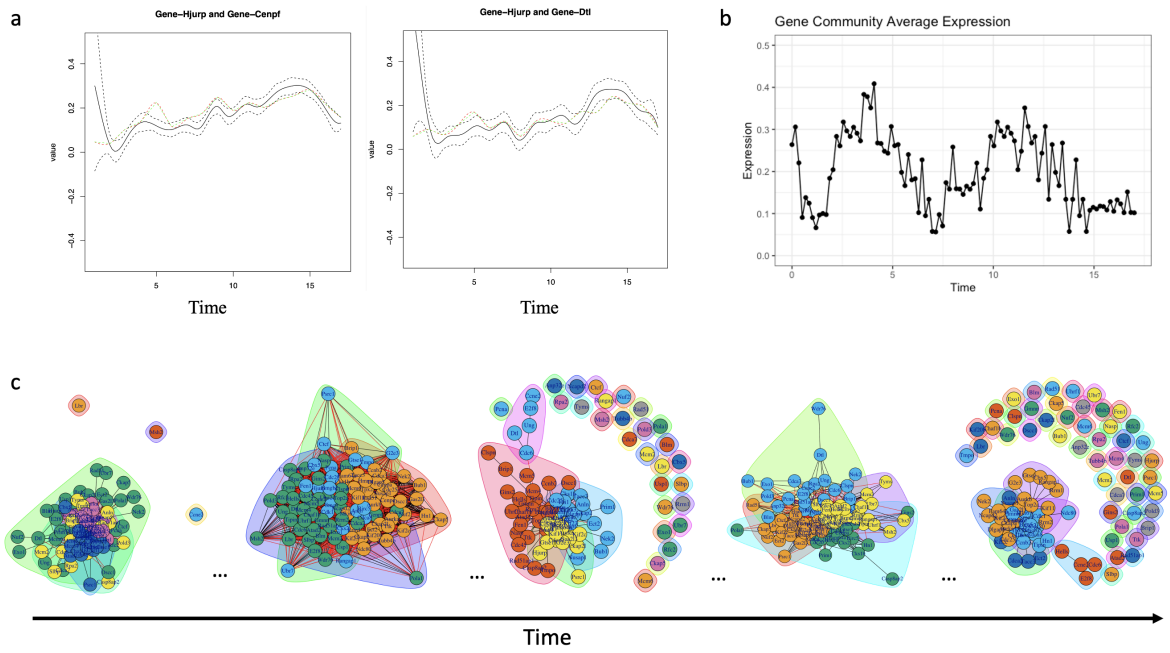

Figure 3.6: Cell progression in cell reprogramming dataset. **a**: Estimated dynamic relationship between pairwise gene Hjurp and Cenpf, Hjurp and Dtl. **b**: Average expression of gene community over time. **c**: Constructed dynamic network based on cell cycle genes.

gene relationship in the interphase. We observed that those dynamic gene-gene relationships consistently follow the natural cell cycle patterns. The periodical patterns validate the effectiveness of the dynamic gene networks by CellST. The effectiveness of CellST dynamic gene networks enables us to study the relationship between genes over time further.

3.2.3 Case study two: individual trajectories on zebrafish embryogenesis process

To further investigate individual cell progression behaviors and gene-gene relationship, we performed CellST on another zebrafish embryogenesis scRNA-seq dataset. This dataset contains 38,731 cells and 11,588 genes of early zebrafish development using Drop-seq (Macosko et al., 2015). Samples in the dataset are from high blastula stage (3.3 hours postfertilization, just after transcription from the zygotic genome

begins), when most cells are pluripotent, to six-somite stage (12 hours postfertilization, shortly after the completion of gastrulation), when many cells have differentiated into different cell types. The detail of data preprocessing is included in the method section. Since the cell type information is unknown in the dataset, cells in the dataset were clustered into 22 cell clusters (Figure 3.7b). We observed that cells were clustered together at the beginning high blastula stage and differentiated into different cell clusters in later development stages, which is consistent with the original paper. Therefore we treat these cell clusters as different cell types and built individual cell trajectories to observe individual cell progression behaviors.

We applied the CellST method to build individual cell trajectories (Figure 3.7a-3.7b) and reflect unique individual cell development behaviors. Unlike the bulk cell trajectory, the CellST individual cell trajectories achieved full cell development paths coverage for all cells. The full coverage indicates that the individual cell trajectories can reveal less frequent cell development patterns overlooked by the bulk cell trajectory. The CellST constructed individual cell trajectories throughout the stages and illustrated the unique individual cell development behaviors. The individual cell trajectories return each cell's potential cell development paths into different cell clusters throughout the 12 developmental stages. We compared the CellST trajectories to the average bulk single cell trajectory constructed by Monocle3 (Cao et al., 2019; McInnes et al., 2018). The starting points of each Monocle3 trajectory are given based on the developing stage (time points). The Monocle3 method constructed two average single cell bulk trajectories (Figure 3.7c and 3.7d), which are not consistent with the natural cell developing stages.

Furthermore, as cells developed into nine different cell clusters at the 12.0-6-somite stage (last developmental stage), we classified those trajectories according to the cell clusters in the last developmental stage. We constructed the dynamic gene networks (Figure 3.8b) for each group of individual cell trajectories. In those dynamic networks, we observed a few genes that behave significantly different from other genes (Figure 3.8b). For instance, DBX1A and ALPL.1 are two genes that appeared to behave differently in cluster 11 and cluster five in the last developmental stage. The functions of DBX1A gene are in regulating cell differentiation and developmental process (Gaudet et al., 2010; Gribble et al., 2007) and the functions of ALPL.1 gene are in regulating osteoblast differentiation and skeletal system process (Foster et al., 2005;

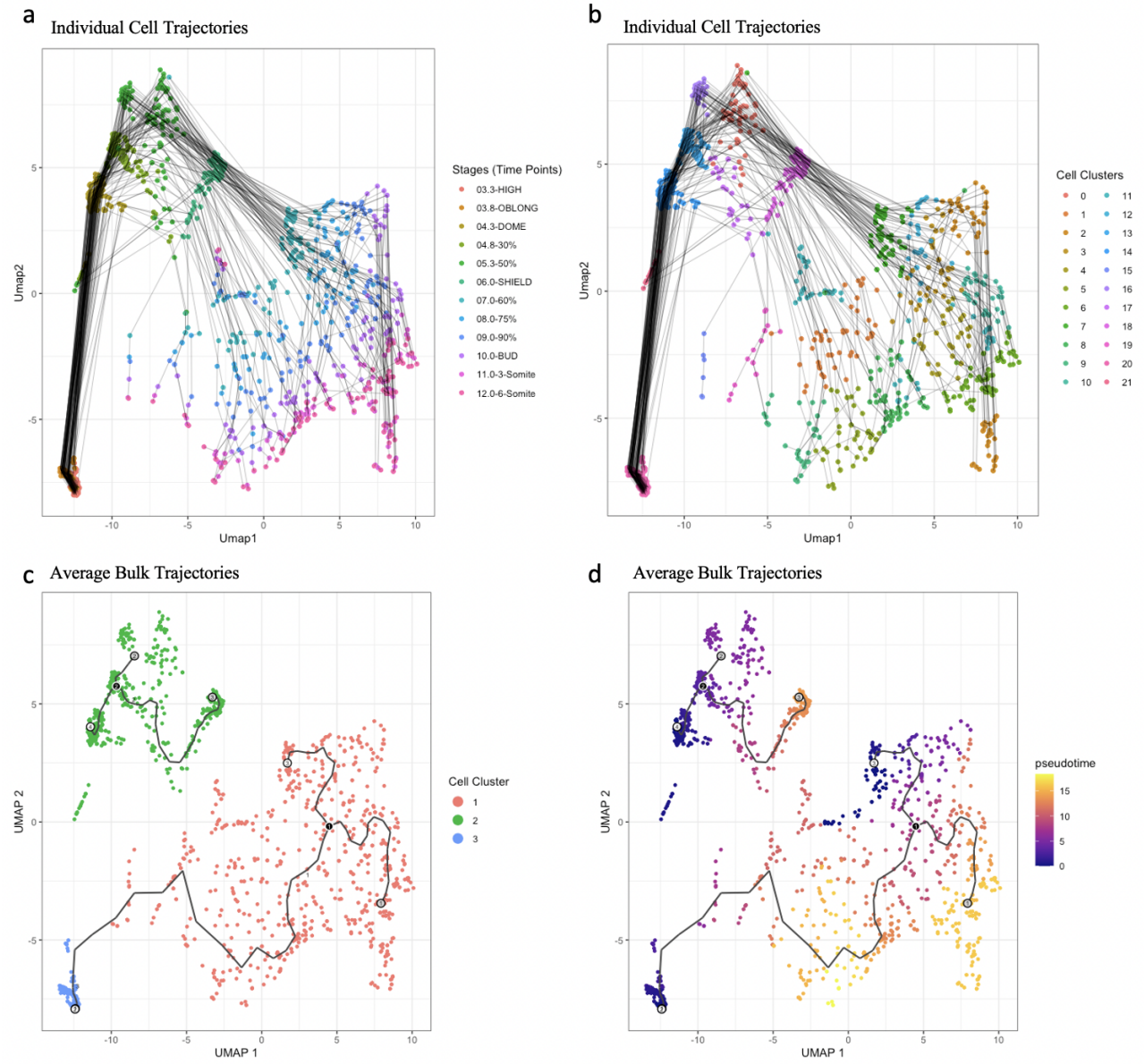

Figure 3.7: **a-b**: Individual cell trajectory in different development stages (**a**) and different cell clusters (**b**). **c-d**: The bulk pseudo time cell trajectories constructed by the Monocle3.

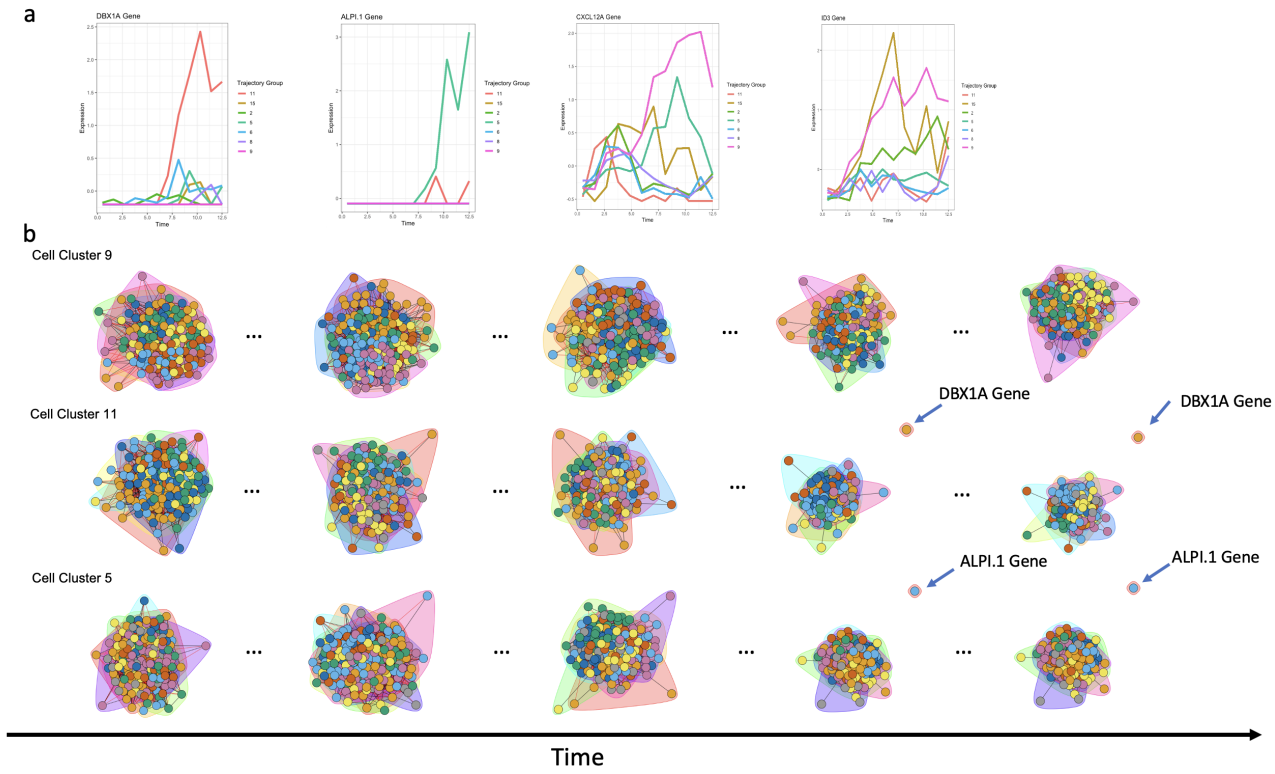

Figure 3.8: **a**: Dynamic gene networks constructed by CellST for different cell clusters at the last developing stage (12.0-6-somite stage). **b**: The behaviors of differentially expressed genes in different cell clusters.

Mornet et al., 1998). We further visualized the expressions of the two critical genes in the original 12 developmental stages. The expressions of DBX1A and ALPL.1 genes are significantly higher in cluster 11 and cluster five accordingly than in other cell clusters, consistent with the discovery in the CellST dynamic networks. Additionally, we performed functional deferentially expressed gene tests based on the CellST individual cell trajectories. We discovered a total of 268 differentially expressed genes in this Danio rerio cell development process dataset. For example, gene CXCL12A and ID3 are two differentially expressed genes in different cell clusters (Figure 3.8a). Next, we performed gene ontology annotations to those genes (table 3.1) and the function of those genes is highly related to regulating the cell development process.

Those results proved that the cell trajectories and dynamic gene networks in the CellST method are accurate and can be used to discover critical genes in cell progression. We also demonstrated the CellST

Table 3.1: Top six gene functional annotation clusters sort by p-values.

Gene Ontology (GO) annotations	Count	Enrichment Score	P-value
Developmental protein	119	5.808	2.090009e-54
DNA-binding	155	3.814	3.071105e-46
Multicellular organism development	126	4.357	9.100393e-44
DNA binding	192	2.891	3.617413e-41
Regulation of transcription	190	2.808	1.614839e-38
Homeobox	81	6.190	2.992261e-38

individual cell trajectories have a full coverage on different cell development behaviors. Those trajectories reflect unique gene expression patterns when cells develop into different cell types.

3.3 Individual cell trajectories and dynamic gene networks

In this section, we introduce the Cell Smooth Transformation (CellST) method. Through CellST method, we illustrate the way to construct the individual cell trajectories and dynamic gene networks for time course scRNA-seq data.

3.3.1 Individual cell trajectories

To conduct the individual cell trajectories, we first link the cells at different time points to construct the cells' correspondence information between time points. We then smooth the gene expressions pattern for each gene along time, and extract the "mean curve" of all individual gene expression patterns in a single individual cell trajectories to obtain the general gene expression pattern.

Cell linking by optimal transport

Linking cells and construct cell's correspondence between time points can be turned into a problem of domain adaptation. Specifically, we denote the normalized gene expressions for cell i at time t as a d-dimensional vector $\mathbf{x}_{t_i}$; each dimension of $\mathbf{x}_{t_i}$ represents a gene expression [1]. We further denote

$\mathbf{X}_t = \{\mathbf{x}_{t_i}\}_{i=1}^{n_t}$, where n_t indicates the number of cells at time t in scRNA-seq dataset. Our goal is to learn the transformation between the domain spaces through aligning the distribution of $\mathbf{X}_t$ to $\mathbf{X}_{t+1}$. In general, the optimal transport cell linking process can be summarized as the following three steps: (1): Estimate empirical distributions μ_t and μ_{t+1} from $\mathbf{X}_t$ and $\mathbf{X}_{t+1}$. (2): Find an optimal transport coupling map $\mathbf{T}$ from μ_t to μ_{t+1}. (3): Apply $\mathbf{T}$ to obtain individual cell couplings from $\mathbf{X}_t$ to $\mathbf{X}_{t+1}$. In the cell-linking case, the transport map $\mathbf{T}$ from μ_t to μ_{t+1} can be denoted as $\mathbf{T}(\mathbf{X}) = \Sigma\mathbf{X}$, where $\mathbf{X}_t = (\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, \ldots, \mathbf{x}_{t_n})^T$ and Σ is an $n_{t+1} \times n_t$ transformation matrix. The optimal transport coupling map $\mathbf{T}$ then can be calculated through the following optimization problem,

$$\min_{\Sigma} \sum_{i=1}^{n_t} \sum_{j=1}^{n_{t+1}} c(\mathbf{x}_{t_j}, \mathbf{x}_{t+1_i}) \Sigma_{i,j}, \tag{3.1}$$

where $c(\mathbf{x}_{t_j}, \mathbf{x}_{t+1_j})$ can be interpreted as the energy required to transform an individual cell from $\mathbf{x}_j^t$ to $\mathbf{x}_i^{t+1}$. We link cells according to $\mathbf{T}$ that minimizes the cost of transporting cells from one time points to another. More details of the cell linking estimation are in the supplementary file.

Individual cell trajectories by smoothing spline models

After we linked the cells from different time points, we can obtain the individual cell couplings at time points t and $t+1$. We then link cells for all time points based on the cell couplings to construct each cell's coarse individual cell trajectories across the timeline. Those cell trajectories are smoothed to reduce the estimation variance in CellST by utilizing the smoothing spline models. Smoothing spline models are a versatile family of smoothing methods that are suitable for both uni-variate and multivariate problems (Gu, 2013). We focus on the uni-variate problem and briefly illustrate the basic idea of smoothing spline models. To construct the proposed smoothed cell trajectories, we use equation 3.11 to model the behavior patterns of the gene expression along the individual cell trajectory (Gu & Ma, 2005). Let t represent the time points in time course dataset and g_i represent the gene expression for the aligned individual cell trajectories. As the genes are co-expressed with each other, we can model the gene behavior patterns using

a smoothing spline mix-effect model with $\{g_i, t_i\}_{i=1}^{d}$ as the observations:

$$g_i = \eta\left(t_i\right) + \mathbf{z}_i^T \mathbf{b} + \varepsilon_i, i = 1, \ldots, d, \tag{3.2}$$

where the regression function $\eta\left(t_i\right)$ is assumed to be a smooth function on the genes domain space in a cell. $\eta\left(t_i\right)$ are the fixed effects and $\mathbf{z}_i^T \mathbf{b}$ are the random effect with $\mathbf{b} \sim N(\mathbf{0}, B)$ and $\varepsilon_i \sim N\left(0, \sigma^2\right)$. The random effects are used to account for the co-expressed genes in one individual cell trajectory. The model terms $\eta(x)$ or $\eta(x) + \mathbf{z}^T \mathbf{b}$ are estimated using the penalized (unweighted) least squares method. Details of estimation steps are showed in the supplementary file. Since there are d gene expression patterns over t time points for each of the individual cell trajectory, the smoothing spline model estimates one expression patterns for individual cells and smooth the expression patterns.

3.3.2 Dynamic gene (co-expression) networks

We consider the connection of two genes to be dynamic and the relationship may smoothly change. Suppose we want to study the dynamic relationship of the lth gene and sth gene, where $1 \leq l, s \leq p, l \neq s$. Denote $X_i^{\langle l \rangle}(t)$ and $X_i^{\langle s \rangle}(t)$ as the lth gene and sth gene's expression values of individual cell trajectories i from previous steps, and $i = 1, \cdots, n$. By taking lth gene as the response and sth gene as the covariate, we consider the functional concurrent linear model,

$$X_i^{\langle l \rangle}(t) = \beta^{\langle l,s \rangle}(t) X_i^{\langle s \rangle}(t) + \varepsilon_{i,t}^{\langle l,s \rangle} \tag{3.3}$$

where $\beta^{\langle l,s \rangle}(t)$ models the dynamic linear relationship between two genes, $\varepsilon_{i,t}^{\langle l,s \rangle}$s are i.i.d. random errors with mean zero and constant variance. We estimate $\beta^{\langle l,s \rangle}(t)$ by minimizing the following penalized least squares functional,

$$\frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(X_i^{\langle l \rangle}\left(t_{ik}\right) - \beta^{\langle l,s \rangle}\left(t_{ik}\right) X_i^{\langle s \rangle}\left(t_{ik}\right) \right)^2 + \lambda J(\beta^{\langle l,s \rangle}) \tag{3.4}$$

where J is a quadratic functional denoting the roughness penalty on $\beta^{\langle l,s\rangle}$. The details of estimating minimization equation 3.17 can be found in the supplementary file. To measure the strength of all pairs of genes' relationships, we estimate the $\{\beta^{\langle l,s\rangle}\}$, for $l, s = 1, \cdots, p, l \neq s$. Based on those $\{\beta^{\langle l,s\rangle}\}$ and their confidence bands, we build the dynamic gene networks. The nodes indicate the genes and the edges indicate the measures of gene-gene relationships.

3.3.3 Functional differentially expressed genes test

We integrate an functional ANOVA test method (Górecki & Smaga, 2019) in our framework to estimate deferentially expressed genes using smoothed individual cell trajectories. For each gene in those individual cell trajectories, we consider independent vectors of random function $\mathbf{X}_{ki}(t) = (X_{ki1}(t), \ldots, X_{kid}(t))^\top$, where k indicates number of trajectory groups, i indicates cells and d indicates number of genes in one individual cell trajectory, defined over the interval I. In the multivariate analysis of variance problem for functional data (FMANOVA), we have to test the null hypothesis as follows:

$$\begin{aligned} H_0 &: \boldsymbol{\mu}_1(t) = \cdots = \boldsymbol{\mu}_k(t), t \in I \\ H_A &: \boldsymbol{\mu}_1(t) \neq \cdots \neq \boldsymbol{\mu}_k(t), t \in I \end{aligned} \tag{3.5}$$

The Wilk's lambda test statistics for testing significant different genes are approximated using fdANOVA method (Górecki & Smaga, 2019). The null distribution of test statistics is approximated by $F_{(l-1)\kappa,(n-l)\kappa}$-distribution, κ were estimated by the naive and biased-reduced methods (J. Zhang, 2014). The p-value is given by $P\left(F_{(l-1)\kappa,(n-l)\kappa} > F_n\right)$, where F_n denotes the test statistic. P-values for all genes tested were corrected by Benjamini & Yekutieli method (Benjamini & Yekutieli, 2001).

3.3.4 Gene ontology analysis

The Gene Ontology and function annotation process for all differentially expressed genes from functional ANOVA test and dynamic gene networks are performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 (Dennis et al., 2003). The clustering stringency used as

default medium, which keeps the balanced results. Other parameters including similarity term overlap = 4, similarity threshold = 0.35, initial group members = 4 and multi-linkage threshold = 50%. In DAVID, Fisher's Exact test is adopted to measure the gene-enrichment in annotation terms. Fisher's Exact p-values are computed by summing probabilities p over defined sets of tables (Prob $= \sum_A \rho$) and the resulting p-values were modified by Benjamini Hochberg correction (Benjamini & Hochberg, 1995).

3.3.5 Single cell RNA-Seq data processing

For simulation datasets, the cell linking simulation data were generated to benchmark the correct cell coupling pairs using splatter scRNA-Seq simulator (Zappia et al., 2017). The mean expression of genes are simulated from a Gamma distribution and the expression variation in the counts per cell are simulated from a log-normal distribution. The expression outlier probability is 0.05 and the differentially expressed probability is 0.4. The time course simulation dataset is generated with dynamic pathway patterns using CancerInSilico (Sherman et al., 2019). In all scenarios for dimension reduction, we generated 160 cells at each time points and 200 genes per cell. The simulated total experimental time is 72 hours and the cell noise rate is 0.1.

For the mouse reprogramming single cell development datasets, the genes in pathways were determined based on co-expressions with a given gene of interest. For each gene, co-expression signatures were computed by finding the set of genes with expressions in cells that are highly correlated with the gene of interest. We compute log-transformed normalized gene expression values for both simulation datasets using Scater R package (McCarthy et al., 2017). We selected 2000 top variable genes for real datasets and 200 for simulated datasets using the variance of standardized values, which were calculated by the Find Variable Features function in the R package Seurat (Satija et al., 2015).

3.4 Theoretical details in CellST method

In this supplementary section, we introduce more details for the cell smoothing transformation (CellST) method. We illustrate the detailed formulations to construct the individual cell trajectories and dynamic gene networks for time course scRNA-seq data. Furthermore, we provide a supplementary figure for the real data analysis.

3.4.1 Cell linking method using optimal transport

Regard cells at different time points as cells with same genes of different domain spaces. Linking cells at different time points is then turns to a problem of domain adaptation. Specifically, we denote the normalized gene expression for cell i at time t as a d-dimensional vector $\mathbf{x}_i^t$; each dimension of $\mathbf{x}_i^t$ represents a gene expression [2]. We further denote $\mathbf{X}_t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$, where n_t indicates the number of cells at time t in single cell RNA-seq dataset. Our goal is to learn the transformation between the domain spaces through aligning the distribution of $\mathbf{X}_t$ to $\mathbf{X}_{t+1}$.

As a powerful tool to learn the transformation from one probability measure to another, optimal transport has been applied to solve the domain adaptation problem (Courty et al., 2014). We thus apply optimal transport to obtain the domain adaptive coupling between $\mathbf{X}_t$ and $\mathbf{X}_{t+1}$. In other words, we transform the cell linking problem into the Monge's original formulation of the optimal transport problem corresponds to minimizing the cost for transporting a gene expression distribution μ_t and μ_{t+1} using a map $\mathbf{T}$:

$$\min_T \int_t c(x, T(x)) \mathrm{d}\mu_t(x), \quad \text{where} \quad T\#\mu_t = \mu_{t+1} \tag{3.6}$$

In equation 3.6, μ_t and μ_{t+1} are probability measures of $\mathbf{X}_t$ and $\mathbf{X}_{t+1}$ in $\mathbb{R}^d$, where d is the number of dimensions. We define the optimal transport map $\mathbf{T} : \mathbb{R}^d \to \mathbb{R}^d$, where $\mathbb{R}^d$ can be interpreted as the domain space for $\mathbf{x}_i^t$ or $\mathbf{x}_i^{t+1}$. In this optimal transport problems, one constraint for the transportation map $\mathbf{T}$ from a measure μ_t to a measure μ_{t+1} is the so-called "measurement-preserving", i.e., $\mathbf{T}\#\mu_t = \mu_{t+1}$. Here, $\#$ represents the push-forward operator, such that for any measurable $\mathbf{x} \subset \mathbb{R}^d$, $\mathbf{T}\#\mu_t(x) =$

$\mu_t(\mathbf{T}^{-1}(x))$. Among all the measurement-preserving maps, the optimal $\mathbf{T}$ is the one that minimizes the transportation cost.

Since we can only observe gene expressions for sample cells at each time points, we focus on the case where the measures are discrete. The measures μ_t and μ_{t+1} for gene features at time points t and $t+1$ are defined as:

$$\mu_t = \frac{1}{N}\sum_{i=1}^{N}\delta_{t_i} \quad \text{and} \quad \mu_{t+1} = \frac{1}{N}\sum_{j=1}^{N}\delta_{t+1_j} \tag{3.7}$$

where δ_x is the Dirac measure at location $x \in \mathbb{R}^d$ and where the position of the supporting points are $\mathbf{X}_t = \{\mathbf{x}_{t_i}\}_{i=1}^{n_t}$, where $\mathbf{x}_{t_i} \in \mathbb{R}^d$. Denote $\mathbf{X}_t = (\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, \ldots, \mathbf{x}_{t_n})^T$. In discrete cases, the transport $\mathbf{T}$ from μ_t to μ_{t+1} can be denoted as $\mathbf{T}(\mathbf{X}) = \Sigma\mathbf{X}$, where Σ is an $n_{t+1} \times n_t$ matrix. In this paper, we consider the equal-size mapping, i.e., $n_t = n_{t+1}$. Notice that in this case, the transport between $\mathbf{X}_t$ and $\mathbf{X}_{t+1}$ is a one-to-one assignment with permutation, Σ then can be regarded as a "permutation" matrix with the (i,j)th element:

$$\Sigma_{i,j} = \begin{cases} 1 & \text{if } \mathbf{T}(\mathbf{x}_{t_j}) = \mathbf{x}_{t+1_i}, \\ 0 & \text{otherwise} \end{cases}. \tag{3.8}$$

Furthermore, the transportation cost $C(\mathbf{T})$ defined in (3.6) can be calculated as:

$$C(\mathbf{T}) = \sum_{i=1}^{n_t}\sum_{j=1}^{n_{t+1}} c\left(\mathbf{x}_{t_j}, \mathbf{x}_{t+1_i}\right)\Sigma_{i,j} \tag{3.9}$$

where $c(\mathbf{x}_{t_j}, \mathbf{x}_{t+1_j})$ can be interpreted as the energy required to transform an individual cell from the stage as $\mathbf{x}_j^t$ to the stage as $\mathbf{x}_i^{t+1}$. The optimal transport map $\mathbf{T}$ then can be calculated through:

$$\min_{\Sigma}\sum_{i=1}^{n_t}\sum_{j=1}^{n_{t+1}} c(\mathbf{x}_{t_j}, \mathbf{x}_{t+1_i})\Sigma_{i,j}. \tag{3.10}$$

In general, the optimal transport cell linking process can be summarized as the following three steps:

- Estimate empirical gene feature distributions μ_t and μ_{t+1} as in (3.7).

- Find an optimal transport map $\mathbf{T}$ from μ_t to μ_{t+1} through (3.10).
- Apply $\mathbf{T}$ to obtain the cell-to-cell coupling from $\mathbf{X}_t$ to $\mathbf{X}_{t+1}$.

3.4.2 Individual cell trajectories by smoothing spline models

After we linked the cells from different time points, we can obtain the individual cell couplings at time points t and $t+1$. We then link cells for all time points based on the cell couplings to construct each cell's coarse individual cell trajectories across the timeline. Those cell trajectories are smoothed to reduce the estimation variance in CellST by utilizing the smoothing spline models. Smoothing spline models are a versatile family of smoothing methods that are suitable for both uni-variate and multivariate problems (Gu, 2013). We focus on the uni-variate problem and briefly illustrate the basic idea of smoothing spline models. To construct the proposed smoothed cell trajectories, we use equation 3.11 to model the behavior patterns of the gene expression along the individual cell trajectory. Let t represent the time points in time course dataset and g_i represent the gene expression for the aligned individual cell trajectories. As the genes are co-expressed with each other, we can model the gene behavior patterns using a smoothing spline mix-effect model with $\{g_i, t_i\}_{i=1}^n$ as the observations (Gu & Ma, 2005):

$$g_i = \eta\left(t_i\right) + \mathbf{z}_i^T\mathbf{b} + \varepsilon_i \tag{3.11}$$

$i = 1, \ldots, n$, where the regression function $\eta\left(t_i\right)$ is assumed to be a smooth function on the genes domain space in a cell. $\eta\left(t_i\right)$ are the fixed effects and $\mathbf{z}_i^T\mathbf{b}$ are the random effect with $\mathbf{b} \sim N(\mathbf{0}, B)$ and $\varepsilon_i \sim N\left(0, \sigma^2\right)$. The random effects are used to account for the co-expressed genes in one individual cell trajectory. The model terms $\eta(x)$ or $\eta(x) + \mathbf{z}^T\mathbf{b}$ will be estimated using the penalized (unweighted) least squares method through the minimization of:

$$\frac{1}{n}\sum_{i=1}^{n}\left(g_i - \eta\left(t_i\right) - \mathbf{z}_i^T\mathbf{b}\right)^2 + \frac{1}{n}\mathbf{b}^T\Sigma\mathbf{b} + \lambda J(\eta) \tag{3.12}$$

where $J(\eta)$ is used to quantify the smoothness of η, and λ is the smoothing parameter controlling the trade-off between the goodness-of-fit and the smoothness of η (Gu, 2013; Wahba, 1990). Consider the minimization of the least squares estimation (equation 3.12) for η in a d-dimensional space span $\{\xi_1, \ldots, \xi_q\}$. Functions in the space can be expressed as:

$$\eta(x) = \sum_{j=1}^{d} c_j \xi_j(x) = \boldsymbol{\xi}^T(x)\mathbf{c} \tag{3.13}$$

Plugging equation 3.13 into equation 3.12, thus η can be estimated by minimizing:

$$(\mathbf{g} - R\mathbf{c} - Z\mathbf{b})^T(\mathbf{g} - R\mathbf{c} - Z\mathbf{b}) + \mathbf{b}^T\Sigma\mathbf{b} + n\lambda\mathbf{c}^T Q\mathbf{c} \tag{3.14}$$

With the standard formulation of penalized least squares regression, the minimization of equation 3.12 is performed in a so-called reproducing kernel Hilbert space $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$ in which $J(\eta)$ is a square seminorm, and the solution resides in the space $\mathcal{N}_J \oplus \text{span}\{R_J(x_i, \cdot), i = 1, \ldots, n\}$, where $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ is the null space of $J(\eta)$ and $R_J(\cdot, \cdot)$ is the so-called reproducing kernel in $\mathcal{H} \ominus \mathcal{N}_J$. The solution has an expression:

$$\eta(x) = \sum_{i=1}^{m} d_\nu \phi_\nu(x) + \sum_{i=1}^{n} \tilde{c}_i R_J(x_i, x) \tag{3.15}$$

where $\{\phi_\nu\}_{\nu=1}^m$ is a basis of $\mathcal{N}_J$. It follows that $R = (S, \tilde{Q})$, where S is $n \times m$ with the (i, ν)th entry $\phi_\nu(x_i)$ and $\tilde{Q}$ is $n \times n$ with the (i, j) th entry $R_J(x_i, x_j)$. In the smoothing spline model, the estimation of η is highly related to the choosing of the smoothing parameter λ. We choose the smoothing parameter λ by Generalized Cross-Validation (GCV) (Gu, 2013; Wahba, 1990). Since there are d gene expression patterns over t time points for each of the individual cell trajectory, the smoothing spline model estimates one expression patterns for individual cells and smooth the expression patterns.

3.4.3 Dynamic gene networks

We consider the connection of two genes to be dynamic and the relationship may smoothly change. Suppose we want to study the dynamic relationship of the lth gene and sth gene, where $1 \leq l, s \leq p, l \neq s$. Denote $X_i^{\langle l \rangle}(t)$ and $X_i^{\langle s \rangle}(t)$ as the lth gene and sth gene's expression values of individual cell trajectories i, and $i = 1, \cdots, n$. By taking lth gene as the response and sth gene as the covariate, we consider the functional concurrent linear model,

$$X_i^{\langle l \rangle}(t) = \beta^{\langle l,s \rangle}(t) X_i^{\langle s \rangle}(t) + \varepsilon_{i,t}^{\langle l,s \rangle} \tag{3.16}$$

where $\beta^{\langle l,s \rangle}(t)$ models the dynamic linear relationship between two genes, $\varepsilon_{i,t}^{\langle l,s \rangle}$s are i.i.d. random errors with mean zero and constant variance. We estimate $\beta^{\langle l,s \rangle}(t)$ by minimizing the following penalized least squares functional,

$$\frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(X_i^{\langle l \rangle}(t_{ik}) - \beta^{\langle l,s \rangle}(t_{ik}) X_i^{\langle s \rangle}(t_{ik}) \right)^2 + \lambda J(\beta^{\langle l,s \rangle}) \tag{3.17}$$

where J is a quadratic functional denoting the roughness penalty on $\beta^{\langle l,s \rangle}$. Throughout this paper, we consider $J(\beta) = \int_\Gamma (\beta^{(m)})^2 dt$. We assume that the unknown function $\beta^{\langle l,s \rangle}$ is smooth and resides in a reproducing kernel Hilbert Space $\mathcal{H}$. $\lambda > 0$ is the smoothing parameter balancing the trade-off between the goodness-of-fit and penalties.

By decomposing the space $\mathcal{H}$ as $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_1 = \{\beta : J(\beta) = 0\}$ is the null space of $J(\beta)$, and $\mathcal{H}_1$ is the orthogonal complement of $\mathcal{H}_1$ in $\mathcal{H}$. With the representer theorem (Wahba, 1990), the optimizer of 3.17 can be written as

$$\hat{\beta}_\lambda^{\langle l,s \rangle}(t) = \sum_{v=1}^{m} d_v \psi_v(t) + \sum_{i}^{n} \sum_{k}^{K} c_{ik} R_1(t_{ik}, t) \tag{3.18}$$

where $\{\psi_v\}_{v=1}^m$ is the basis function of the m-dimensional null space $\mathcal{H}_0$, and $R_J(\cdot,\cdot)$ is the reproducing kernel of $\mathcal{H}_1$. d_v and c_{ik} are the coefficients. By Plugging equation 3.18 to equation 3.17, we can yield the estimation of $c = (c_1, \cdots, c_{1K}, \cdots, c_{n1}, \cdots, c_{nK})^T$ and $d = (d_1, \cdots, d_{1K}, \cdots, d_{n1}, \cdots, d_{nk})^T$, which follows,

$$\begin{aligned} \boldsymbol{c} &= \left(\mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{S}\left(\mathbf{S}^T\mathbf{M}^{-1}\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{M}^{-1}\right)\mathbf{X}^{\langle s\rangle}\boldsymbol{X}^{\langle l\rangle} \\ \boldsymbol{d} &= \left(\mathbf{S}^T\mathbf{M}^{-1}\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{M}^{-1}\boldsymbol{X}^{\langle l\rangle} \end{aligned} \tag{3.19}$$

where $\mathbf{X}^{\langle s\rangle} = diag((X_1^{\langle s\rangle T}, \cdots, X_n^{\langle s\rangle T}))$ with the vector $X_i^{\langle s\rangle T} = (X_i^{\langle s\rangle}(t_{i1}), \cdots, X_i^{\langle s\rangle}(t_{iK}))^T$, $X^{\langle l\rangle} = (X_1^{\langle l\rangle T}, \cdots, X_n^{\langle l\rangle T}))^T$ with the vector $X_i^{\langle l\rangle} = (X_i^{\langle l\rangle}(t_{i1}), \cdots, X_i^{\langle l\rangle}(t_{iK}))^T$, $\mathbf{S} = (\mathbf{S}_1^T, \cdots, \mathbf{S}_n^T)^T$ with the (k,v)th entry of the $K \times m$ matrix $\mathbf{S}_i$ equals to $\psi_v(t_{ik})X_i^{\langle s\rangle}(t_{ik})$, $\mathbf{M} = \mathbf{X}^{\langle s\rangle}\mathbf{Q}\mathbf{X}^{\langle s\rangle} + n\lambda\mathbf{I}$ and $\mathbf{Q}$ is the $nK \times nK$ block matrix with the (i,j)th block is the $K \times K$ matrix with the (k,u)th entry equals to $R_1(t_{ik}, t_{ju})$. Thus, the estimation of $\beta^{\langle l,s\rangle}(t)$ can be written as

$$\hat{\beta}^{\langle l,s\rangle}(t) = \boldsymbol{\psi}^T\boldsymbol{d} + \boldsymbol{\xi}^T\boldsymbol{c} \tag{3.20}$$

where $\psi = (\psi_1(t), \cdots, \psi_m(t))^T$ and $\xi = (R_1(t_{11}, t), \cdots, R_1(t_{1K}, t), \cdots, R_1(t_{n1}, t), \cdots, R_1(t_{nK}, t))^T$. Note that $\beta^{\langle l,s\rangle}(t)$ models the dynamic linear relationship between lth gene and sth gene, and $\beta^{\langle l,s\rangle}(t_0) = 0$ means the correlation between gene l and gene s to be o at the time point t_0. Naturally, for the time point t_0, we need a threshold $\gamma(t_0)$ to decide whether make a connection between these two genes by checking whether $|\hat{\beta}^{\langle l,s\rangle}(t_0)| > \gamma(t_0)$. We then derive the $100(1-\alpha)\%$ confidence band of $\beta^{\langle l,s\rangle}(t)$ to decide the connection threshold. We adapt the Bayes model in (Gu, 2013) and get the posterior variance of $\beta^{\langle l,s\rangle}(t)$ satisfies

$$\mathrm{Var}\left[\beta^{\langle l,s\rangle}(t) \mid \mathbf{X}, \mathbf{X}^{\langle l\rangle}\right] = \frac{\sigma^2}{nK\lambda}\left(R_1(t,t) + \boldsymbol{\psi}^T\left(\mathbf{S}^T\mathbf{M}^{-1}\mathbf{S}\right)^{-1}\boldsymbol{\psi} - 2\boldsymbol{\psi}^T\boldsymbol{d}_\xi - \boldsymbol{\xi}^T\boldsymbol{c}_\xi\right) \tag{3.21}$$

where

$$\begin{aligned} \boldsymbol{c}_\xi &= \left(\mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{S}\left(\mathbf{S}^T\mathbf{M}^{-1}\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{M}^{-1}\right)\mathbf{X}^{\langle s\rangle}\boldsymbol{\xi} \\ \boldsymbol{d}_\xi &= \left(\mathbf{S}^T\mathbf{M}^{-1}\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{M}^{-1}\mathbf{X}^{\langle s\rangle}\boldsymbol{\xi} \end{aligned} \tag{3.22}$$

Using equation (3.21), we can estimate the posterior variance of $\beta^{\langle l,s\rangle}(t_0)$ and write as $\gamma^{\langle l,s\rangle}(t_0)$. Thus, by setting the significance level at α, we make a connection for gene l and gene s at time point t_0 if $|\hat{\beta}^{\langle l,s\rangle}(t_0)| > z_{\alpha/2}\gamma^{\langle l,s\rangle}(t_0)$ and $|\hat{\beta}^{\langle s,l\rangle}(t_0)| > z_{\alpha/2}\gamma^{\langle s,l\rangle}(t_0)$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile for standard normal distribution.

3.5 Individual cell trajectories in different pathways

In the mouse cell reprogramming dataset, we constructed individual cell trajectories for gene sets in six different pathways. Those trajectories patterns were compared with Waddington-OT method.

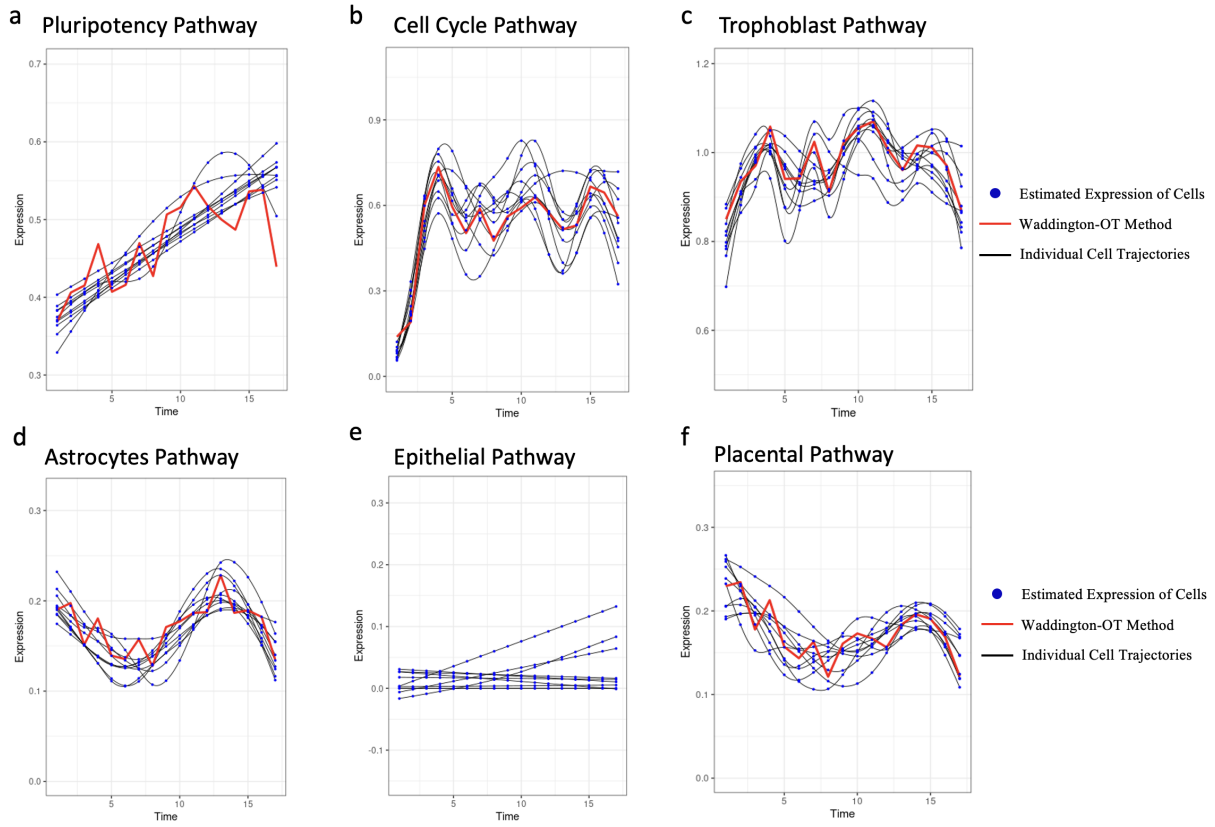

Figure 3.9: Individual cell trajectories constructed by CellST for all pathways estimated for the mouse reprogramming dataset

References

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D. et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, *182*(1), 145–161.

An, S., Ma, L. & Wan, L. (2019). Tsee: An elastic embedding method to visualize the dynamic gene expression patterns of time series single-cell rna sequencing data. *BMC genomics*, *20*(2), 224.

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T. et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, *13*(3), 229–232.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, *57*(1), 289–300.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 1165–1188.

Burrows, N., Bashford-Rogers, R. J., Bhute, V. J., Peñalver, A., Ferdinand, J. R., Stewart, B. J., Smith, J. E., Deobagkar-Lele, M., Giudice, G., Connor, T. M. et al. (2020). Dynamic regulation of hypoxia-inducible factor-1α activity is essential for normal b cell development. *Nature Immunology*, *21*(11), 1408–1420.

Cannoodt, R., Saelens, W. & Saeys, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *European journal of immunology*, *46*(11), 2496–2506.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J. et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), 496–502.

Chen, H., Albergante, L., Hsu, J. Y., Lareau, C. A., Bosco, G. L., Guan, J., Zhou, S., Gorban, A. N., Bauer, D. E., Aryee, M. J. et al. (2019). Single-cell trajectories reconstruction, exploration and mapping of omics data with stream. *Nature communications*, *10*(1), 1–14.

Courty, N., Flamary, R. & Tuia, D. (2014). Domain adaptation with regularized optimal transport. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 274–289.

Dai, K., Damodaran, K., Venkatachalapathy, S., Soylemezoglu, A. C., Shivashankar, G. & Uhler, C. (2020). Predicting cell lineages using autoencoders and optimal transport. *PLoS Computational Biology*, *16*(4), e1007828.

De Meo, P., Ferrara, E., Fiumara, G. & Provetti, A. (2011). Generalized louvain method for community detection in large networks. *2011 11th International Conference on Intelligent Systems Design and Applications*, 88–93.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. & Lempicki, R. A. (2003). David: Database for annotation, visualization, and integrated discovery. *Genome Biology*, *4*(9), 1–11.

Foster, L. J., Zeemann, P. A., Li, C., Mann, M., Jensen, O. N. & Kassem, M. (2005). Differential expression profiling of membrane proteins by quantitative proteomics in a human mesenchymal stem cell line undergoing osteoblast differentiation. *Stem Cells*, *23*(9), 1367–1377.

Gaudet, P., Livstone, M. & Thomas, P. (2010). Annotation inferences using phylogenetic trees.

Górecki, T. & Smaga, Ł. (2019). Fdanova: An r software package for analysis of variance for univariate and multivariate functional data. *Computational Statistics*, *34*(2), 571–597.

Gribble, S. L., Nikolaus, O. B. & Dorsky, R. I. (2007). Regulation and function of dbx genes in the zebrafish spinal cord. *Developmental dynamics: An Official Publication of the American Association of Anatomists*, *236*(12), 3472–3483.

Grün, D. & van Oudenaarden, A. (2015). Design and analysis of single-cell sequencing experiments. *Cell*, *163*(4), 799–810.

Gu, C. (2013). *Smoothing spline anova models* (Vol. 297). Springer Science & Business Media.

Gu, C. & Ma, P. (2005). Optimal smoothing in nonparametric mixed-effect models. *The Annals of Statistics*, *33*(3), 1357–1379.

Guo, J., Grow, E. J., Yi, C., Mlcochova, H., Maher, G. J., Lindskog, C., Murphy, P. J., Wike, C. L., Carrell, D. T., Goriely, A. et al. (2017). Chromatin and single-cell rna-seq profiling reveal dynamic signaling and metabolic transitions during human spermatogonial stem cell development. *Cell Stem Cell*, *21*(4), 533–546.

Hrvatin, S., Hochbaum, D. R., Nagy, M. A., Cicconet, M., Robertson, K., Cheadle, L., Zilionis, R., Ratner, A., Borges-Monroy, R., Klein, A. M. et al. (2018). Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nature neuroscience*, *21*(1), 120–129.

Ji, Z. & Ji, H. (2016). Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, *44*(13), e117–e117.

Klimovskaia, A., Lopez-Paz, D., Bottou, L. & Nickel, M. (2020). Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature Communications*, *11*(1), 1–9.

Ko, M. E., Williams, C. M., Fread, K. I., Goggin, S. M., Rustagi, R. S., Fragiadakis, G. K., Nolan, G. P. & Zunder, E. R. (2020). Flow-map: A graph-based, force-directed layout algorithm for trajectory mapping in single-cell time course datasets. *Nature protocols*, *15*(2), 398–420.

Lawson, D. A., Bhakta, N. R., Kessenbrock, K., Prummel, K. D., Yu, Y., Takai, K., Zhou, A., Eyob, H., Balakrishnan, S., Wang, C.-Y. et al. (2015). Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature*, *526*(7571), 131–135.

Liu, Z., Lou, H., Xie, K., Wang, H., Chen, N., Aparicio, O. M., Zhang, M. Q., Jiang, R. & Chen, T. (2017). Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nature communications*, *8*(1), 1–9.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M. et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202–1214.

McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. (2017). Scater: Pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, *33*(8), 1179–1186.

McInnes, L., Healy, J. & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mendonsa, A. M., Na, T.-Y. & Gumbiner, B. M. (2018). E-cadherin in contact inhibition and cancer. *Oncogene*, *37*(35), 4769–4780.

Meng, C., Ke, Y., Zhang, J., Zhang, M., Zhong, W. & Ma, P. (2019). Large-scale optimal transport map estimation using projection pursuit. *Advances in Neural Information Processing Systems*, 8116–8127.

Moon, K. R., Stanley III, J. S., Burkhardt, D., van Dijk, D., Wolf, G. & Krishnaswamy, S. (2018). Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, *7*, 36–46.

Mornet, E., Taillandier, A., Peyramaure, S., Kaper, F., Muller, F., Brenner, R., Bussiere, P., Freisinger, P., Godard, J., Le Merrer, M. et al. (1998). Identification of fifteen novel mutations in the tissue-nonspecific alkaline phosphatase (tnsalp) gene in european patients with severe hypophosphatasia. *European Journal of Human Genetics*, *6*(4), 308–314.

Nawy, T. (2013). Single-cell sequencing. *Nature methods*, *11*(1), 18.

Pavel, M., Renna, M., Park, S. J., Menzies, F. M., Ricketts, T., Füllgrabe, J., Ashkenazi, A., Frake, R. A., Lombarte, A. C., Bento, C. F. et al. (2018). Contact inhibition controls cell survival and proliferation via yap/taz-autophagy axis. *Nature Communications*, *9*(1), 1–18.

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A. & Trapnell, C. (2017). Single-cell mrna quantification and differential analysis with census. *Nature methods*, *14*(3), 309.

Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., Larson, D. R. & Zhao, K. (2017). Ctcf-mediated enhancer-promoter interaction is a critical regulator of cell-to-cell variation of gene expression. *Molecular Cell*, *67*(6), 1049–1058.

Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature biotechnology*, *37*(5), 547–554.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5), 495–502.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P. et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, *176*(4), 928–943.

Shapiro, E., Biezuner, T. & Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, *14*(9), 618–630.

Sherman, T. D., Kagohara, L. T., Cao, R., Cheng, R., Satriano, M., Considine, M., Krigsfeld, G., Ranaweera, R., Tang, Y., Jablonski, S. A. et al. (2019). Cancerinsilico: An r/bioconductor package for combining mathematical and statistical modeling to simulate time course bulk and single cell gene expression data in cancer. *PLoS Computational Biology*, *14*(4), e1006935.

Skinnider, M. A., Squair, J. W. & Foster, L. J. (2019). Evaluating measures of association for single-cell transcriptomics. *Nature Methods*, *16*(5), 381–386.

Spiller, D. G., Wood, C. D., Rand, D. A. & White, M. R. (2010). Measurement of single-cell dynamics. *Nature*, *465*(7299), 736–745.

Stegle, O., Teichmann, S. A. & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, *16*(3), 133–145.

Sun, C., Wang, H., Ma, Q., Chen, C., Yue, J., Li, B. & Zhang, X. (2021). Time-course single-cell rna sequencing reveals transcriptional dynamics and heterogeneity of limbal stem cells derived from human pluripotent stem cells. *Cell & Bioscience*, *11*(1), 1–12.

Tanay, A. & Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. *Nature*, *541*(7637), 331–338.

Tomasetti, C., Durrett, R., Kimmel, M., Lambert, A., Parmigiani, G., Zauber, A. & Vogelstein, B. (2017). Role of stem-cell divisions in cancer risk. *Nature*, *548*(7666), E13–E14.

Tong, A., Huang, J., Wolf, G., van Dijk, D. & Krishnaswamy, S. (2020). Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. *arXiv preprint arXiv:2002.04461*.

Torii, K., Kubota, A., Araki, T. & Endo, M. (2020). Time-series single-cell rna-seq data reveal auxin fluctuation during endocycle. *Plant and Cell Physiology*, *61*(2), 243–254.

Traag, V. A., Waltman, L. & van Eck, N. J. (2019). From louvain to leiden: Guaranteeing well-connected communities. *Scientific reports*, *9*(1), 1–12.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome research*, *25*(10), 1491–1498.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, *32*(4), 381.

Tritschler, S., Büttner, M., Fischer, D. S., Lange, M., Bergen, V., Lickert, H. & Theis, F. J. (2019). Concepts and limitations for learning developmental trajectories from single cell genomics. *Development*, *146*(12).

Van den Berge, K., De Bezieux, H. R., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S. & Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications*, *11*(1), 1–13.

Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G. & Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, *360*(6392), 981–987.

Wahba, G. (1990). *Spline models for observational data* (Vol. 59). SIAM.

Wang, J.-L., Chiou, J.-M. & Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, *3*, 257–295.

Yuan, Y. & Bar-Joseph, Z. (2021). Deep learning of gene relationships from single cell time-course expression data. *Briefings in Bioinformatics*, *22*(5), bbab142.

Zappia, L., Phipson, B. & Oshlack, A. (2017). Splatter: Simulation of single-cell rna sequencing data. *Genome Biology*, *18*(1), 1–15.

Zhang, J. (2014). Analysis of variance for functional data. *Monographs on Statistics and Applied Probability*, *127*, 127.

Zhang, J., Zhong, W. & Ma, P. (2020). A review on modern computational optimal transport methods with applications in biomedical research. *arXiv preprint arXiv:2008.02995*.

Chapter 4

Novel Analysis Pipelines with Applications in Bioinformatics, Human Dynamic Big Data and Biological Behaviors Dataset

4.1 Introduction in locomotor behavior and human dynamic

With the rapid development of science and technology, large and complex data have been generated in many science areas especially with biological science and human health care. In this chapter, I propose a series of novel analysis pipelines that focus to elucidate the complex large scale data generated from wearable device that measures human dynamic as well as modern biological dataset including locomotor behavior data and Next-Generation Sequencing (NGS) dataset.

Locomotor behavior data The locomotor behavior data of zebrafish is of high-throughput, time-related and involves both experimental and biological variables. Its systematic studies have provided new insights into neurobiology, pharmacology, and toxicology. However, the complexity of these locomotor

data has created challenges in data analysis, which may potentially limit the advancement of neurobehavioral studies. To address the challenges brought by high-throughput behavior data, we have established a coherent statistical analysis framework for analyzing such data. In this section, I compared the time-related behavior profiles of zebrafish in several commonly-used scenarios. This study addressed the normalization need by establishing an approach based on linear-regression modeling. The model was established using a dataset of visual-motor response (VMR) obtained from several strains of wild-type (WT) zebrafish collected at multiple stages of development. This normalization approach explicitly modeled the effect of some systematic variations on VMR, such as the light emitted by the machine and biological replicates. This method also normalizes the activity profiles of different conditions to a common baseline. This approach is versatile, as it can incorporate different normalization needs as separate factors. In this section, I performed the Hotelling's T-squared test and the High-Dimensional Hypothesis test for zebrafish drug screening.

Human dynamic data The human dynamic trajectories data collected through wearable devices contain continuously precise GPS and physical activities. Such data can be used to study the dynamic patterns of human behavior (Barabasi, 2005). Employees working full-time in Georgia state were recruited for the study. Participants were asked to wear GPS devices and Physical Activity Monitors at the same time for up to two weeks to capture geolocation data aligned with their physical activities. GPS data were sampled at 30-seconds epochs and merged with the accelerometer data using the Personal Activity and Location Measurement System with the default settings. Data contain steps, longitude, latitude, elevation change, and activity intensity on each 30s time segment (Gay et al., 2018; Gay et al., 2017). Candidates were separated into the office worker group and non-office worker group based on their job functions.

The goal is to learn the dynamic patterns in the two groups of candidates. The best way to represent human dynamic patterns is to build trajectory networks for candidates (Sun et al., 2019). The nodes are candidate's frequent visiting places such as home, workplaces, favorite grocery stores, or restaurants. The nodes also contain information about the candidate's visiting time and physical activities in the nodes. One candidate can be represented by one network graph. We propose a two-layer graph convolutional

network (GCN) framework for graph classification on the data and achieve $> 85\%$ on the testing accuracy (Kipf & Welling, 2016; Schlichtkrull et al., 2018; Ying et al., 2018).

To train and learn the graph neural network in such a large-scale human dynamic network efficiently, we first batch multiple graphs together to form a mini-batch in each training epoch. A batch of graphs can be viewed as a large graph that has many disjointed connected components. Secondly, we implement and deploy the proposed GNN model on a decentralized computing platform (G. Zhang et al., 2019; Zhao et al., 2019; Zhu et al., 2019). To optimize the representation of a target node in a graph, we only need to look up the representations of its direct neighbors. In this situation, the human dynamic network G has to be formed as adjacency lists (i.e., each record contains one target node and all of its neighbors together) first. Then the whole set of adjacency lists is divided into several parts and stored in the memory of several machines. Hence, the neighbor lookup procedure for a certain node will only happen in one machine, which contributes to shorten the communication time between different machines and make the training procedure efficient

Arabidopsis root microbiota metagenomics study. Plants are naturally associated with root microbiota, which are microbial communities influential in hosting fitness. Thus, it is important to understand how plants control root microbiota. Epigenetic factors regulate the readouts of genetic information and consequently many essential biological processes. However, it has been elusive whether RNA-directed DNA methylation (RdDM) affects root microbiota assembly. In this paper(**kaushal2021dicer**), I analyzed the metagenomics data and investigate root microbiota of Arabidopsis mutants defective in the canonical RdDM pathway, including dcl234 that harbors a triple mutation in the Dicer-like proteins DCL3, DCL2, and DCL4, which produce small RNAs for RdDM. I performed gene analysis using shotgun sequencing of the root microbiome. The results demonstrate an important role of the DCL proteins in influencing root microbiota through integrated regulation of plant defense, cell wall compositions, and root exudates. Moreover, the canonical RdDM is dispensable for Arabidopsis root microbiota. These findings establish a connection between root microbiota and plant epigenetic factors and highlight the complexity of plant regulation of root microbiota.

Monterey Bay marine metagenomics study. Monterrey Bay is part of the California Current ecosystem. Our collaborator from Marine Sciences Dept at UGA collected a total of 83 microbial samples over two months from Monterey Bay. Simultaneously, various environmental features were recorded, including temperature, salinity, oxygen, nutrients, and several other physicochemical and biological parameters. To perform metagenomics analysis, contigs were assembled using reads in all samples together and binned by our lab's analysis tool (MetaGen). I discovered 416 metagenomic bins with their relative abundances across the 83 samples. To assign each contig bin into an individual genome, I predicted genes and other marker information. Then I created a multiple sequence alignment based on the identified maker genes to determine the most likely genomes and classify the contig bins into those genomes. Each binned metagenome-assembled genome is considered a species. I successfully assigned 241 contigs bins into known genomes in bacteria and archaea from the database (GTDB). Among those identified genomes, I observed the dynamic relative abundance of *Rhodobacteraceae*, which were proposed to be associated with white syndrome disease in coral. Moreover, I observed bacteria from the family of *Flavobacteriaceae*, which were recorded to be associated with rainbow trout fry syndrome or bacterial cold-water disease in marine life.

The rest of this chapter is organized as the following: Section 2 illustrates a project that propose a new normalization method for zebrafish behavior dataset. Section 3 proposed a new classification analysis using graph-neural network and smoothing spline models. Section 4 shows some exploratory bioinformatics analysis pipeline based on long reads sequencing technology which are called the third generation sequencing.

4.2 Multivariate analysis on large-scale behavioural data collected from zebrafish

4.2.1 Zebrafish behavioural data

Neuroscience research has been revolutionized by experimental approaches that can collect behavioural data simultaneously from multiple individual animals, including worms (Swierczek et al., 2011), fruit flies (Branson et al., 2009), rodents (Alexandrov et al., 2015) and zebrafish (Bruni et al., 2014). When these animals are also perturbed by genetical or pharmacological means, their resulting behavioural data would reveal the underlying neural circuitry that drives the behaviour (Bruni et al., 2014; Rihel et al., 2010), or reveal new drugs for treating neurological diseases (Alexandrov et al., 2015; Bruni et al., 2014; Ganzen et al., 2017). However, these behavioural data are complex in structure and pose many challenges to data analysis. These challenges must be resolved by appropriate statistical approaches to extract accurate information from the behavioural data.

To illustrate the data complexity and analytical challenges, we will outline a popular high-throughput approach for analysing zebrafish behaviour, the visual motor response (VMR). This is a locomotor response displayed by zebrafish larvae upon drastic light onset (Light-On) or offset (Light-Off) (Emran et al., 2007; Emran et al., 2008; Ganzen et al., 2017; L. Zhang et al., 2012). In a typical VMR experiment, zebrafish larvae are arranged in a 96-well plate and stimulated by a controlled light source in a lightproof chamber. These larvae can have different genotype or are exposed to different chemical treatments. Their resulting swimming activities are recorded and summarized as number of detected pixels moved in successive frames in the video, or as absolute displacement (Liu et al., 2015). These larvae are usually subjected to multiple trials of Light-On and Light-Off over the course of a long period of time (i.e. technical repeats). The experiment is often repeated using independent samples (i.e. biological repeats). The activity of larvae is then extracted from the video, which in turn results in a huge matrix of activity values of many larvae over time.

This experimental design poses challenges to data analysis by traditional statistical approaches including t-test and ANOVA (Scott et al., 2016) because they cannot not handle time-series data (i.e. data with time-dependency). Consequently, the VMR data have been analysed by advance approaches including repeated-measured ANOVA (de Esch et al., 2012; Fernandes et al., 2012; Kopp et al., 2018; Vignet et al., 2013) that can handle samples that are repeatedly measured and correlated in time. Our group has also established Hotelling's T-squared test (Liu et al., 2015), multivariate analysis of variance (MANOVA) (Liu et al., 2015), and generalized linear mixed model (GLMM) (Liu et al., 2017) for VMR data analysis. These approaches take into consideration of unique features of VMR data, such as time dependency among the VMR of individual animals and joint property of all VMR profiles. They also incorporate potential sources of batch effect in the analysis, and allow for proper comparisons between different sample groups.

These analyses, however, do not address another intrinsic issue of VMR data: these data are collected from individual larvae subjected to systematic variations that require normalization. For example, under a particular intensity setting of stimulating light, the larvae in different wells of the 96-well plate may receive slightly different light intensities from the machine. This issue is created by the physical constraint of light generation. Inside the machine, the stimulating light is generated by arrayed LEDs. Since they generate light as point source, they will not evenly illuminate all wells even with a diffuser. When the larvae in the plate are exposed to slightly different light intensities, their resulting VMR may be slightly different. Another example of systematic variation is biological replication. When an experiment is repeated, the biological samples may subject to unwanted variations, including day-day variation in the quality of the embryos, even when they are collected from the same parents. These systematic variations must be corrected by normalization, an approach to adjust values measured on different scales to the same scale for meaningful comparisons between different conditions. In this study, we present a normalization approach for VMR data based on linear-regression modeling. This model-based normalization handles different types of systematic biases separately or together, which allows users to choose specific variations to normalize in their studies. This approach complements the aforementioned statistical analyses for

VMR data. Together, they establish an essential framework for analysing high-throughput behavioural data with a similar structure.

4.2.2 Statistical and machine learning analysis

Experimental data The VMR data analysed in this paper were previously collected (Liu et al., 2015) and were downloaded from the Harvard Dataverse http://dx.doi.org/10.7910/DVN/HTXXKW. The dataset comprises activities collected from three wild-type (WT) zebrafish strains: AB, TL and TLAB. For each strain, the VMR data were collected daily from 3 days post-fertilization (dpf) to 9 dpf, using a standard experimental scheme (see S1 Fig) (Emran et al., 2007; Emran et al., 2008; Gao et al., 2014; Gao et al., 2016; Liu et al., 2015; L. Zhang et al., 2016). In this scheme, the larvae were arrayed in a 96-well plate. The plate was placed in a Zebrabox system (ViewPoint Life Sciences, Lyon, France) and received light stimulus from a light-controlling unit positioned under the plate. The light intensity of each well was measured by an ILT950 spectrometer (International Light Technologies, Peabody, MA). During an experiment, the plate was first dark-adapted for 3.5 hours (hrs). It was then subjected to three consecutive periods of light onset (Light-On) and light offset (Light-Off). Each of those periods lasted for 30 minutes (mins). Several variables that might affect larval activities were controlled (Liu et al., 2015). For instance, all experiments were conducted at the same time of the day with the same type of 96-well plate. Each strain was also individually analysed on separate plates. The research protocol was reviewed and approved by the Purdue Animal Care and Use Committee (PACUC). The approved protocol number is 1201000592.

Statistical analysis The larval activity was summarized as Burst Duration, the fraction of frames in each second of the video data that a larva moved(Liu et al., 2015). The larvae were first registered by the recoding software in the video frame as grey pixels. These pixels were compared between different frames. A larva was declared moving in a frame if their registered pixels moved more than a preset threshold. The activity for each larva (i.e. Burst Duration) was reported as the fraction of moving frames in each second. The normalization in this study was done by linear-regression model. We will first define the group and

explanatory variables in the model, and then describe the general framework of the model. Group and explanatory variables: Group variables were used to indicate different normalization conditions in the model, so that normalization can be conducted for each condition separately or for all conditions together. The group variables used in the normalization model include biological variations—Strains: AB, TL and TLAB; and Stage: 3–9 dpf. The group variables also include technical repeats—three consecutive periods of light onset (Light-On) and light offset (Light-Off).

The main explanatory variables are: (1) light intensity: measured from each well of the 96-well plate, and (2) biological replicates: two biological replicates were conducted for each experiment. Linear-regression model: The linear-regression model has the following general form:

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j + \epsilon_{ij} \tag{4.1}$$

where y_{ij} denotes the observed activity of the ith zebrafish larva in group j for $i = 1, \ldots, n_j$; x_{ij} denotes a column vector of explanatory variables for the corresponding larva; β_j represents a column vector containing the parameters of the linear-regression model for the group j, and ϵ_{ij} is the error term. The group j is coded to analyze corresponding specific subset of the data. For example, when $j =$ strain-AB & Light-On, the model used the observations from the AB strain during the Light-On period for normalization. Our model also assumed a simple linear relationship between the response and predictors. The statistical model was analysed using R software. The analysis computing scripts can be found at the GitHub repository https://github.com/zhanzmr/Normalization_Zebrafish.

We used principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) to evaluate the results of the data normalization, as described below. Principal Component Analysis (PCA) (Pearson, 1901) is a statistical multivariate analysis tool for dimensionality reduction and data visualisation. PCA takes the possibly correlated multivariate data matrix as input, uses an orthogonal transformation to produce a set of linearly independent output called principal components (PCs). This transformation projects the high-dimensional data in to a low-dimensional space composed of PCs. PCA defines a new orthogonal coordinate system that best describes the intrinsic variability of the data. The

variability contains the statistical information of the data set that we need to retain during the normalization procedure. Usually, the high-dimensional data can be visualised by the plotting the first two or three PCs, which usually capture much of the total variability of the data. In the PCA plot, the shape and relative locations of the data points represent the variability of the original multivariate data, and should not substantially change in a good normalization procedure.

In this study, we used PCA to analyse the multivariate VMR data before and after the integrated normalization. The data consist of the time-series activity profiles of individual larva from different stages from 2 seconds before light onset to 3 seconds after light onset. The multivariate VMR data (X) were orthogonally transformed by eigendecomposition, which aims to find an orthonormal matrix P where Y = PX such that is diagonalized. The principal components of X are the rows of P, or equivalently the eigenvectors of XXT. The PCA results were plotted in a 2D-PCA plot using the first two PCs. Each sample point on the plot represented the activity time profile of one individual larva and was also coloured according to its corresponding developmental stage. The PCA analysis was implemented using R software.

T-distributed stochastic neighbor embedding (t-SNE) is a dimensionality reduction and visualisation tool designed to aid the analysis of multivariate data (Van der Maaten & Hinton, 2008). It uses stochastic neighbor embedding, a nonlinear transformation, to reduce the dimension of the data. This method visualises the high-dimensional data by giving each sample point a location in a two-dimensional map, which can potentially reveal underlying relationship between data points as clusters.

We used the same data as in the PCA analysis for t-SNE analysis with parameter perplexity equals to 30. The data consist of the time-series activity profiles of individual larva from different stages from 2 seconds before light onset to 3 seconds after light onset. The main algorithm of t-SNE consists of the following steps: First, we constructed the probability distribution of pair-wise similarity between any pair of samples to define the neighbors for each sample. Similar samples had a higher probability to be picked, while dissimilar points had a lower probability to be picked. Second, in the low-dimensional map of t-SNE, we defined a similar probability distribution for the low-dimensional points similar to each other.

Finally, we iteratively improved the low-dimensional representation to minimize the Kullback–Leibler divergence between the two distributions so that they looked as closely alike as possible. The results were then plotted on 2D t-SNE map with each point representing one individual larva and coloured according to its corresponding developmental stage. The t-SNE analysis was implemented using R software with the R package "Rtsne".

4.2.3 Analysis results

In this study, we used the linear-regression model to conduct normalization of VMR data. We will first outline the approach for normalization of three different needs, and then illustrate how to integrate several normalizations needs together in an integrated analysis.

Example 1: Normalization of larval activities obtained from individual wells of a 96-well plate. In the VMR experiment, zebrafish larvae were arrayed individually in different wells of the 96-well plate. They were then subjected light stimulation emitted by the light-controlling unit with LED arrays. Since these LEDs were point light source, the larvae in different wells would receive slightly different light intensities, even though the emitted light was scattered by a diffuser. To illustrate the light variation, we measured the light intensities received in the wells of the 96-well plate when the light-intensity output of the machine was set at 100% (Fig 4.1). The wells in the center received higher light intensity than those in the corners. This difference in light intensities likely initiated the larvae to display a different level of VMR. Since this difference was not caused by biological difference, it must be removed by proper normalization for downstream analysis. To estimate the effect of light-intensity variation between different wells on VMR, we fit a linear-regression model (1) as follows:

$$\text{activity}_{ij} = \beta_{0j} + \beta_{1j}\,\text{light.intensity}_{ij} + \epsilon_{ij} \tag{4.2}$$

where activity is the observed VMR, i denotes the ith observation (i.e. larva), j denotes the group number (i.e. strain, stage, and technical repeats), and $light.intensity$ is the value of predictor variable

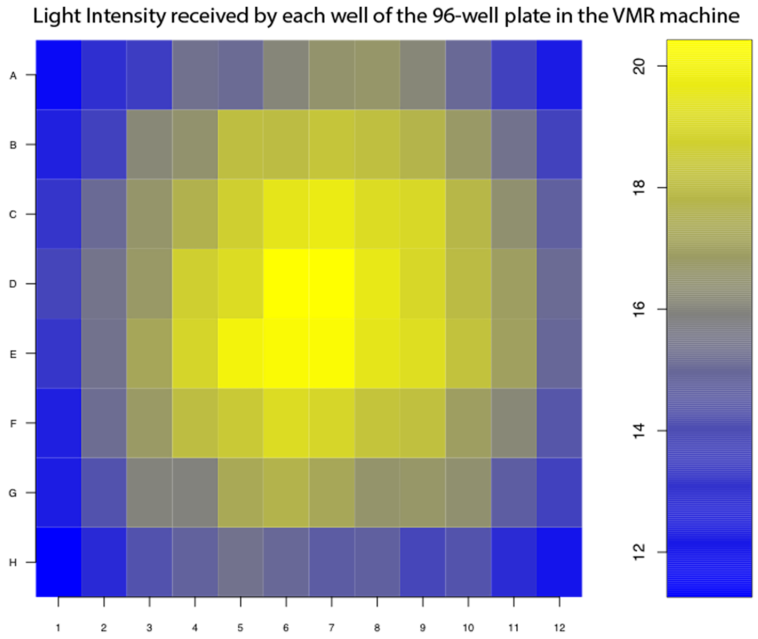

Figure 4.1: The heat map indicates the light intensity (in W/m2) received by each of the 96-well pate in the VMR machine, when the light-intensity output was set at 100%. The higher light intensities are represented by yellow colours, whereas the lower light intensities are represented by blue colours.

for light intensity. The parameters of the model are β_{0j} and β_{1j}. The parameter β_{0j} is the intercept of the regression line of the group j, which represents the mean response $E(activity)_j$ when light intensity is zero. The parameter β_{1j} is the slope of the regression line, which indicates the light intensity effect for the group j, i.e. the change in the mean activity $E(activity)_j$ per unit increase in light intensity for group j. The random error term for the ith observation and jth group is denoted as ϵ_{ij}, which is the deviation of the observed activity from the (unobservable) mean activity. This model estimated the effect of light-intensity variation between different wells in the 96-well plate on VMR for all different groups.

Then, we calculated the regression residual, the difference between the observed VMR ($activity_ij$) and the estimated activity caused by light-intensity variation across different wells (light-activity $_{ij}$). This residual represents the normalized activity of the ith larvae from jth group after removing the light-intensity effect from different wells. This subtraction would occasionally introduce negative activity values, which were corrected by adding an offset value μ_{offset} to keep all normalized activities positive: $\mu_{offset} \geq |$

min (light-activity -aj $_{ij}$) | for all i,j. Together, these calculations would yield light-normalized-activity (i.e. activity $_{ij}$ − light-activity $_{ij}$ + μ_{offset}), which could be used for downstream analysis.

In this example, we fit the linear-regression model **??** with the VMR data obtained from 6-dpf TL larvae from -30 s to 30 s after light onset (i.e. Light-On Stimulus Trials (yellow bars) in Fig 4.7). This group of data was denoted as $j = 1$ in the estimated model:

$$\text{light-activity}_{i1} = 1.907 \times 10^{-2} - 1.998 \times 10^{-4}\, \text{light.intensity}_{i1}, \tag{4.3}$$

where light-activity $_{i1}$ is the estimated activity for ith observation and group 1, the estimates of the regression coefficients are: $\widehat{\beta_{01}} = 1.907 \times 10^{-2}$ with standard errors $s\left(\widehat{\beta_{01}}\right) = 3.715 \times 10^{-4}$. $\widehat{\beta_{11}} = -1.998 \times 10^{-4}$ with standard errors $s\left(\widehat{\beta_{11}}\right) = 3.174 \times 10^{-5}$. Since the slope coefficient of the light intensity, $\widehat{\beta_{11}}$, was significantly different from zero (p-value = 3.12×10^{-10}), the variation of light intensity across wells positively influenced larval VMR. The fitted model effectively normalized and removed the effect of light-intensity variation on larval activities, which became more uniformly distributed across the 96-well plate (Fig 2).

Example 2: Normalization of batch effect. Many VMR experiments require the analysis of more than 96 larvae, exceeding the capacity of a 96-well plate that would be analysed in the same VMR machine per run. Consequently, these larvae were analysed sequentially on different 96-well plates in the same VMR machine, or in parallel in different VMR machines. These experimental schemes created batch variations on larval activity, which can be normalized by the following linear-regression model:

$$\text{activity}_{ij} = \beta_{0j} + \beta_{kj} I(\text{ batch } k) + \epsilon_{ij}, k = 1, \ldots, K \tag{4.4}$$

where $activity$ is the response, i is the ith observation (i.e. larva), j is the group number (i.e. strain, stage, and technical repeats), and $I(batchk)$ is the indicator function such that $I(batchk) = 1$ when the activity data come from batch k, otherwise $I(batchk) = 0$. We assume there are K levels of batch

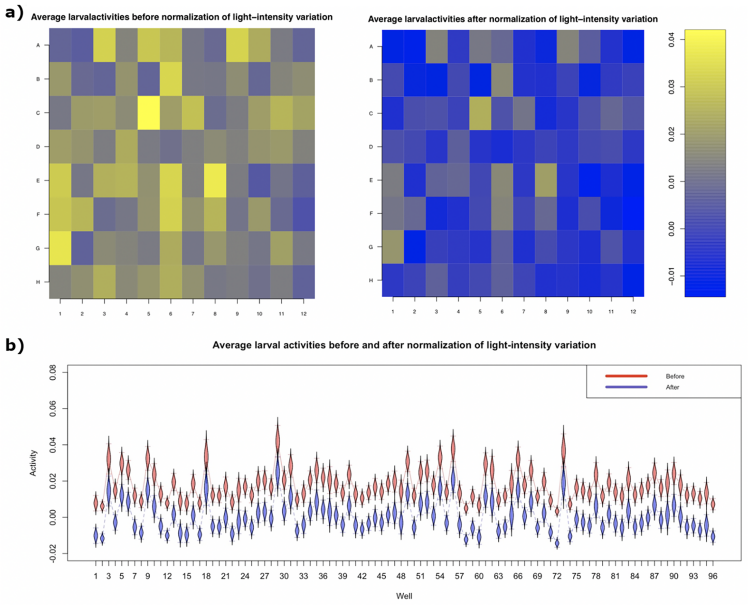

Figure 4.2: (a) Heatmaps showing average larval activities in each well of the 96-well plate before (left) and after (right) normalizing the light-intensity variation across the plate. These larval activities were extracted from 6-dpf TL larvae from 1 to 30s after light onset. (b) A boxplot of average larval activities before (red) and after (blue) light-intensity normalization.

effect to be removed. The parameters of the model are β_{0j} and β_{kj} for $k = 1, \ldots, K$, and ϵ_{ij} denotes the random error. The parameter β_{0j} is the grand mean, which represents the mean normalized activity, $E(activity)j$, for all observations from group j; the parameter β_{kj} is the batch effect for batch k, which is the deviation from the grand mean due to the batch effect of batch k for the group j.

To illustrate our approach for normalizing batch effect, we analysed our VMR dataset that contained two biological replicates conducted on different days in the same VMR machine. We used a subset of the VMR data obtained from 6-dpf TL larvae from -30 s to 30 s after light change, where we denoted as group $j = 1$. We modeled the two replicates as two batches: replicate1 and replicate2. They were treated as separate explanatory variables that took two possible values: 1 and 0 . "1" indicated the activity data belong to this replicate, whereas "0" indicated the activity data did not belong to this replicate. Therefore, replicate $1+$ repliacte2 $= 1$ for each pair of i and j. This subset of data was used to fit a model: batch-activity $_{i1} = 1.207\times10^{-2} - 1.586\times10^{-3}$ replicate $_{i1} + 1.586\times10^{-3}$ replicate 2_{i1}, where the batch-activity

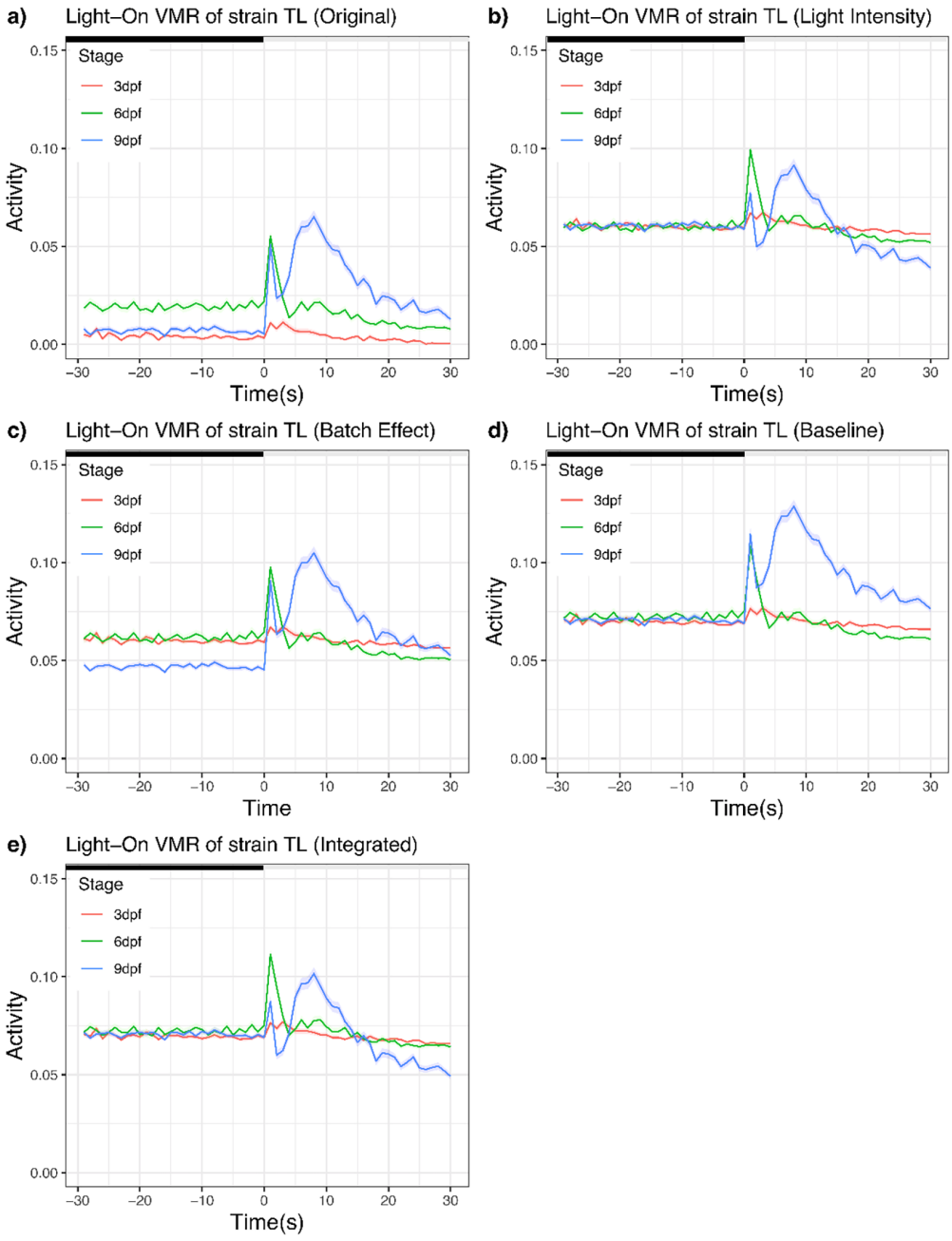

Figure 4.3: (a) Original data without any normalization. (b) Light-intensity normalization. (c) Batch-effect normalization. (d) Baseline normalization. (e) Integrated normalization. In all plots, the activities of larvae at different stages were plotted from 30 seconds before light onset to 30 seconds after light onset. The solid traces show the mean activities (red trace: 3dpf, green trace: 6dpf, and blue trace: 9dpf), whereas the ribbons surrounding these activity traces indicate the corresponding standard error of the pointwise mean activity. The offset value μ offset for (b) to (e) was 0.06.

y_{i1} is the estimated activity for ith observation in group 1. The estimates of the regression coefficients are: $\widehat{\beta_{01}} = 1.207 \times 10^{-2}$ with standard errors $s\left(\widehat{\beta_{01}}\right) = 2.885 \times 10^{-4}$; $\widehat{\beta_{11}} = -1.586s \times 10^{-3}$ with standard errors $s\left(\widehat{\beta_{11}}\right) = 4.090 \times 10^{-4}$. Since we implemented the zero-sum constraint on β_{11} and β_{21}, i.e. $\beta_{11} + \beta_{21} = 0$, we have $\widehat{\beta_{21}} = 1.586 \times 10^{-3}$. The slope coefficient of the light intensity $\widehat{\beta_{11}}$ was significantly different from zero (p-value $= 1.05 \times 10^{-4}$), which indicates that the activities from two replicates are significantly different.

Then, we calculated a regression residual to remove the batch effect from biological replicates. batch-batch variation (batch-activity $_{ij}$). This residual represents the normalized activity of the ith larvae from jth group after removing the batch-batch effect. This subtraction would occasionally introduce negative activity values, which were again corrected by adding an offset value μ_{offset} to make all normalized activities positive: $\mu_{offset} \geq |\min(\text{ batch-activity } ij)|$ for all i, j. These calculations would yield batch-normalized-activity (ije. activity ij $_{ij}-$ batch-activity $_{ij} + \mu_{offiet}$), which could be used for downstream analysis.

To illustrate the effect of batch normalization on the VMR profiles, we again plotted the same Light-On dataset for TL strain (Fig 4.3c). Compared with the unnormalized data (Fig 4.3a), the batch-normalized activities now share the same mean activity across time. In other words, if we summarize each curve in Fig 4.3 into its corresponding mean value, they will have the same mean value after batch normalization.

Example 3: Normalization to a common baseline We previously designed a Hotelling's T-squared test (Liu et al., 2015) to compare VMR between two samples in a specific time frame. One of the most important comparisons was the time frame after light change, as this could reveal the difference in light sensation between groups. This statistical comparison allowed us to evaluate not only visual impairment in fish mutants, but also drug improvement of their impaired vision (Ganzen et al., 2017; Liu et al., 2015). However, the success of this comparison relied on an implicit assumption: the two samples displayed comparable activities before light change. In reality, different samples often displayed varying baseline activities (Fig 4.3a). This baseline variation must be normalized for an effective comparison of two samples by the Hotelling's T-squared test. The baseline can be the grand mean activity across all conditions from

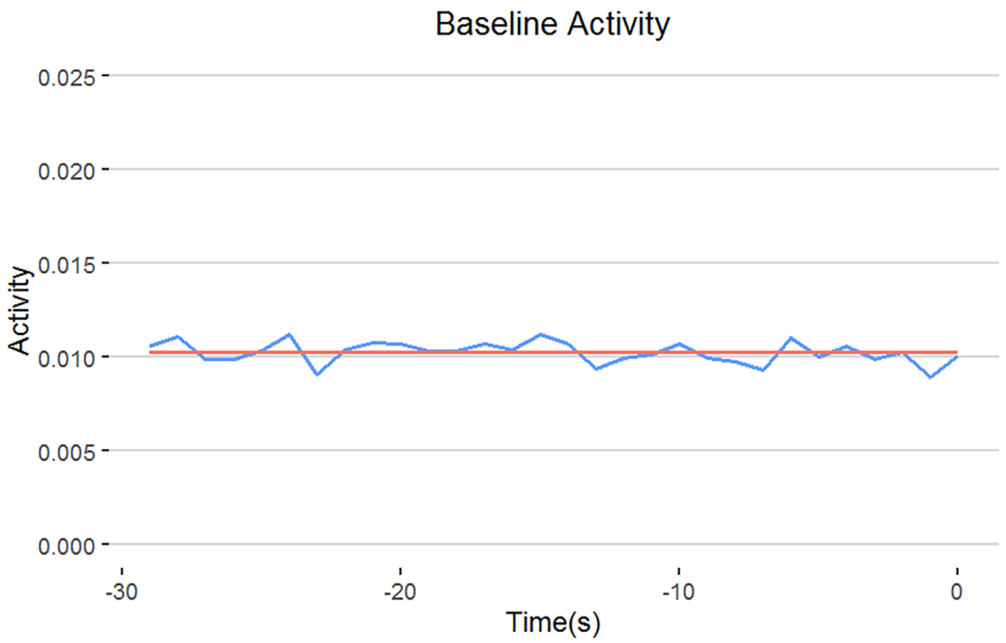

Figure 4.4: In this study, we proposed to use the average activity of last 30 seconds from the 3.5-hour adaptation period (i.e. regions under red bar in S1 Fig) as baseline for normalization. The blue line indicates the mean activity for each second, whereas the red line indicates the grand mean of all activities in the whole 30-second period. Since the two lines are highly comparable, this suggests the grand mean of the activities is very stable and can be used for baseline normalization.

a specific time period immediately before light change, for example the last 30 seconds from the 3.5-hour adaptation period before the light change (i.e. regions under the first red bar in Fig 4.7), because the larvae should be acclimatized and would be more stable after several hours of adaptation. In fact, the grand-mean activity 30 seconds before the light change was 0.01024 across all strains and stages (Fig 4.4, red line). It was around the average activities per individual second during the same 30-second period (Fig 4.4, blue line), which were stable. Hence, the grand-mean activity could be used for baseline normalization.

The VMR of different groups were then normalized by adjusting the averaged activities of each group to 0.01024. This was achieved by the following steps: First, we obtain a baseline normalization factor by fitting a linear-regression model with only intercept term:

$$activity_{ij} = \beta_{BaselineNormFactor_j} + \epsilon_{ij} \tag{4.5}$$

where activity ij is the ith observation in the jth group. The parameter of the model, $\beta_{BaselineNormFactor_j}$, can be estimated as $\hat{\beta}_{BaselineNormFactor_j} = ave\,(darkActivity_j) - 0.01024$, where ave($darkActivity_j$) denotes the average activity from the 30-second time period before the light change for group j. Then, we calculated a regression residual, the difference between the observed VMR ($activity_i j$) and the baseline normalization factor $\hat{\beta}_{BaselineNormFactor_j}$. Since the calculation might yield negative values for activities, we again corrected that by adding an offset value μ offset to make all baseline normalized activities positive: $\mu_{\text{offset}} \geq \left|\min\left(\hat{\beta}_{BaselineNormFactor_j}\right)\right|$ for all j. Together, these calculations would yield $baselineNormalizedActivity)ij\left(i.e \quad activity_i j - \hat{\beta}_{baselineNormFactor_j} + \mu_{offset}\right)$. After this baseline normalization, all groups will have the same group average, 0.01024 + μ offset, as the baseline (Fig 4.3d). These baseline-normalized activities could be used to perform the Hotelling's T-squared test.

Integrated normalization of VMR data In the previous sections, we demonstrated how to normalize different variables of the VMR experiments by linear-regression models. In practice, these variables should be normalized all at once. The resulting residuals from the model would be free from systematic variations and can be used to reveal true biological difference between different samples. To illustrate the value of our normalization approach, we will normalize all three variables outlined in the earlier examples using the same Light-On VMR data of TL strain at 3, 6, and 9 dpf again. The integrated normalization had three steps: First, it normalized light-intensity variations in the 96-well plate; Second, it used the residuals from step 1 as the response variable to normalize the batch effect; Third, it used the residuals from step 2 to perform a baseline normalization. In this step, an offset value μ offset = 0.06 was applied. The result of integrated normalization is shown in Fig 4.3e. The normalized data were then used for statistical comparisons by the Hotelling's T-squared test (Liu et al., 2015) (Fig 4.5). In this example, we analysed three seconds around the light change to highlight the effect of integrated normalization.

Before integrated normalization (Fig 4.3a), the activity of 6-dpf larvae before light onset was significantly different from that of the 3-dpf and 9-dpf larvae (Fig 4.5, $p < 0.0001$), whereas the activities of 3-dpf and 9 dpf larvae were comparable (Fig 4.5, p = 0.121). After light onset, the 3-dpf larvae did not display much activities and was significantly different from the 6-dpf larvae and 9-dpf larvae (Fig 4.5, $p < 0.0001$).

Before integrated normalization		
Comparison	**Test statistic (p-value)**	
Stage (dpf)	**Before light onset (-2–0 s)**	**After light onset (1–3 s)**
3 vs. 6	115 (0.0000)	208 (0.000e+00)
3 vs. 9	5.83 (0.121)	535.7 (0.000e+00)
6 vs. 9	82.8 (0.0000)	13.6 (3.671e-3)
After integrated normalization		
Comparison	**Test statistic (p-value)**	
Stage (dpf)	**Before light onset (-2–0 s)**	**After light onset (1–3 s)**
3 vs. 6	13.2 (0.0132)	117.66 (0.000e+00)
3 vs. 9	0.188 (0.9795)	66.04 (8.071e-14)
6 vs. 9	11.4 (0.0152)	94.02 (0.000e+00)

Figure 4.5: The top table contains the test results before integrated normalization, whereas the bottom table contains the test results after integrated normalization. The corresponding activity plots can be found in Fig 4.3a and 4.3e respectively. In both tables, we presented the comparisons of VMR three seconds before light onset and three seconds after light onset.

The 6-dpf and 9-dpf larvae, however, displayed a strong Light-On VMR in the first three seconds after light onset that were relatively comparable to each other (Fig 4.3a; Fig 4.5, $p < 3.671e\text{-}3$). The situation was quite different after the integrated normalization (Fig 4.3e). The normalization brought the activities before light onset to a more comparable level and changed the shape of the activity profiles after light onset. In particular, the peak activity of 6-dpf larvae was now substantially higher than that of the 9-dpf larvae (Fig 4.5, $p < 0.0001$).

Evaluation of model-based normalization for VMR data Any effective normalization approach should demonstrate two properties that would make the normalized data reveal the underlying information better than the original data. First, the normalization approach should not change the intrinsic variability of the data. Data variability is the extent to which sample points vary in a data distribution. The change of data variability is an indicator of whether the normalized data have been distorted or not. Our normalization procedure should maintain data variability since linear-regression modelling focus on the mean of the data. Second, the normalization approach should help find a clear and concrete grouping pattern for data from different classifications. These two properties were integral components of

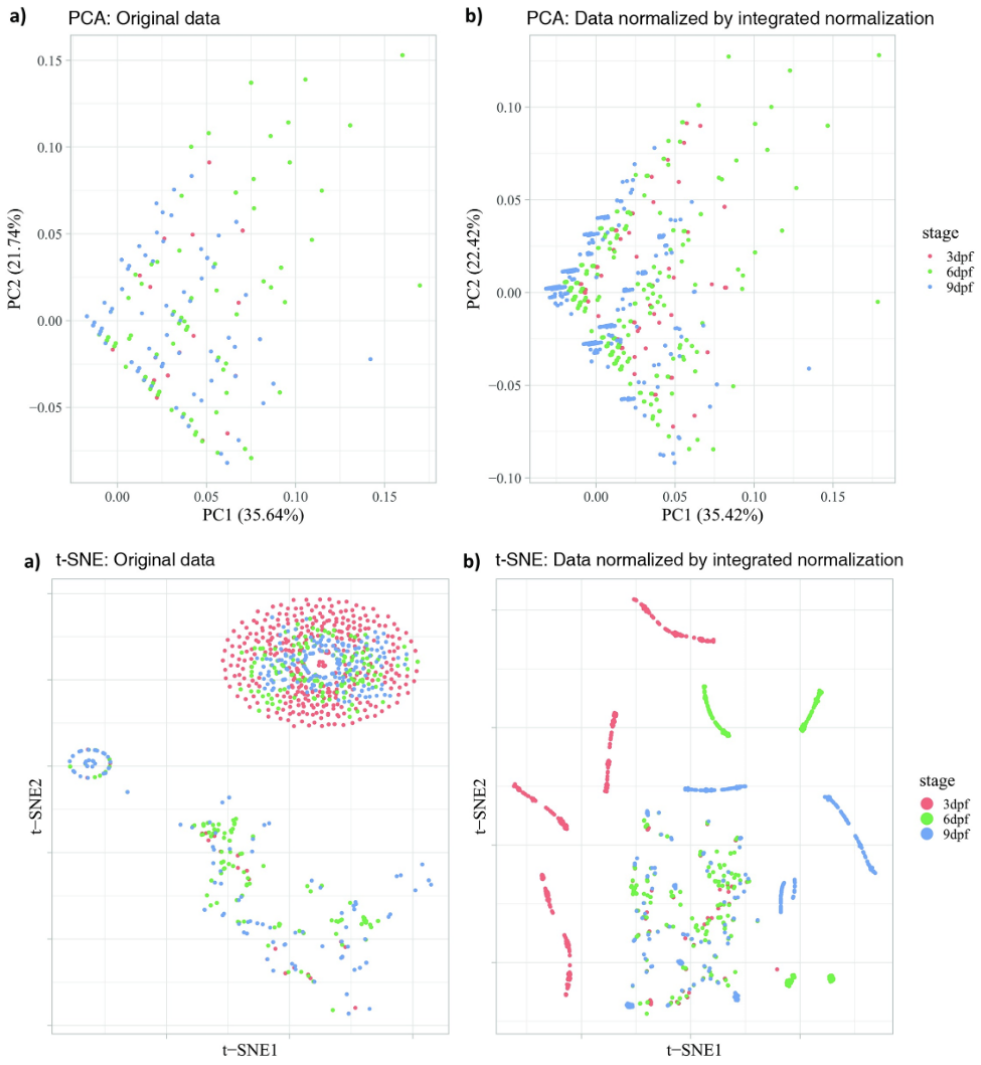

Figure 4.6: In this study, we proposed to use the average activity of last 30 seconds from the 3.5-hour adaptation period (i.e. regions under red bar in S1 Fig) as baseline for normalization. The blue line indicates the mean activity for each second, whereas the red line indicates the grand mean of all activities in the whole 30-second period. Since the two lines are highly comparable, this suggests the grand mean of the activities is very stable and can be used for baseline normalization.

our model-based normalization for VMR data, as illustrated by PCA (Fig 4.6 top) and t-SNE (Fig 4.6 bottom).

We visualised the VMR data before and after normalization by PCA. This method transforms the multidimensional data into fewer orthogonal dimensions called principal components (PCs) that are uncorrelated with each other. Each PC captures the largest possible variance compared to the next one. In Fig 4.6 top, we plotted the first two PCs that captured more than 55% of the data variance. The plots show that i) the normalized dataset has a similar triangular shape compared to the raw data; ii) the relative location of the individual data points are similar; and iii) the variance explained by PC1 and PC2 are similar before and after normalization (Fig 4.6 top a: 35.64% and 21.74% vs. Fig 5b: 35.42% and 22.42%). These together suggest that the intrinsic variability of the data was maintained by our normalization method.

To reveal the clustering of larva from different stages, we further visualised the VMR dataset before and after normalization by t-SNE. This method transforms the multidimensional dataset into low dimensional space by converting the distances in multidimensional space between sample points into probabilities that represent their similarities. In Fig 4.6 bottom, we plotted the 2D t-SNE map. Before normalization, the data points from different developmental stages were either scattered randomly on the plot or were aggregated together (Fig 4.6 bottom a). After normalization, data points from different developmental stages were clearly clustered together and separated from other stages (Fig 4.6 bottom b). There were some data points from different stages aggregated in the middle bottom of the figure, probably reflecting the larvae involved displayed similar behavioural pattern, for example, moving little or not at all during the experimental period. The t-SNE map in Fig 4.6 bottom b shows a clearer clustering of larvae from the same stages that display similar behavioural patterns. This clear clustering of data from similar stages indicate that our normalization approach can reveal patterns in the data closer to the biological nature.

4.2.4 Conclusion on zebrafish behavior data

High-throughput approaches for collecting behavioural data have revolutionized neuroscience research, when the collected data are properly analysed. These data are often multi-dimensional, as they are con-

tinually and repeatedly collected from multiple individuals under different kinds of perturbations. One such experimental approach is called VMR. This assay collects swimming responses from many zebrafish larvae arranged in 96-well plates over time, which make the data correlated in time and by location. If the data are collected in very short time frame in seconds, some larvae may not move. The resulting data will then contain many zero values, which creates a data-imbalance problem. These features of the VMR data cannot be dealt with by traditional analyses including the t-test and AVONA. In previous studies, we addressed the time-dependency issue by the Hotelling's T-squared test (Liu et al., 2015), and the data-imbalance problem and location-correlation issue by the GLMM (Liu et al., 2017). These new analyses enable proper statistical analysis of VMR data for the first time. Nonetheless, these statistical analyses did not address another fundamental issue of these high-throughput behavioural data: the experiments are often subjected to systematic variations. If these variations are not accounted for, they would affect the performance of the aforementioned statistical analyses. To address this analytical gap, we established an approach to normalize the systematic errors by linear-regression modeling.

Our normalization approach modeled the relationship between larval activities (response) and uncontrolled systematic variations (explanatory predictors). The resulting regression residuals were then used as the normalized activities. This approach was flexible because it could easily handle different types of uncontrolled experimental conditions by adding separate terms in the normalization model. For example, it handled continuous variables such as light intensity (Figs 4.2 and 4.3b) and baseline activities (Fig 4.3d), and categorical variables such as biological replicates (Fig 4.3c) These variables can also be normalized in one integrated model to remove the effect of multiple systematic errors at once (Fig 4.3e). The linear-regression model can also be adapted to different sample groups (genotypes strains, and/or stages), and enabled normalization of selected subset of data.

By removing systematic biases, new patterns can be revealed from the normalized data. For example, the integrated normalization (Fig 4.3e) removed the difference in activities due to variation in light intensity between different wells of the 96-well plate. This has changed the activity profile of the individual stages. In addition, the normalization also brought the activities before light onset to a comparable level, essentially

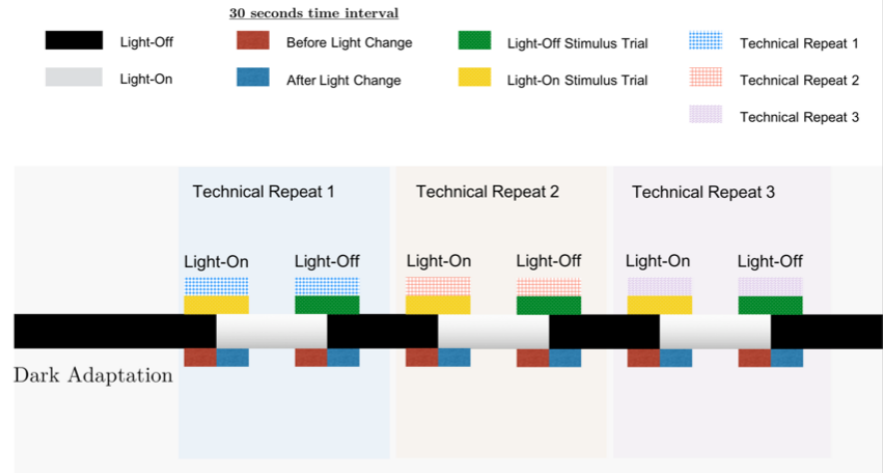

Figure 4.7: Supplementary Figure VMR experimental scheme. This scheme was used to collect the dataset used in this analysis. In the scheme, the larvae were first dark adapted for 3.5 hrs (long black bar on the left). Then, they were subjected to three consecutive trials of light onset (grey bars) and light offset (short black bars). Each light-on or light-off session lasted for 30 mins. Three technical repeats were also performed in each biological replicate; two biological replicates were performed for each condition. In this study, we extracted the data from 30 s before light change (red bars; not to scale) to 30 s after light change (blue bars; not to scale) for statistical analyses. In some cases, we further restricted the analysis to from 3 s before light change to 3 s after light change.

assuming that was the baseline activities for different groups of larvae. This assumption may not be applicable to all cases, but it can be used to assess the extent of relative level of larval response upon light onset. In our case, this normalization clearly shows that the 6-dpf TL larvae responded to light simulation much stronger than the 9-dpf TL larvae, a conclusion that can only be drawn after appropriate normalization. The new patterns revealed from the normalized data likely reflect the underlying biological pattern clearer, as the model-based normalization did not alter the data structure and could cluster data in the same categories better (Fig 4.6).

Our model-based normalization had several limitations. First, it only handled continuous responses of larval movement. This limitation can be resolved by using generalized linear model to deal with categorical response variable. Second, the linear-regression model mainly focuses on the linear relationship between the response and the explanatory variables. This can be partly resolved by adding higher-order terms of

explanatory variables. The model can also be generalized through nonparametric regression techniques, in which no assumption is made on the relationship between response mean and predictors to any specific class, including linear or quadratic class. This approach can be useful to model the effect of light intensity as a function of visual sensitivity which operates over several log units. Third, it does not consider the temporal dependency during the normalization. This can be resolved by adding time-series terms to the linear-regression model or generalizing it to time-series regression model.

To conclude, our study has implemented the linear-regression model to normalize VMR data. The normalized data can then be used in downstream analyses including the Hotelling's T-squared test(Liu et al., 2015) and the GLMM for statistical comparisons between sample groups. This model-based normalization can be integrated into our framework for VMR data analysis in the following workflow: (1) Normalization using linear-regression model; (2) Comparing the larval activities of different groups using Hotelling's T-squared test; (3) Using GLMM to model the relationship between responses and candidate predictors; and (4) Combining the results from (2 & 3) to interpret larval activities. This framework facilitates the dissection of the underlying circuitry that drives VMR, and in turn the identification of true biological factors that affect the behaviour. We also expect our normalization and analysis framework applies to other high-throughput behavioural data with a similar structure, and can unveil new insights into neurobehaviour.

4.3 Graph neural network and non-parametric regression analysis on human dynamics dataset

4.3.1 Introduction to human dynamic

Human dynamics aim to understand human behaviors using analytical models. It has received substantial attention in the security and defense area not only for its potential in detecting human anomalies but also for its capability in containing potential disastrous damages and mental horror in the human society. The

recent development in wearable devices has revolutionized daily life and inaugurated a new era in human dynamics study. Advocated by Barabási who demonstrates that certain human behavior can be modeled quantitatively using proxy tools (Barabasi, 2005). A large amount of human dynamics papers have been published with a wide variety of proxies. Wearable device data is a major source of proxies that can be used to understand spatial-temporal trends from which we can identify abnormal patterns in human dynamics. The abnormal patterns can be used as an indicator for disasters.

4.3.2 Data collection

Data collected through the wearable devices contain continuously precise GPS and physical activities. Such data can be used to study the dynamic patterns of human behavior. Employees working full-time at city or county governments in the state of Georgia were recruited for the study. Participants were asked to wear GPS devices and Physical Activity Monitors at the same time for up to two weeks to capture geo-location data aligned with their physical activities. The devices are shown in Figure 4.8(a). GPS data were sampled at 30-seconds epochs and merged with the accelerometer data using the Personal Activity and Location Measurement System using the default settings. Data contain many segments and resting notes, many steps per segment, longitude, and latitude, and activity intensity based on slope and elevation change per segment. The dataset has a total of 46 variables including spatial coordinate (Latitude, Longitude), activities estimates, heart-rate, estimated type of trip, and light intensity. There was a total of 78 participants with both sufficient Physical Activity Monitors and GPS devices data available (Gay et al., 2018; Gay et al., 2017). Figure 4.8(b) illustrates some trajectories of these wearable device data.

4.3.3 Graph classification using convolutional network (GCN)

Our goal is to learn the dynamic patterns in the two groups of candidates. The two groups are office based workers group and non-office based works group. In this section, We split the dataset into training and testing. The training dataset contains 44 candidates and the testing dataset contains 10 candidates. The first step is to transform the candidate's spatial trajectories into different networks. Then, we build a graph

Table 4.1: Represent variables description

Variable Name	Description	Type	Value
dateTime	Date and Time of day in 24 hour format	String	yyyy-mm-dd hh:mm:ss
lat	Latitude of participant's location at this time	Double	Degrees -180.0, -180.0 if unknown
lon	Longitude of participant's location at this time	Double	Degrees -180.0, -180.0 if unknown
duration	Duration of epoch (in seconds)	Integer	30
distance	Distance traveled during the epoch (in meters)	Integer	0 - no max
speed	Speed of travel during the epoch (in km/hour)	Double	0.0 - 9999.0
tripMode	Estimated Mode of Transportation when moving within a trip	String	Pedestrian Bicycle Vehicle Others -1 if unkown (no data) -2 when marked as "non wearing"
vectorMag	Vector magnitude during the interval as calculated by the device	Double	0 - 9999 -1 - unknown (no data)
lux	Value reported by the ambient light sensor	Integer	0 - 130,000 (for ActiGraph devices)

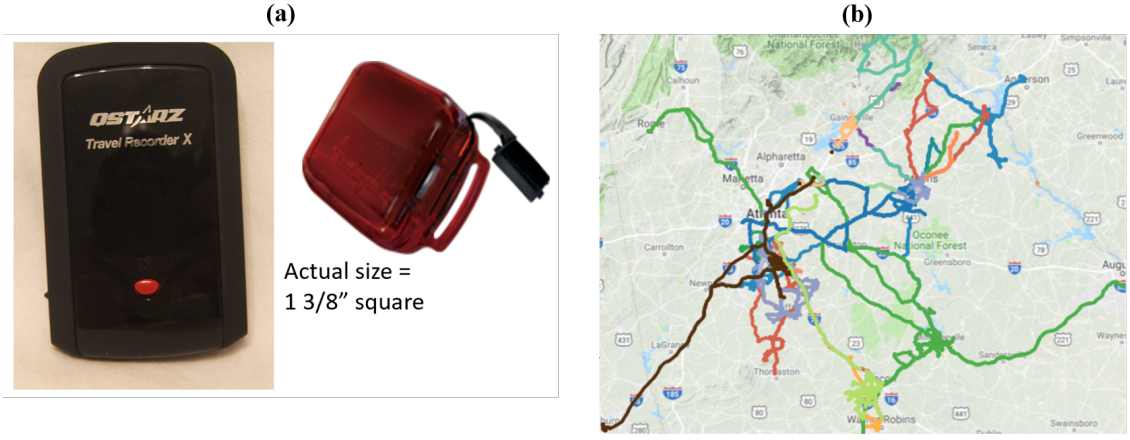

Figure 4.8: (a) The QStarz BT-1000XT Device. (b) Illustration of sample trajectories in Atlanta-Athens area in the personal wearable device data. Different color indicate different person's trajectories.

neural network model to classify each group of candidates and evaluate the result based on the testing dataset.

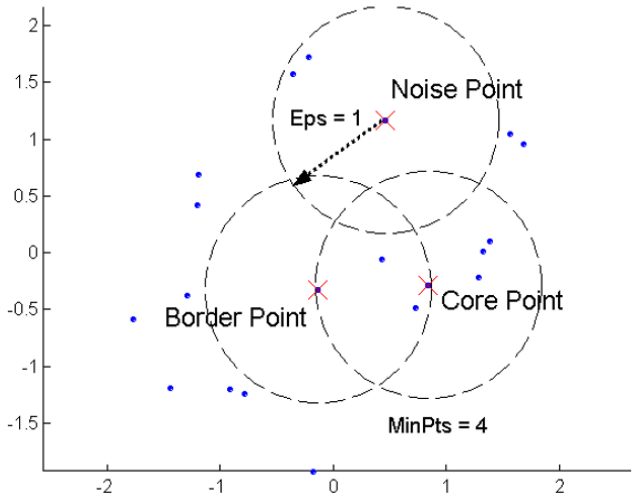

Figure 4.9: The Density-based spatial clustering of applications with noise (DBSCAN) algorithms.

DBSCAN clustering on spatial points The best way to represent human dynamic patterns is to build trajectory networks for candidates (Sun et al., 2019). The density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in data mining and machine learning (Figure 4.9). The density at the core point is defined as the number of points within a circle of Radius Eps (ϵ) from the core point. The dense region defined as the circle with radius ϵ contains at least a minimum number of points (MinPts) for each point in the cluster.

The main concept of the DBSCAN algorithm is to locate regions of high density that are separated from one another by regions of low density. The clusters of each candidate indicate the frequently visiting place such as home workplaces, favorite grocery stores, or restaurants during this period. We collect the output clusters of each candidate and set it as the node in our trajectory networks. The nodes also contain information about the candidate's visiting time and physical activities in the nodes.

Figure 4.10 shows the output of spatial clusters for an office based worker and a non-office based worker. Figure 4.10 a) is an example of office based worker geo-location trajectories. The worker travel

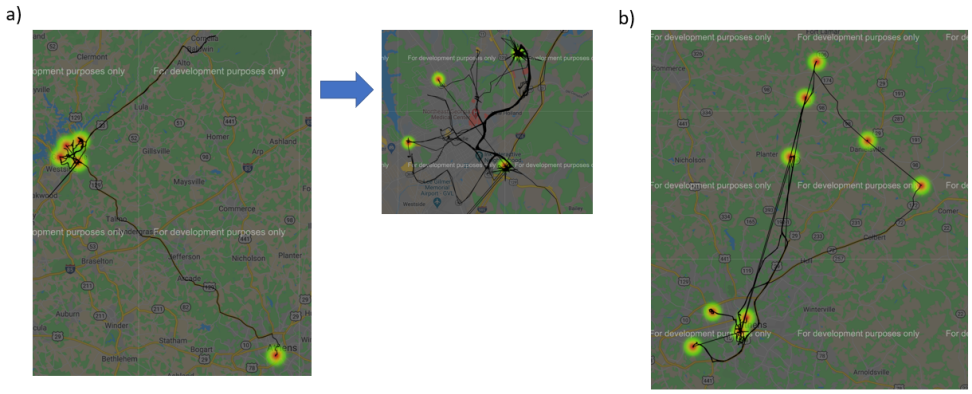

Figure 4.10: Trajectories network for candidates: **a)** An example of office based worker geo-location trajectories. The right figure enlarge the area in left figure where all clusters are gather together. **b)** An example of non-office based candidate's geo-location trajectories.

frequently from one cluster to another. Those places might indicate the worker's home and workplace. As a result, one candidate can be represented by one network graph and some examples of the candidate's network are showing in Figure 4.11.

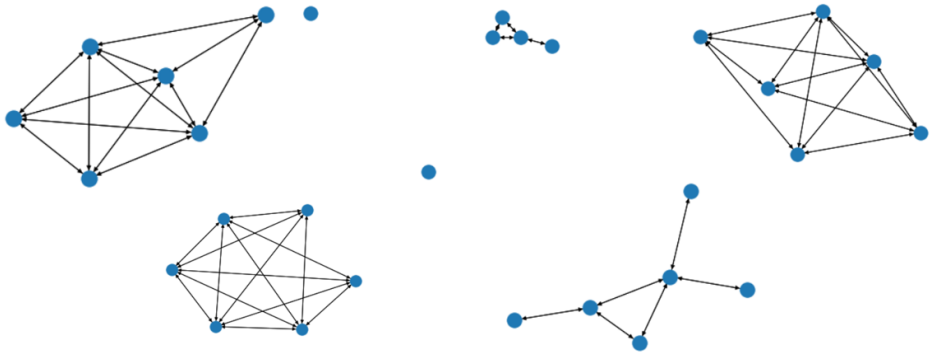

Figure 4.11: Trajectories network for candidates: Examples of both office based and non-office based worker's trajectories network.

Graph classification using GCN Graph Convolutional Network (GCN) is a very powerful neural network architecture for machine learning on graphs. Graph classification (Kipf & Welling, 2016; Schlichtkrull et al., 2018; Ying et al., 2018) is also an important problem with applications across many fields, such as

bioinformatics, chemoinformatics, social network analysis, urban computing, and cybersecurity. Applying graph neural networks to this problem have been a popular approach recently (Duvenaud et al., 2015). For this dataset, we propose a two-layer graph convolutional network (GCN) framework for graph classification of all candidate's trajectories network. The goal of this GCN model is to learn a function of signals/features on different graphs. We take the following as inputs:

1. A feature description x_i for every node i; summarized in a $N \times D$ feature matrix X (N: number of nodes, D: number of input features)

2. A representative description of the graph structures in matrix form; in the form of an adjacency matrix A. Since we input many graphs together, we put graphs together to form a large diagonal adjacency matrix, which we use here to indicate the graph structures.

The model will produce graph-level output Z (an $N \times F$ feature matrix, where F is the number of output features per graph. This is different than node-level output where F is the number of output features per node. We add a pooling operation before outputting the results. For each convolutional layer, it can be written as a non-linear function:

$$H^{(l+1)} = f\left(H^{(l)}, A\right) \tag{4.6}$$

with $H(0) = X$ and $H(L) = Z$, L indicates the number of layers. We estimate the output feature matrix $H^{(l+1)}$ using Layer-wise propagation rule:

$$f\left(H^{(l)}, A\right) = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{4.7}$$

with $\hat{A} = A + I$, where I is the identity matrix and $\hat{D}$ is the diagonal node degree matrix of $\hat{A}$. In order to train and learn the graph neural network for these human dynamic networks efficiently, we also need to batch multiple graphs together to form a mini-batch in each training epoch. A batch of graphs can be viewed as a large graph that has many disjointed connected components. Figure 4.13 shows an overview

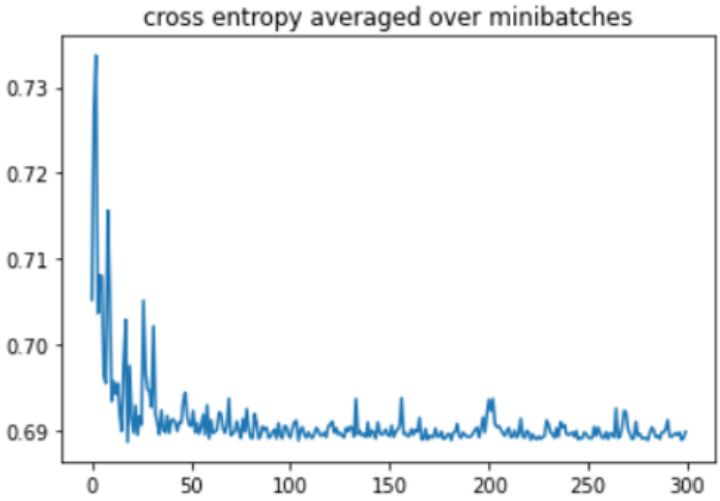

Figure 4.12: The Loss of each epoch in the training process.

of the model. The two-layer graph convolutional network (GCN) framework is showing in figure 4.13 a), this framework goes through two identical graph convolutional layers with ReLu as the activation function. Figure 4.13 b) shows the input is a batched list of graphs. Then the model goes through graph convolutional layers. Before the final classification layer, we add a graph readout to average over all node features for each graph in the batch. Then, the soft classification layer will output the results for each trajectories graph, not for each node.

Our results perform pretty well as we achieve $> 85\%$ on the testing accuracy. Figure 4.12 shows the measure of loss for each epoch in the model training process. The loss drops significantly in the first few training epochs.

4.3.4 Analysis using smoothing spline ANOVA model

Differences between occupational worker and non-occupational workers

One common hypothesis on human behavior is that the average activity level during working hours for occupational workers is higher than non-occupational workers. By analyzing the wearable data, we aim to

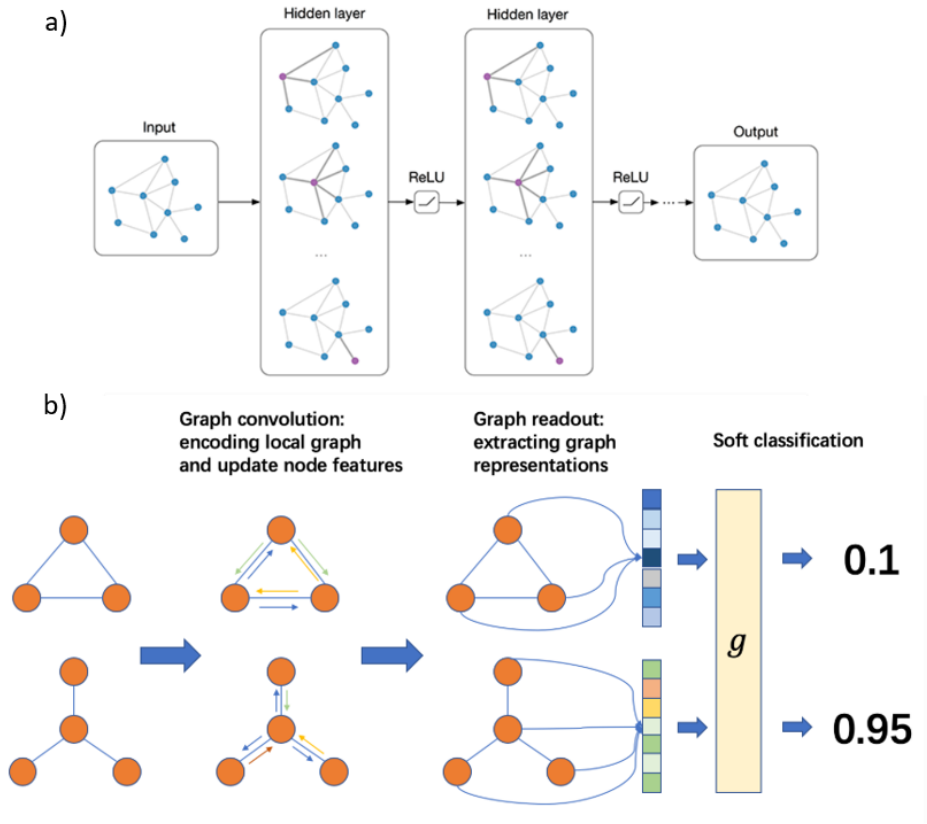

Figure 4.13: a) Typical model for graph convelution netwotk model. **b)** The workflow of our GCN graph classification model.

quantify the difference. One challenge when dealing with the data is denoising. Because of the sensitivity of the wearable devices, almost every movement during the experiment will be recorded, even a very slight one. Such sensitivity results in a extremely noisy data. We can hardly extract any difference between the occupational workers and non-occupational workers from the original data. Figure 4.14 shows the activity levels of two participants during working hours, i.e. 9 AM - 5 PM on weekdays, one is occupational and the other is non-occupational. Notice that the data is a time series, and there is a strong periodicity, thus we first consider transforming the data onto a frequency domain.

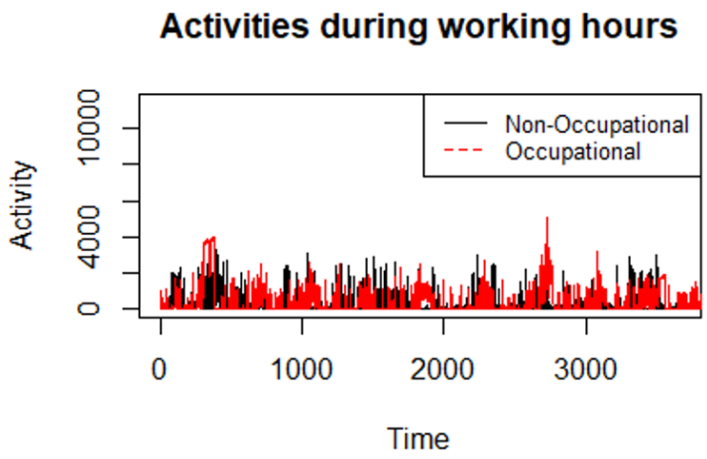

Figure 4.14: Activity levels of two workers during working hours (9 AM - 5 PM), the red line is for occupational worker and the black one is for non-occupational worker.

Frequency domain A time-domain graph shows how a signal changes over time. Frequency-domain graph shows how much of the signal lies within each given frequency band over a range of frequencies. We use the following Fourier Transformation formula in order to transform activities onto the Frequency Domain.

$$\tilde{x}_\nu = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t e^{-\mathrm{i}2\pi t\nu/T}, \quad \nu = 0, 1, \ldots, T-1$$

where is is the activities estimates and t is the time in the time period we choose previously. After the transformation, we can then denoise the data using certain smoothing method.

Fitting smoothing spline ANOVA models Smoothing spline ANOVA models are a versatile family of smoothing methods that are suitable for both uni-variate and multivariate problems (Gu, 2013).

$$y_i = \eta(x_i) + \varepsilon_i, i = 1, \ldots, n \tag{4.8}$$

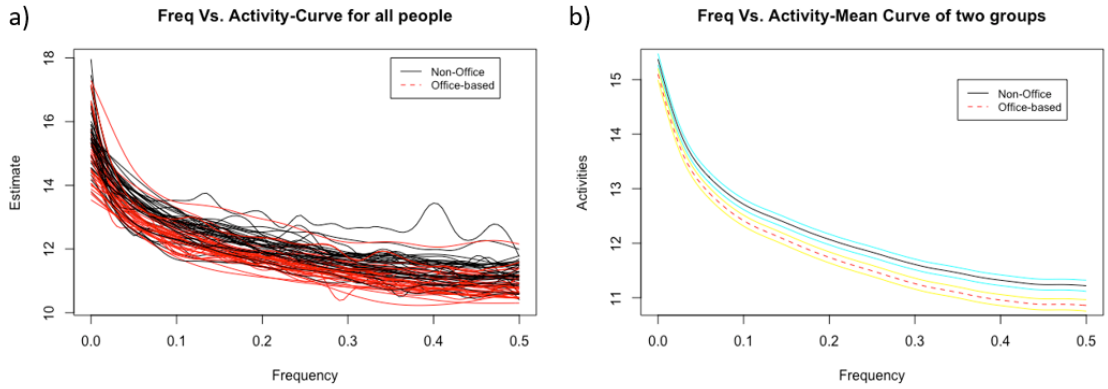

Figure 4.15: a) here total 74 curves in this graph. The black curves represent the participants whose work is non-office based. The red curves represent the participants whose work is office based. **b)** The two mean curves represent two groups of candidates using SSANOVA model.

We estimate $\eta(x)$ via penalized least squares:

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \eta\left(X_i\right)\right)^2 + \lambda\int_0^1 \eta''(X)^2 dx$$

For each of the participant, The first step is to estimate the frequency spectrum of the participant's physical activities by using Fourier Transformation. Because the spectral density is an even function, so we only needs to estimate the frequency from (0, 0.5). After Fourier Transformation, a cubic spline model can now be fitted to the log periodogram via gamma regression. We define $X = (x_1, x_2, ..., x_{101})^T$ to be a sequence contain 101 numbers uniformly from 0 to 0.5. Each X_i represents each level of frequency from 0 to 0.5. We predict estimates activities for each x_i via the cubic spline model. Figure 4.15 (a) shows that the distribution curves for each participant after predict classified by non-office and office based worker in frequency spectrum.

We can clearly see the physical activities difference between office based and non-office based works. Figure 4.15 (b) has only two curves represent two categories in all the participants. This graph shows that if the frequency level is fixed, non-office workers always have higher activities than office based workers.

Oppositely, if the activity is fixed, non-office workers tend to achieve this level more frequently than office-based workers.

References

Alexandrov, V., Brunner, D., Hanania, T. & Leahy, E. (2015). High-throughput analysis of behavior for drug discovery. *European journal of pharmacology*, *750*, 82–89.

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*(7039), 207.

Branson, K., Robie, A. A., Bender, J., Perona, P. & Dickinson, M. H. (2009). High-throughput ethomics in large groups of drosophila. *Nature methods*, *6*(6), 451–457.

Bruni, G., Lakhani, P. & Kokel, D. (2014). Discovering novel neuroactive drugs through high-throughput behavior-based chemical screening in the zebrafish. *Frontiers in pharmacology*, *5*, 153.

de Esch, C., van der Linde, H., Slieker, R., Willemsen, R., Wolterbeek, A., Woutersen, R. & De Groot, D. (2012). Locomotor activity assay in zebrafish larvae: Influence of age, strain and ethanol. *Neurotoxicology and teratology*, *34*(4), 425–433.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A. & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 2224–2232.

Emran, F., Rihel, J., Adolph, A. R., Wong, K. Y., Kraves, S. & Dowling, J. E. (2007). Off ganglion cells cannot drive the optokinetic reflex in zebrafish. *Proceedings of the National Academy of Sciences*, *104*(48), 19126–19131.

Emran, F., Rihel, J. & Dowling, J. E. (2008). A behavioral assay to measure responsiveness of zebrafish to changes in light intensities. *JoVE (Journal of Visualized Experiments)*, (20), e923.

Fernandes, A. M., Fero, K., Arrenberg, A. B., Bergeron, S. A., Driever, W. & Burgess, H. A. (2012). Deep brain photoreceptors control light-seeking behavior in zebrafish larvae. *Current Biology*, *22*(21), 2042–2047.

Ganzen, L., Venkatraman, P., Pang, C. P., Leung, Y. F. & Zhang, M. (2017). Utilizing zebrafish visual behaviors in drug screening for retinal degeneration. *International Journal of Molecular Sciences*, *18*(6), 1185.

Gao, Y., Chan, R. H., Chow, T. W., Zhang, L., Bonilla, S., Pang, C.-P., Zhang, M. & Leung, Y. F. (2014). A high-throughput zebrafish screening method for visual mutants by light-induced locomotor response. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*(4), 693–701.

Gao, Y., Zhang, G., Jelfs, B., Carmer, R., Venkatraman, P., Ghadami, M., Brown, S. A., Pang, C. P., Leung, Y. F., Chan, R. H. et al. (2016). Computational classification of different wild-type zebrafish strains based on their variation in light-induced locomotor response. *Computers in biology and medicine*, *69*, 1–9.

Gay, J. L., Buchner, D. M. & Smith, J. (2018). Occupational physical activity opposes obesity: A cross-sectional modern replication of the morris 1953 london busmen study. *Journal of occupational and environmental medicine*.

Gay, J. L., Buchner, D. M., Smith, J. & He, C. (2017). An examination of compensation effects in accelerometer-measured occupational and non-occupational physical activity. *Preventive medicine reports*, *8*, 55–59.

Gu, C. (2013). *Smoothing spline anova models* (Vol. 297). Springer Science & Business Media.

Kipf, T. N. & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kopp, R., Legler, J. & Legradi, J. (2018). Alterations in locomotor activity of feeding zebrafish larvae as a consequence of exposure to different environmental factors. *Environmental science and pollution research*, *25*(5), 4085–4093.

Liu, Y., Carmer, R., Zhang, G., Venkatraman, P., Brown, S. A., Pang, C.-P., Zhang, M., Ma, P. & Leung, Y. F. (2015). Statistical analysis of zebrafish locomotor response. *PloS one*, *10*(10), e0139521.

Liu, Y., Ma, P., Cassidy, P. A., Carmer, R., Zhang, G., Venkatraman, P., Brown, S. A., Pang, C. P., Zhong, W., Zhang, M. et al. (2017). Statistical analysis of zebrafish locomotor behaviour by generalized linear mixed models. *Scientific Reports*, *7*(1), 1–9.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, *2*(11), 559–572.

Rihel, J., Prober, D. A., Arvanites, A., Lam, K., Zimmerman, S., Jang, S., Haggarty, S. J., Kokel, D., Rubin, L. L., Peterson, R. T. et al. (2010). Zebrafish behavioral profiling links drugs to biological targets and rest/wake regulation. *Science*, *327*(5963), 348–351.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I. & Welling, M. (2018). Modeling relational data with graph convolutional networks. *European Semantic Web Conference*, 593–607.

Scott, C. A., Marsden, A. N. & Slusarski, D. C. (2016). Automated, high-throughput, in vivo analysis of visual function using the zebrafish. *Developmental Dynamics*, *245*(5), 605–613.

Sun, Y., He, T., Hu, J., Huang, H. & Chen, B. (2019). Socially-aware graph convolutional network for human trajectory prediction. *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 325–333.

Swierczek, N. A., Giles, A. C., Rankin, C. H. & Kerr, R. A. (2011). High-throughput behavioral analysis in c. elegans. *Nature methods*, *8*(7), 592–598.

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Vignet, C., Bégout, M.-L., Péan, S., Lyphout, L., Leguay, D. & Cousin, X. (2013). Systematic screening of behavioral responses in two zebrafish strains. *Zebrafish*, *10*(3), 365–375.

Ying, Z., You, J., Morris, C., Ren, X., Hamilton, W. & Leskovec, J. (2018). Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 4800–4810.

Zhang, G., He, H. & Katabi, D. (2019). Circuit-gnn: Graph neural networks for distributed circuit design. *International Conference on Machine Learning*, 7364–7373.

Zhang, L., Chong, L., Cho, J., Liao, P.-C., Shen, F. & Leung, Y. F. (2012). Drug screening to treat early-onset eye diseases: Can zebrafish expedite the discovery? *The Asia-Pacific Journal of Ophthalmology*, *1*(6), 374–383.

Zhang, L., Xiang, L., Liu, Y., Venkatraman, P., Chong, L., Cho, J., Bonilla, S., Jin, Z.-B., Pang, C. P., Ko, K. M. et al. (2016). A naturally-derived compound schisandrin b enhanced light sensation in the pde6c zebrafish model of retinal degeneration. *PloS one*, *11*(3), e0149663.

Zhao, K., Xiao, W., Ai, B., Shen, W., Zhang, X., Li, Y. & Lin, W. (2019). Aligraph: An industrial graph neural network platform.

Zhu, R., Zhao, K., Yang, H., Lin, W., Zhou, C., Ai, B., Li, Y. & Zhou, J. (2019). Aligraph: A comprehensive graph neural network platform. *Proceedings of the VLDB Endowment*, *12*(12), 2094–2105.