Pathway-based Lineage Analysis of Time-course Single-Cell RNA Sequencing Data

by

Thinh Ngoc Tran

A thesis submitted in conformity with the requirements
for the degree of Master of Science

Department of Molecular Genetics
University of Toronto

© Copyright by Thinh Ngoc Tran 2019

Pathway-based Lineage Analysis of Time-course Single-Cell RNA Sequencing

Thinh Ngoc Tran

Master of Science

Department of Molecular Genetics
University of Toronto

2019

Abstract

Time-series single-cell RNA sequencing (scRNAseq) can capture heterogeneity in cell states and transitions during dynamic biological processes, such as development and differentiation. Many trajectory inference methods have been developed to order cells by their progression through a dynamic process and infer the cells' movement trajectory. These methods, however, do not consider time information when ordering cells. In this thesis, I present a novel method, called Tempora, that uses pathway expression profiles and experiment timepoint information to infer the lineage relationships among different cell populations captured in time-series scRNAseq experiments. Tempora accurately inferred developmental lineages and important time-dependent signaling pathways in human skeletal myoblast differentiation and murine cerebral cortex development time-series scRNAseq data. These results demonstrate the power of using time information, when available, to supervise trajectory inference, as well as suggests that pathway expression profiles are an informative alternative to gene expression profiles in representing individual cells for scRNAseq analysis.

Acknowledgments

First and foremost, I would like to thank my thesis advisor, Dr. Gary Bader, whose continuous support and tremendous patience have made this work possible. From bi-weekly progress meetings to endless emails that keep me updated with the literature, Dr. Bader has not only introduced me to the wide intriguing world of computational biology but also patiently guided me through every step of the research process and helped me solve any problem that stopped me in my track. I would forever be grateful for the extra miles Dr. Bader has gone to help me succeed, even with the limited amount of time I have got to spend in his lab.

My gratitude also goes to my thesis committee members, Dr. Gordon Keller and Dr. Ian Scott, for their thoughtful feedback and advice throughout my progress. I have also grown tremendously as a result of my conversations and collaborations with the wonderful postdocs and fellow graduate students in the Bader and Keller Lab. In particular, I am indebted to Dr. Maria Abou Chakra, whose perspectives and approaches to science inspire me not only at work but also in life. Maria, thank you for the wisdom, the coffee and the faith that you have in me – they have lifted me up from so many low points over the past years and helped remind me why I chose this path in the first place. I am also thankful for the advices I have received from Brendan Innes, Owen Whitley, Dr. Ruth Isserlin, Dr. Shraddha Pai, as well as other members of the Bader Lab who have graciously answered my Slack inquiries from day one. I am grateful for Dr. Stephanie Protze, Dr. Jee Hoon Lee, Donghe Yang and other members of the cardiac team at the Keller Lab for their help in acquainting me with cardiac biology, feeding me with food for thought and clarifying my confusion about the first and second heart fields many, many times. Finally, my big thanks to Patrick Scopa, Anna Corpus and Mary White, who have consistently helped me navigating changes in my calendar and ensured my forms were signed on time.

Last but not least, I would like to acknowledge my family and friends, who have stuck with me through ups and downs as I completed this thesis. To my parents and sister – thank you for your unswerving belief in me and what I do, even when I couldn't explain to you my research in proper Vietnamese (or a form of English that wouldn't confuse you). Finally, Nghiem – thank you for helping me see the light in the worst of time and taking care of me when I forgot to do so. This would not have been as wonderful of a journey without you.

Table of Contents

List of Figures

Chapter 1
Introduction

1 Dynamic processes in biology

Living organisms are biological systems that maintain their existence and wellbeing through a network of continuous processes to respond to stimuli, regenerate and reproduce[1]. These dynamic processes involve constant changes of constituents at different scales, from molecular changes in metabolic processes to emergence of differentiated cell types from stem cells during development. They are initiated and regulated by multiple interactions among regulators, including genes, RNA, proteins, posttranslational modifications and epigenetics[2]. The understanding of these processes, both by themselves and as parts of an integrated network, is instrumental in understanding how organisms function and elucidating the mechanisms of disease.

To develop a comprehensive understanding of dynamic biological processes and how they are regulated, it is necessary to identify the components involved, characterize their functions and connections to each other, as well as observe how they evolve over time. Two main approaches are often used to tackle these tasks. The reductionist approach involves studying separate components of a process and their connections, which have been successful in elucidating the mechanisms of multiple processes, such as cellular respiration and electrochemical signaling within the nervous system[3,4]. However, since a biological components can participate in multiple processes, it is necessary to study these components in the network context to fully understand their roles in different processes as well as the interactions between these processes[3,5]. For example, signaling pathways can have one or more components in common that leads to their co-activation, a phenomenon known as pathway crosstalk[6]. A complete understanding of how the common components can affect changes necessitates a systems-level investigation that considers both signaling pathways as well as the spatiotemporal context of the components' activity[7]. Beyond elucidating the mechanisms of how processes work, the network approach is also useful in studying how processes are regulated, which often involves multiple genes or gene products engaging in feedback and feed-forward mechanisms. The elucidation of gene regulatory networks (GRNs) involved in the regulation of multiple processes in different species as well as their time-variant activities has furthered our understanding of how processes such as immune reactions are regulated[3]. The development of technology to survey the structure and quantity of genes and gene

products at the omics level has enabled novel findings regarding the composition and behaviors of complex dynamic systems[3,7].

1.1 Developmental processes

One of the central dynamic processes in biology is development, in which a multicellular organism is created through the growth and differentiation of cells in an embryo. Development in many multicellular organisms follows basic, well-conserved principles, both at the cellular and molecular levels, to achieve the right quantities and types of cells in the adult body[8]. During development, a fertilized egg divides to grow in size and number, creating an embryo with genetically identical stem cells. These cells then undergo differentiation to give rise to distinct cell fates, thus contributing to different lineages in the organism. The differentiation process is regulated by a multitude of inputs, including various signaling pathways, epigenetics and environmental cues. These inputs lead to the differential expression of genes in offspring cells, which result in the distinct morphologies and functions that characterize different cell types.

The development of skeletal muscles exemplifies a typical developmental process. Skeletal muscles in the body develop from the somites, embryonic segments formed from paraxial mesodermal cells[9,10]. Cells in the dorsal part of the newly formed somites become muscle progenitors by activating myogenic regulatory factors, including Myf5, Mrf4 and MyoD[9]. These progenitors, known as myoblasts, proliferate extensively, then exit the cell cycle before differentiating into mononucleated myotubes[9,11]. These mononucleated cells then fuse to form multinucleated myofibers, which assemble into a continuous layer of muscle known as the myotome[9,11]. This process happens in two phases during development: an embryonic phase (embryonic day 10.5 (E10.5) to E12.5 in mouse) to form primary muscle fibers and a fetal phase (E14.5 to E17.5 in mouse) to form secondary fibers[9]. While primary muscles are mainly slow-twitch fibers (Type I), secondary muscles mostly become fast-twitch Type II fibers[10]. The patterning of the somites is tightly regulated by multiple gradients of different signaling pathways, including Notch, FGF and Wnt signaling, while the muscle regulatory factors as well as Shh, Wnt and BMP signals secreted by the surrounding environment regulate myogenesis from myoblasts[10,11]. Many questions about skeletal muscle development remain, particularly regarding the specification of myogenic progenitors in the somites, the integration of signaling pathways that control myogenesis, as well as the similarities in regulatory mechanisms of adult muscle

regeneration and embryonic skeletal muscle development, which are common themes in developmental biology that necessitate further studies[9,10].

The question of how development can generate cell type diversity in multicellular organisms from homogeneous pools of stem cells has been of great interest, as some of these mechanisms are possible drivers of cancer and other development-related diseases. Efforts to answer this question have led to findings of important components in the specification mechanism and how their activities are coordinated during development, such as the spatiotemporal patterning of neural stem cells by cascades of transcription factors (TFs) that leads to different neuronal types in *Drosophila*[12,13] or the network of Sox TFs in murine development[14]. The development of high-throughput methods, such as CRISPR screens and automated image analysis of fluorescently-labeled progenitor nuclei has furthered investigations of various development regulatory circuits and their mechanisms in the network context[15-17].

Despite the advances we have made in understanding the mechanisms of cell fate specification, many questions remain unanswered, particularly in regard to the temporal trajectory and regulation of the process[18]. Time plays a central role in cell fate specification as it balances proliferation and differentiation to ensure the right number and type of cells are produced, patterns a developmental process into smaller windows in which multiple cell types can be specified, and orchestrates multiple developmental processes in an organism, yet little is known about how stem cells keep track of time and progress accordingly[18]. This question is challenging to explore, as cellular heterogeneity often hinders the study of temporal development at the single cell level. The development of novel techniques in fluorescence labeling, live imaging and next-generation sequencing, however, are starting to enable the measurement of temporal changes at greater resolution and allow us to start answering some of these long-standing questions[19-22].

2 Next-generation sequencing

2.1 Overview

Information about the components and regulation of all biological processes is encoded in genes and non-coding DNA regions of an organism, collectively known as the genome. Genome sequencing allows for the determination of the DNA sequences and structures that make up a genome, thus shedding light into the mechanisms of life and diseases. The complete human

genome, sequenced between 1990 and 2003 with Sanger sequencing, has become a tremendous resource to better understand the genome structure and variations[23]. Sanger sequencing, however, is expensive and not scalable to a large number of genes[24]. The development of next generation sequencing (NGS) platforms, starting with Roche 454 in 2005 and followed by Solexa's Genome Analyzer, Applied Biosystems' SOLiD and others, addressed the problem of high cost and low efficiency in traditional Sanger sequencing, thus spearheading the study of genomics in the next decade[24,25]. NGS has found tremendous applications in biomedical research and medicine, ranging from the reconstruction of phylogenetic trees between species[26] to genome-wide association studies to identify variants associated with certain traits[27] and whole genome sequencing for cancer diagnosis and treatment[28,29], among others.

2.2 Transcriptome sequencing

The development of NGS techniques has furthered studies of not only the genome but also the transcriptome, which is the collection of mRNAs, non-coding RNAs and small RNAs in a cell[30]. Beyond the transcriptional structures of individual genes, the transcriptome also carries information about a cell's activities in certain conditions, since the set of transcribed genes in a cell in part determines its states and active processes[30]. Transcriptomic studies are therefore suitable to investigate gene expression changes during dynamic processes. Prior to the development of NGS techniques, these studies were done by hybridization-based microarrays, which require the prior knowledge of genomic sequences and can be contaminated by cross-hybridization[30,31]. The use of NGS sequencing to sequence the transcriptome, known as RNA sequencing (RNAseq), has addressed these problems and enabled high-throughput quantification of gene expressions. RNAseq starts with the synthesis of complementary DNA strands (cDNA) of all available transcripts in the cells, followed by the fragmentation of cDNA and ligation of adaptors to cDNA fragments, a process known as library preparation[31]. The prepared library can then be sequenced on an NGS platform.

After sequencing, raw reads are analyzed for their GC contents as well as the presence of adaptors and duplicated reads to filter out any outliers[32]. Reads are then aligned to a reference genome or annotated transcriptome, and the expression levels of all transcripts are quantified as proportional to the number of cDNA fragments that align to each gene after correcting for transcript length and total number of reads[33]. Differential gene expression analysis, which involves

using statistical tests to quantify the changes in gene expression levels between different groups, can be done to identify genes that are up- or downregulated in certain conditions[34]. The large suite of computational and statistical tools designed to analyze RNAseq data has enabled researchers to extract important biological insights from their data that further our understanding of human diseases and normal development[33]. The breakthroughs in RNAseq technique development has also led to the creation of large databases containing transcriptomic profiles of different tissue types (Encyclopedia of DNA Elements[35]) and cancer tissues (The Cancer Genome Project[36], International Cancer Genome Consortium[37]), as well as repositories of high-throughput gene expression and functional genomics datasets (Gene Expression Omnibus[38], European Nucleotide Archive[39]), all of which provide the community with massive resources to pursue biological questions of interest.

3 Single-cell RNA sequencing

3.1 Overview

Even though traditional RNAseq has provided researchers unparalleled insights into gene expression changes in different conditions and during development, it can only generate an average population transcriptome and thus discard cell-to-cell transcriptional variability. This measurement is sufficient to understand biological systems under the assumption that all cells in a tissue are homogenous in their gene expression profiles. However, typical tissues are composed of many cell types; for example, mammalian skeletal muscles consist of different contractile fiber types, multiple connective tissues such as tendon and ligaments, as well as the extracellular matrix[40,41]. Cellular heterogeneity within a tissue has been characterized over the past years, especially in the context of tumors and their microenvironments, which comprise multiple cell types, as well as within developing tissues, where cell progress through development at different rates. The heterogeneity within a tissue is biologically meaningful, as it can influence cell fate decisions as well as responses to environmental stimuli[42]. Measuring the transcriptomes of single cells, therefore, is desirable to better understand such cellular heterogeneity and its role in dynamic processes. Furthermore, single-cell transcriptomics is valuable in identifying and studying rare yet important populations of cells, such as adult stem cells[43,44] and circulating tumor cells[45].

Gene expression studies at the single-cell level were first accomplished by a PCR-based protocol to synthesize and amplify cDNA transcripts from samples as small as a single cell that is

unbiased in respect to transcript length and abundance[46]. Later, microfluidic devices allowed for the automated isolation of single cells and enabled the quantification of gene expression with quantitative polymerase chain reaction (qPCR) or hybridization-based microarrays[47,48]. However, single-cell transcriptome sequencing still faced the challenges in efficiently isolating single cells and amplifying a small amount of RNA in each cell[48]. In 2006, Tang *et al.* improved the cDNA amplification process commonly used in microarrays to capture full-length cDNAs from single cells and successfully sequenced the transcriptomes of individual cells in a mouse embryo, effectively developing the first single cell RNA sequencing (scRNAseq) protocol[47]. Over the next few years, a plethora of scRNAseq methods have been developed, differing in techniques used to handle single cells as well as their transcripts[48]. These differences give each method unique strengths and weaknesses that make them suitable for specific applications.

3.2 Methods

Despite the diversity in approaches and chemistries, the framework for all high-throughput scRNAseq methods is similar. Single cells are first isolated, either by fluorescent-activated cell sorting (FACS), microfluidic chips or droplets, then cells are lysed with lysis buffer to release all transcripts. MARS-Seq[49] and SMART-seq2[50] use FACS to sort cells directly into wells on a 96- or 384-well plates. Fluidigm C1 uses an integrated microfluidic chip, which consists of a system of connected micrometer-sized channels and capture chambers, to isolate up to 800 single cells in one experiment[51]. These methods, however, are limited by the number of cells they can capture and their bias for certain cell sizes, thus restricting their applications in the study of larger and more heterogeneous cell populations[52-54]. Droplet microfluidics methods, such Drop-seq[52], inDrop[53] and 10X Genomics Chromium[54], overcome these limitations by encapsulating single cells in aqueous droplets, along with beads that contain barcoding oligonucleotide primers to label which cell each transcript comes from (Figure 1). Each barcoding primers also consists of a short unique molecular identifier (UMI) to identify PCR duplicates and a poly-T 3' end to select for transcript fragments at the 3' end, which are necessary for the 3'-tag sequencing process (Figure 1 and discussed below)[52-54]. After isolation, cells are lysed within the droplets and RNA molecules in each cell are barcoded with the oligonucleotide primers during reverse transcription. The droplets are then broken to generate a pooled library of all cDNA molecules to be reverse transcribed, amplified and sequenced (Figure 1)[55]. These droplet-based methods can vary in their cell capture efficiency: while Drop-seq can only capture 12.8% of the cells in a sample, 10X

Genomics Chromium can capture approximately 50% and inDrop up to 75% of cells[52-54]. However, the main drawback of these methods lies in their low mRNA capture efficiency: inDrop can only capture about ~7% transcripts in a cell, which means it cannot reliably detect genes with fewer than 20-50 transcripts[56]. These methods can therefore miss important genes that are expressed at lower levels[56,57].

After isolation and reverse transcription, single-cell libraries are sequenced with NGS platforms[48]. There are two main protocols to sequence RNA from single cells: full-length and 3'-tag RNA sequencing. Full-length RNAseq, used by SMART-seq2 and Fluidigm C1, is similar to the bulk RNAseq technique, in which all transcripts in a cell are fragmented and sequenced, then gene expression levels are quantified by the number of reads that align to certain genes in a in reference genome[50]. The majority of scRNAseq methods use 3'-tag RNAseq, which only amplifies and sequences one fragment in the 3' region of each transcript, selected by hybridization of the 3'-poly-A tail of the transcript with a poly-dT primer on the barcode DNA[48,58]. Before fragmentation, each transcript is tagged with a unique molecular identifier (UMI), a short (~5-10 bp) random sequence that is part of the barcode DNAs used to label individual beads in droplet-based platforms[52-54]. After PCR amplification, molecules with the same UMI are assumed to come from the same input transcript[59]. While originally designed to detect PCR amplification bias, UMIs have also been used to directly quantify gene expression in scRNAseq through counting the number of distinct UMIs that map to a gene[59]. Both full-length and 3'-tag RNAseq have distinct advantages and disadvantages. While full-length RNAseq gives more information about transcript structures, including splicing sites and isoforms, it requires higher sequencing depth and thus increases cost. 3'-tag RNAseq reduces the number of reads required per sample and thus serves as a low-cost alternative to full-length scRNAseq, while still provides sufficient information to quantify relative gene expression[58,60].

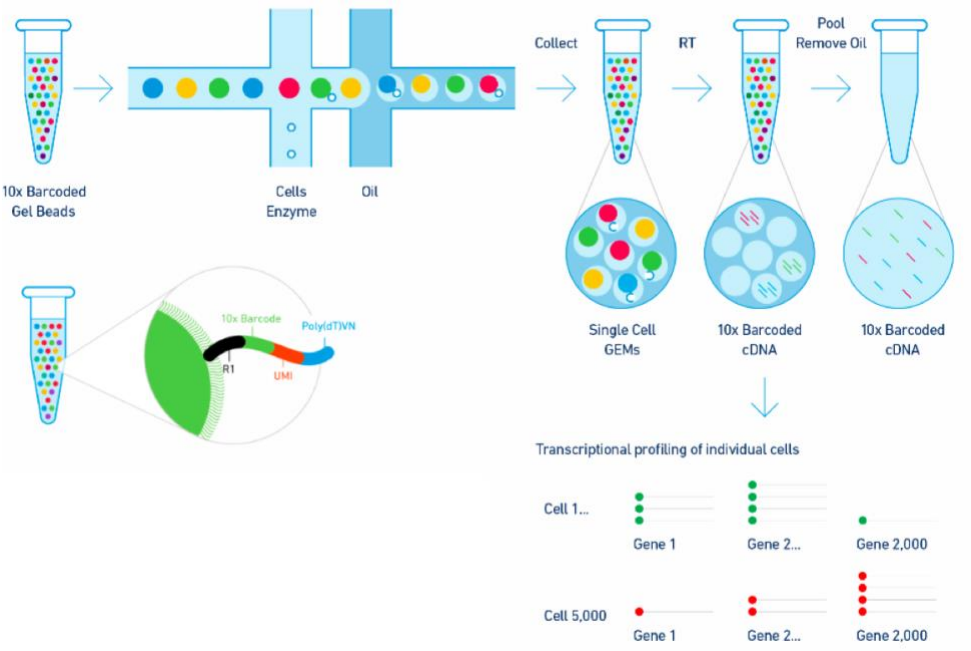

Figure 1. Overview of barcoded bead design (bottom left) and scRNAseq protocol using 10X Genomics Chromium, a popular platform using droplets to isolate cells. Figure adapted from[61].

Current scRNAseq techniques still face multiple limitations in detecting low-abundance transcripts (sensitivity) and quantifying gene expression (accuracy)[48,62]. While droplet-based techniques can capture more cells, they suffer from low sensitivity as the number of genes they detect tends to be lower compared to lower throughput methods, which leads to high number of drop-outs and zero inflation, which is an excess of zero counts in the data[53,60]. 10X Genomics Chromium can detect on average 4,500 genes and ~20,000 transcripts in their V2 chemistry, which is approximately 14-15% of all transcripts in a cell, while SMART-seq2 can detect 7,500 genes on average per cell[60]. The development of new chemistry for droplet-based techniques, such as 10X Genomics Chromium's V3 chemistry which claims to detect up to 32% of all transcripts per cell, is expected to improve the sensitivity of these methods[60,63]. On the other hand, non-UMI methods such as SMART-seq2 suffer from noisier gene expression quantification, since varying amounts of cDNA are lost during reverse transcription and amplification[58]. As small differences in transcript abundance can lead to tremendous biological consequences, especially for TFs or long non-coding RNAs in the context of development, improvements on sensitivity and accuracy are necessary so that scRNAseq can more faithfully capture the transcriptional profiles underlying cell states and processes.

3.3 Bioinformatics analysis of scRNAseq data

3.3.1 Overview

Bioinformatics analysis of scRNAseq data shares much in common with that of bulk RNAseq data, starting with quality control, read alignment to the reference genome and transcript quantification, followed by normalization and other downstream analyses to answer specific biological questions (see 2.2)[55]. However, due to the much larger number of data points and higher level of variance, comprising both biological differences and technical noise, separate bioinformatics pipelines, such as Seurat[64], Scater[65] and SCANPY[66], are being developed to analyze scRNAseq data, aiming to reveal novel biological insights while accounting for technical drawbacks. The basic pipeline includes preprocessing of count data, such as filtering out low quality cells and batch effect correction, followed by normalization, dimension reduction, clustering and differential gene expression analysis[55,67]. Similar to the diversity in experimental platforms, many computational methods for scRNAseq are developed to handle one or multiple of the aforementioned tasks, each with their own strengths and weaknesses[67].

3.3.2 Quality control and batch effect correction

Multiple filtering steps are first applied to the count data in order to ensure that low quality reads and damaged cells are eliminated from further analysis. In the first step of quality control, cells with abnormally high mitochondrial RNA content, which ranges from higher than 5% of total mRNA in normal cells or higher than 30%-50% in cells with high energy requirements such as cardiomyocytes and hepatocytes[68,69], are filtered out, since a high concentration of mitochondrial RNA reads indicates compromised cell membranes and thus damaged cells[70]. Doublets, which are samples resulting from two or more cells being captured and sequenced in the same droplet, are next filtered out by looking for cells with abnormally large library sizes[70]. These outlier cells, if not removed, can influence downstream analyses by clustering together due to their library sizes. Low abundance genes are also removed as they are likely to be contaminant and can drive downstream clustering[70].

ScRNAseq datasets generated in different experiments and sequencing runs can be affected by batch effect, in which technical and experimental variations, such as sequencing depth, experiment time, reagents and instruments, lead to biases in gene expression values[71,72].

Uncorrected batch effect can reduce power of downstream statistical analyses, leading to a higher number of false negatives in differential expression analyses[72]. Furthermore, explicit modeling of batch effects is required in order to directly compare datasets obtained in different conditions. Multiple algorithms have been developed to account for these variations, including mutual nearest neighbors (MNN)[73], canonical correlation analysis (CCA)[64] and Harmony[74]. These methods project cells into a common reduced dimensionality space (see 3.3.3) to identify shared cell types and correct for batch effects[64,71,73,74]. The output of these methods is a batch-corrected gene expression matrix, which results from multiplying the input expression matrix with cell-specific correction factors inferred by the respective algorithms.

3.3.3 Dimensionality reduction

ScRNAseq data are high dimensional, as each gene captured represents a dimension along which cells can vary. It is difficult to meaningfully visualize all data points in such a high-dimensional space; thus, a dimensionality reduction (DR) method is often required to project scRNAseq data to a lower dimensional space for visualization and downstream analyses[48,75]. Principal component analysis (PCA), one of the most popular DR methods, finds linear combinations of variables, i.e. genes in scRNAseq, that explain the largest amount of variance in the original data[76]. Each linear combination represents a principal component onto which cells can be projected. Even though PCA performs well in identifying important dimensions that contribute to variance for downstream analyses such as clustering (see 3.3.5), as a linear transformation tool, it cannot capture the nonlinear relationships in scRNAseq data[75]. Other DR methods have been developed to overcome this issue, most notably t-Stochastic Neighbor Embedding (t-SNE)[77,78] and Uniform Manifold Approximation and Mapping (UMAP)[79]. t-SNE maps data points, i.e. cells in scRNAseq data, from a high dimensional space to a 2-dimensional space to so that the closeness of data points in both spaces is preserved, which allows for the visualization of these cells on a 2D space[77,78]. The preservation of local structure makes tSNE useful in visualizing cell clusters, even though the algorithm itself is not designed for clustering. tSNE is still restricted by its ability to map large number of cells (> 30,000 cells) and its tradeoff of global structure preservation[77]. UMAP has emerged as a novel DR method that is able to preserve both global and local data structure in the low-dimensional visualization, allowing users to infer similarity between clusters by their distance on the plot[79]. These methods have made significant improvement in visualization of large scale datasets, such as scRNAseq data (Figure 2).

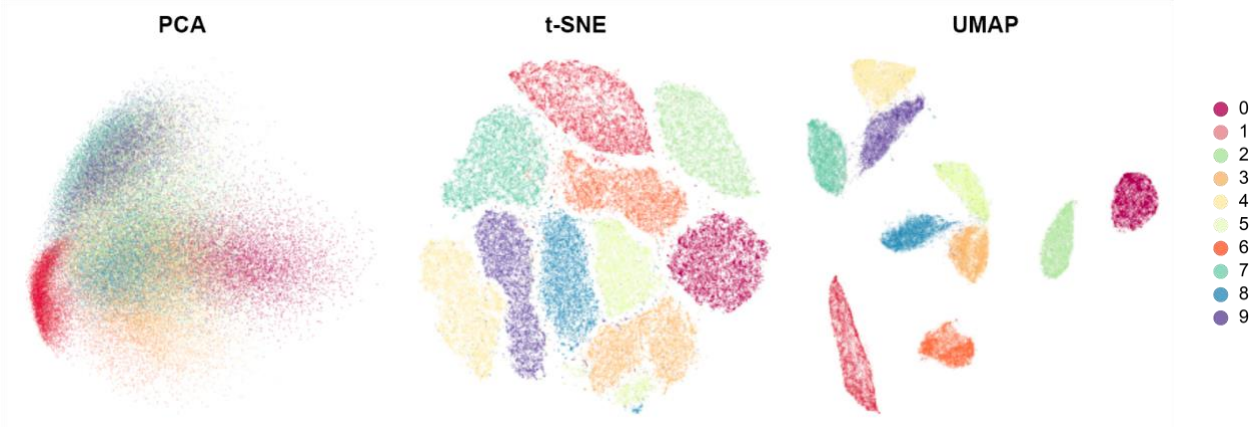

Figure 2. Visualization of 60,000 images of handwritten digits in the MNIST database[80] using the first two components of PCA (left), t-SNE (middle) and UMAP (right). Each dot represents a datapoint (handwritten image), and each cluster of points represents the collection of handwritten images of a digit. PCA is not able to discern the clear clusters of datapoints in the dataset, which tSNE and UMAP can. The clusters in the t-SNE plot are placed close together and the pairwise distances between them do not necessarily signify how different they are. UMAP better preserves global structure in the dataset to identify groups of clusters, whose pairwise distances correlate to their degree of difference. Figure adapted from[79].

3.3.4 Normalization

ScRNAseq data is characterized by features distinct from bulk RNAseq data, including high sparsity and variability[62]. A large proportion of scRNAseq read counts is zero, which occurs both due to low sensitivity of the sequencing techniques (discussed in 3.2) and lack of gene expression in certain cell population[48]. Furthermore, gene expression measurements are affected by systematic biases, such as endogenous mRNA contents, transcript capture and reverse transcription efficiency, number of reads, etc. that vary between cells[81]. The raw transcript counts between cells, therefore, are not on the same scale and cannot be directly compared. The aim of normalization methods is to bring all gene expression measurements to a common scale by removing cell-specific biases[81]. The first common normalization approach, adapted from bulk RNAseq, is library-size normalization, including reads or fragments per kilobase million (RPKM or FPKM) and transcript per kilobase million (TPM). These methods standardize the number of reads across all cells, but mask the differential expressions of genes and thus skew downstream analyses[75,81]. Other bulk RNAseq normalization methods that have been adapted for scRNAseq use with various degree of success is DESeq[82] and Trimmed Mean of M-value (TMM)[83], both of which determine cell-specific normalization factors as the median (DESeq) or weighted mean (TMM) of gene-wise cell-to-reference ratios in each cell. These methods better account for

differentially expressed genes, but perform poorly with the sparse scRNAseq data[55]. Normalization methods specific for scRNAseq have been developed, such as the deconvolution method implemented in the *scran* R package, which pools cells with similar gene expression profiles and library sizes together to normalize[84]. As they are more robust to characteristics of scRNAseq data in their algorithm, these methods outperform others in estimating the true cell-specific scaling factors in simulated datasets[81].

3.3.5 Clustering

One of the primary applications for scRNAseq is to characterize the heterogeneity in a biological sample, which involves detecting the distinct cell subpopulations present. These subpopulations can represent cell types that have been previously studied and catalogued in the Cell Ontology[85], e.g. progenitors, immune cells, cardiomyocytes, or stable cell states, which are physiological conditions that cells can be in, e.g. cycling, proliferating or metabolizing[86]. The computational problem of grouping similar cells into subpopulations is known as clustering, and has been addressed with a diverse group of algorithms, most notably k-means, hierarchical and graph-based clustering[87]. To reduce computational resources required, these clustering methods are usually run downstream of PCA, so all distances calculated are in reduced dimension space. K-means clustering, implemented in SC3[88] and RaceID[89], initiates *k* cluster centers and iteratively assigns cells to the closest cluster center until the distances between the cluster centers and their respective cluster members are minimized. Hierarchical clustering, implemented in Mpath[90] and BackSPIN[91], sequentially splits cells up into smaller clusters or adds cells to form larger clusters based on distances between cells. Both of these algorithms have high computational complexity and are thus not scalable for large scRNAseq datasets, while graph-based clustering algorithms often perform better in terms of speed. SNN-Cliq, implemented in Seurat, first calculates the Euclidean distance between cells, then list *k* nearest neighbors for each cell and determine similarity scores based on the number of common nearest neighbors that each pair has[92]. A graph with cells as nodes and weighted edges as similarity scores is constructed, and clusters are defined based on subgraphs[92].

Despite new advances in algorithm development, challenges regarding clustering in scRNAseq remain. In particular, the process often relies on user-input parameters to determine the number of final clusters, and no gold standard exists to guide the selection of these parameters[87].

Researchers need to select the clustering resolution based on prior knowledge of marker genes, which can be unavailable for certain systems. Tools such as scClustViz[93] and TooManyCells[94] have been developed to help users visualize clusters at different resolutions, analyze clusters' relationships across resolutions as well as investigate marker gene expression, which could ameliorate some difficulties with clustering. However, without a consensus guideline on how these issues should be handled, clustering results from different studies can still remain difficult to compare and replicate.

3.3.6 Differential gene expression (DEG) analysis

A typical analysis of scRNAseq data is to identify genes that are differentially expressed in each subpopulation, which can drive biological processes in certain cell states or determine the fate of certain cell types. Beyond the adaptation of DEG methods from bulk RNAseq data, including edgeR[95] and DESeq[82], many scRNAseq-specific tools have been develop to account for scRNAseq features such as high dropouts, including SCDE[96], MAST[97] and others[98]. These methods share the common framework of modeling read counts or expression values with a probability distribution, then using a statistical test with false discovery rate adjustment to identify DEGs and their significance[98,99]. EdgeR and DESeq model gene expression using the negative binomial distribution and use an exact test to determine DEGs[82,95]. SCDE[96] account for dropouts in scRNAseq data by modeling gene expression using a mixture probabilistic model, while MAST[97] fits a two-part generalized linear model to model the rate and levels of expression of individual gene, then uses a likelihood test for DEGs. Other than these methods, general non-parametric statistical tests, such as the Wilcoxon rank-sum test, have also been used to test for DEGs, with comparable accuracy to aforementioned methods but with lower computational complexity and easier parallelization, resulting in greater computing speed[99].

3.3.7 Pathway enrichment analysis

Along with the long list of genes generated through DEG analysis comes the challenge of interpreting their functions. Pathway enrichment analysis (PEA) provides a solution to this issue by identifying enriched pathways, which are sets of genes that work together to carry out certain biological functions in the cells, from a given list of DEGs[100]. PEA requires a pathway database consisting of annotated gene sets that have been curated from the literature, e.g. Gene Ontology[101], Reactome[102] and KEGG[103]. The enrichment of pathways can be evaluated using overrepresentation

analysis, which identifies gene sets containing more genes in a list of DEGs than expected by random and tests for their significance using statistical tests such as chi-square or Fisher's exact test, or functional class scoring, which calculates an enrichment score for every gene set based on the position of its gene in a ranked list of DEGs and tests for significance of the enrichment score[104]. The second approach has been shown to be more robust in detecting subtler changes in pathway enrichment and is implemented in two popular PEA methods, Gene Set Enrichment Analysis (GSEA)[105] and Gene Set Variation Analysis (GSVA), the single-sample variation of GSEA[106]. The output of these methods is a list of enriched pathways and their associated statistics. Enrichment results can then be visualized as networks with software such as Cytoscape[100] and EnrichmentMap[107], which is useful in identifying closely related pathways that are enriched in a population and determining common themes in enriched pathways[100,108].

3.4 Applications

ScRNAseq has found many applications in studying cellular heterogeneity in different biological contexts, including cataloguing the different cell types that make up normal organ systems, dissecting the tumor microenvironment and other diseased tissues, discovering novel GRNs and splicing patterns, among others[109]. One of the most prominent and ambitious applications of scRNAseq is the Human Cell Atlas, a collaborative effort to create a reference map of all human cell types, defined by their unique transcriptional profiles, in the hope of better understanding human physiology and diseases[110].

Beyond static cellular heterogeneity, scRNAseq can be used to investigate transcriptional changes over time in dynamic systems. However, since cells are destroyed during sequencing, scRNAseq cannot track the same population over time and correlate the measurements to give direct insights on temporal dynamics of a process. It can only generate a snapshot of cells at various states along the process, which allows researchers to understand the continuum of transcriptional changes that a single cell goes through from beginning to end. Computational tools are then required to model this continuum of change and recapitulate cell trajectories based on such continuum (further discussed in section 4 below). Even though this line of thinking lends itself naturally to studies of development (reviewed in 3.5), it is applicable to any dynamic process. In particular, it has been used study the progression of diseases such as osteoarthritis[111] and

dengue[112], cellular responses to influenza infection[113] and lung inflammation[114], chemoresistance in breast cancer[115] and squamous cell carcinoma[116], among others.

3.5 ScRNAseq in studies of developmental processes

ScRNAseq has been commonly used to investigate development, both *in vivo* processes as well as *in vitro* differentiation and regeneration[117]. The continuum of cell states captured in a snapshot scRNAseq experiment of a developing system is useful to identify intermediate stages and study the transcriptional changes that occur during the process. Snapshot scRNAseq studies have been used to investigate multiple aspects of development of the early embryo[118,119], blood[120-123], kidney[124], various areas of the brain including the cortex and the hippocampus[125-127], among other systems. These studies have shed light into some common themes in development: the heterogeneity of progenitor populations, the multilineage priming and plasticity in fate choices of these progenitors, as well as the complex GRNs that regulate cell fate decisions. Furthermore, scRNAseq has been used to compare the developmental trajectories of cells during *in vitro* differentiation and in organoid models with known developmental trajectories, such as the human cerebral cortex[128], kidney[129], liver[130] and intestine[131]. The comparisons between *in vitro* and *in vivo* development help researchers evaluate their differentiation protocols' ability to recapitulate the right phenotype and maturity levels of the desired cell types and improve them so they can be used for future regenerative therapies and models of diseases[132].

Even though snapshot scRNAseq can provide novel insights into development, as a static measurement it cannot fully explain the dynamics of the process[133]. Furthermore, cells undergoing developmental processes are constantly changing as a result of changes in their transcriptomes; therefore, populations that appear earlier or later than the sampling time cannot be studied by snapshot scRNAseq. To capture populations distinct to certain phases of development and study transitions more closely, the system can be sampled and sequenced at different time points. With the expansion of scRNAseq techniques that help decrease sequencing cost, time-series studies of development processes are becoming increasingly common. The scale of these studies range from small experiments of a few hundred cells captured over a small number of time points to study a specific system, to a few hundred thousand cells captured over the course of embryonic development of a whole animal. In particular, time-series scRNAseq has been applied to study the development of many organs, including the lung epithelium[134], areas of the brain including the

cerebral cortex[135], prefrontal cortex[136] and cerebellum[137], pancreatic islets[138], liver[139], fetal germ cells[140], heart[141,142] and skin[143]. On a larger scale, two studies sequenced more than 100,000 cells over 18 time points during the development of the zebrafish embryo and computationally reconstructed the lineage relationships between them[144,145]. These studies resulted in detailed embryonic development maps that highlight the non-linear nature of developmental trajectories as well as novel connections between lineages. Furthermore, time-series scRNAseq is also a powerful tool to compare *in vitro* differentiated cells to *in vivo* cells, as it enables researchers to determine the maturity level of their *in vitro* cells. For example, by comparing cardiomyocytes derived from mouse embryonic stem cells at day 20 with mouse cardiomyocytes sequenced at 8 time points during embryonic development and after birth, DeLaughter *et al.* showed that cells at day 20 in culture are most similar to cardiomyocytes at E14.5[146]. The available time information in time-series scRNAseq, thus, gives these studies more discovery power and allows them to answer questions regarding the dynamics of development that would not otherwise be possible with snapshot scRNAseq experiments.

4 The lineage inference problem in scRNAseq

4.1 Problem overview

When using scRNAseq to study dynamic processes, whether through snapshot or time-series experiments, it is of interest to order cells at different stages along an axis that represents how far along they are on the process under study based on their transcriptional signatures. The ordering problem, commonly termed pseudotime ordering if it is inferred from data without a known temporal ordering, consists of two main parts: the identification of a trajectory representing the paths that cells go through, and the determination of pseudotime value for individual cells along this trajectory. This inferred trajectory allows us to study the sequential changes of gene expression during a process, as well as identify branches and instrumental genes at the branching points.

4.2 Previous work

More than 70 computational methods to order cells along pseudotemporal axes, known as trajectory inference (TI) methods, have been published to date, which employ different strategies to infer lineage and order cells[147]. Most TI methods are developed based on the basic premise that

cells closer in developmental time have more similar gene expression signatures, thus a likely trajectory is a path that maximizes cell-to-cell similarity. To deal with high-dimensional and noisy scRNAseq data, these methods often implement one or more DR techniques before constructing the trajectory in a reduced dimension space that captures most of the meaningful variance in the data. Common strategies for trajectory construction include fitting a minimum spanning tree (MST), which connects all data points in a path that minimizes distance between points, nonlinear dimensionality reduction (DR) that describes the low-dimensional manifold that cells lie on, or graph-based methods. In this section, I provide a brief review of the notable TI methods that have performed well in multiple studies. For a more comprehensive evaluation and review, Dynverse serves as an excellent resource to explore available methods as well as their strengths and weaknesses[147].

The first TI method that spearheaded pseudotime analysis is Monocle, which determines the trajectory by calculating all pairwise Euclidean distances between cells in a reduced dimension space, then connecting all cells in a tree that would minimize the distance between them (minimum spanning tree, or MST). The longest path through this tree is determined as the backbone of the trajectory and cells are then ordered along this path[148]. As the first method of its kind, Monocle has been applied to many studies, including cell fate decisions in hematopoiesis[149], differentiation trajectory of mesencephalic dopamine neurons[150] and asexual maturation of malaria parasites[151]. Despite its proven performance, the accuracy of Monocle's inferred trajectory can be compromised because of its reliance on cell-to-cell distance, which allows outlier populations that do not belong on the same lineage to easily skew the MST[152]. Other TI methods have been developed using other approaches to construct a more robust trajectory. To decrease the complexity of the trajectory space and thus limit the impact of outlier population, methods such as Waterfall, TSCAN and Slingshot first cluster cells, then build an MST on cluster centroids and determine their pseudotime values based on their positions on the MST[127,152,153]. These methods differ in their DR methods (TSCAN and Waterfall use PCA while Slingshot is designed to work with multiple different DR methods), as well as how they determine pseudotime after MST construction (Waterfall and TSCAN orthogonally project cells onto the MST, while Slingshot constructs simultaneous principal curves to project cells onto a nonlinear manifold). Besides MST, another major approach to constructing trajectory is to use nonlinear DR methods, such as principal curves and trees, to identify the manifold along which cells lie. Nonlinear DR methods are more robust to noise than

MST and less likely to overfit small datasets. Diffusion map, which uses a diffusion distance metric to construct the paths that cells take towards differentiation, serves as the base for other TI methods, such as Palantir and Wishbone[154-156]. Monocle 2, released in 2017, uses reverse graph embedding to learn a principal tree, which is a version of principal curve that allows branches through the data in a fully unsupervised manner, which constructs a more accurate lineage without any required input or parameters from the users[157]. Other methods rely on graph theory methods to construct trajectories[89,158]. For example, PAGA[158] makes use of graph partition to robustly detect an arbitrary number of lineages in a dataset with minimal assumptions about the underlying dynamics of the process, while SCUBA[159] measures similarity between two sub-clusters of every large cluster to determine whether cells in the parent clusters have differentiated into two distinct lineages (low similarity of sub-clusters) or continued on the same trajectory (high similarity sub-clusters), thus efficiently identifying bifurcating lineages.

The plethora of available TI methods, with diverse underlying algorithms, results in specific strengths and weaknesses, which renders these methods not one-size-fit-all. For example, some methods can only detect certain topologies of lineages, some make specific assumption about the underlying process, and some do not scale well with large number of cells and genes[147]. Selecting TI methods that work well with a specific dataset remains a difficult task. Furthermore, unlike other established downstream analyses such as differential gene expression or clustering, TI analyses produce diverse outputs with limited means to statistically evaluate each output to select the optimal solution. Methods such as TSCAN and p-Creode have proposed various methods to score inferred trajectories, such as a stability metric that measures how stable the TI method is when applied on subsamples of a dataset, a pseudo-temporal ordering score that correlates the cells' original collection time with the inferred pseudotime values, or a graph distance and topology metric to measure the similarity between two inferred trajectories formalized as graphs[152,160]. However, these scoring methods are usually difficult to apply on outputs of different algorithms as they are highly non-standardized. Integrated platforms such as Dynverse have begun to emerge to assist researchers with selecting the appropriate methods for their data as well as comparing outputs from multiple methods. These cross-method platforms provide standard benchmarks and enable method comparisons, thus identify the common weaknesses of TI methods, which pave the way for the development of novel future methods to address these issues and produce more reliable trajectories.

5 Conclusion

ScRNAseq has emerged as a powerful tool to study cellular heterogeneity and to some extent dynamic processes, especially when applied at multiple time points. The use of scRNAseq to investigate dynamic processes, particularly developmental processes, is made possible by computational methods that order cells at different stages along a pseudotemporal axis. Even though time-series scRNAseq data provides more information on time-dependent changes in cell population composition and their transcriptional signatures, few available tools can analyze them. The majority of TI methods do not work with time-series data, as they do not consider time information in building trajectories even when such information is available. The only TI method for time-series scRNAseq infers a trajectory by connecting cells from earlier time points that express certain transcription factors to later cells that express the known targets of such factors[21]. While this method can provide information about a transcriptional network that controls branching events in a trajectory, it requires a curated database of transcription factors and their targets, which is not available for all species. Furthermore, the additional information regarding transcriptional network provided, though interesting, can be superfluous to researchers who are primarily interested in discovering the ordering of cells in their data. Thus, there is a clear need for a general TI method designed to work with time-series scRNAseq data, which would use time information to supervise the trajectory inference process. Finally, since scRNAseq gene expression measurement is easily affected by read depth and noises, sporadically high or low expression of certain genes can influence the constructed trajectory. Meanwhile, the concerted up- or downregulation of genes in pathways can provide a more biologically meaningful way to measure cell-to-cell similarity and infer trajectories accordingly. Therefore, I hypothesize that time-series scRNAseq studies will benefit from novel TI methods that take advantage of available time information in ordering cells, as well as of pathway information to both reduce noise in gene expression and reveal more biological insights into the cellular activities driving the progression of the dynamic process under study.

Chapter 2
Pathway-based trajectory inference in time-series scRNAseq data of developing systems

This chapter describes Tempora, a new algorithm for trajectory inference (TI) in time-series scRNAseq, and demonstrates its improved performance on two time-series datasets compared to other popular TI methods. I conducted the algorithm design and validation under the guidance of Dr. Gary Bader.

1 Algorithm overview

To address the lack of TI methods designed for time-series scRNAseq data, I developed Tempora, a method to infer developmental lineages in time-series scRNAseq data using pathway enrichment profiles. Since we can make reasonable assumptions about the progress of cells along a differentiation trajectory based on the time they are collected, the trajectory inference problem is simplified to identifying how cells at different time points are connected in a trajectory. To build a more robust trajectory less influenced by small outlier populations, Tempora first clusters cells with similar transcriptional signatures together and infers a lineage that connects cell clusters rather than individual cells. Second, to increase the biological interpretability of the trajectory and reduce the effect of sporadic expression of individual genes on the trajectory, Tempora constructs trajectories based on pathway enrichment profiles of cell clusters.

Tempora takes as input a preprocessed gene expression matrix from time-point scRNAseq experiments, and provides users with tools to assess and correct batch effects as needed. Once the data is clustered, Tempora calculates the average gene expression profiles, or centroids, of all clusters in the data before transforming the data from gene expression space to pathway enrichment space using single-sample pathway analysis (Figure 3). To remove pathways that do not contribute to variance as well as redundancy in representation of some well-studied biological processes, such as the cell cycle, in pathway databases, Tempora applies PCA on the pathway analysis result and selects important PCs that explain at least 85% of the variance in the dataset. Pathways with high loadings on those PCs are used to construct the trajectory in the next step.

To infer the trajectory between cell clusters from different time points, each cluster is characterized by a pathway enrichment profile. I abstract the backbone of the trajectory as a

network of cell clusters, with vertices representing the states/clusters and edges representing transitions between states. To infer this network, Tempora uses ARACNe[161], an established algorithm for predicting cellular network based on mutual information (MI). ARACNE filters the network using the data processing inequality to remove edges with the smallest MI in all triples, which helps remove indirect connections (Figure 3). After constructing the backbone, Tempora makes use of available temporal information from the input data to determine edge directions. First, each cluster is assigned a temporal score corresponding to its cell composition from different time points, so that a cluster containing more cells from an early time point will have a low score and vice versa. The edges are then directed so that their sources will have a lower score than their targets, indicating a transition from an early cell state to a later cell state. The trajectory is visualized with a hierarchical layout.

Tempora features a downstream pathway exploration tool to determine and visualize time-dependent pathways. These pathways are identified by fitting a generalized additive model (GAM) to the enrichment information of each pathway across all clusters and selecting pathways whose expression patterns deviate significantly from the null model of uniform pathway enrichment scores across all time points.

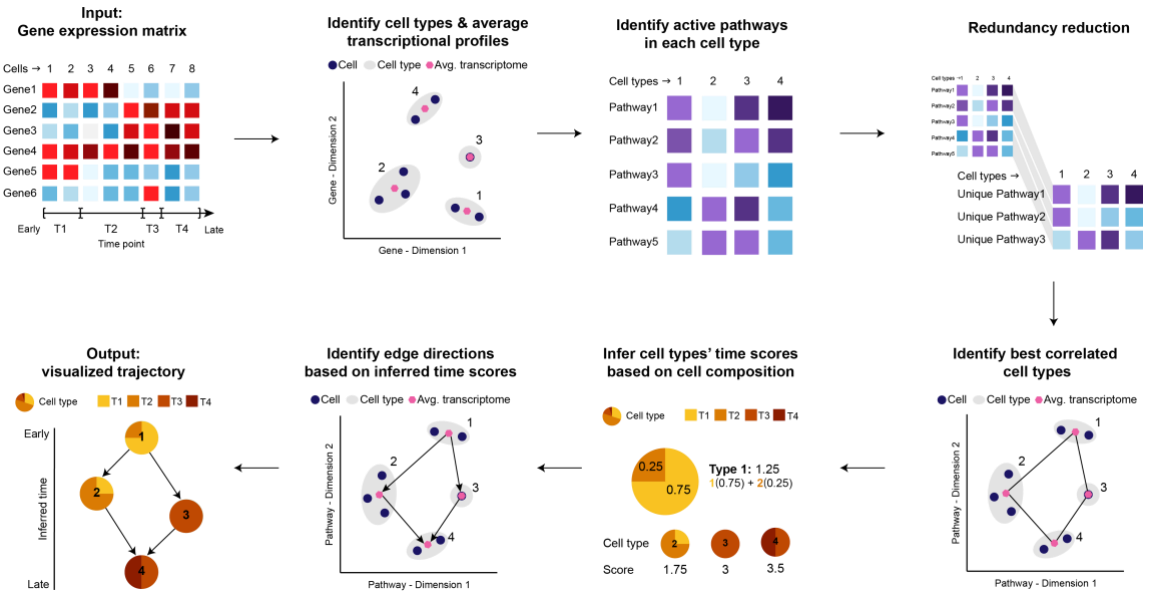

Figure 3. Schematic of the Tempora algorithm.

2 Algorithm components

2.1 Data preprocessing, batch effect correction and clustering

Tempora takes processed scRNAseq data as input, either as a gene expression matrix with separate time and cluster labels for all cells, or a Seurat object containing gene expression data and a clustering result. It is important to note that Tempora does not implement clustering as part of its pipeline and assumes that the researchers have inputed a well-annotated cluster solution into the method. If users indicate that they have not corrected for batch effect, Tempora will run the kBET[162] method to determine if any batch effect is visible and ask users to rerun clustering after correcting for these effects with the Harmony data integration method[74].

2.1.1 Preprocessing of validation datasets

Two time-course scRNAseq datasets were used to validate Tempora, one on the *in vitro* differentiation of human skeletal muscle myoblasts (HSMM) and the other on an *in vivo* sample of early murine cerebral cortex development. HSMM read count data was accessed from the HSMMSingleCell R package[163]. Murine cerebral cortex data were downloaded from GEO at accession number GSE107122[135].

Both datasets were filtered to remove lowly expressed genes (defined as those found in less than 3 cells) and damaged cells with high mitochondrial contents (4 median absolute deviations above the median). After this initial filtering step, the murine cerebral cortex data were further filtered to remove non-cortical cells, as done in the original publication to focus the analysis on the cortical lineage[135]. These included cells expressing *Aif1* (microglia), hemoglobin genes (blood cells), collagen genes (mesenchymal cells), as well as *Dlx* transcription factors and/or interneuron genes (ganglionic eminence-derived cells)[135]. The datasets were then normalized using the deconvolution method implemented in the scran R package, which pools cells with similar gene expression profiles and library sizes together to normalize[84]. Afterwards, cells were iteratively clustered with the SNN-Cliq algorithm[92] implemented in Seurat at increasing resolutions until the number of differentially expressed genes between two neighboring clusters reached 0. I determined the clustering resolution and annotated all clusters by examining expressions of known marker genes using scClustViz[93]. The optimal clustering resolution was chosen to maximize the number of clusters while keeping the number of DE genes between neighboring clusters larger than 0. To

maintain consistency with how the validation datasets were originally analyzed, clusters were annotated using marker gene expressions as indicated in the original publications. The resulting clusters represent cell types or states (see 3.3.5) that are stable over the developmental process, such as apical progenitor cells in murine cerebral cortex development and myoblasts in muscle development.

2.2 Pathway enrichment analysis

Tempora calculates the average gene expression over all cells in a cluster for all clusters as input by the users and determines the pathway enrichment profile of each cluster using Gene Set Variation Analysis (GSVA)[106]. By analyzing scRNAseq data on the cluster level instead of the single-cell level, Tempora amplifies gene expression signals from similar cells in a cluster to alleviate the typical problem of low sensitivity in single cells of popular scRNA-seq experimental methods, as well as reduce the number of nodes in the inferred trajectory, allowing users to interpret it more easily. GSVA calculates the Kolmogorov-Smirnov rank statistics of all gene sets present in a given gene set database and normalizes the enrichment score by calculating the difference between the maximum and minimum deviation from zero of each gene set score. The default gene set database Tempora uses is the Enrichment Map gene set without electronic annotation, filtered to include gene sets between 10 and 500 genes in size according to the recommendation from GSEA[105,107]. The enrichment scores of all G pathways in each cluster make up the cluster's pathway enrichment profile, which is an ID vector of length G.

Since gene set databases have been shown to contain redundant pathways, which share similar names and/or genes[164], strong signals from dependent pathways in the pathway enrichment profiles can mask subtler signals and skew downstream calculation of correlation between clusters. To ameliorate this problem, Tempora uses PCA to reduce redundancy in the clusters' pathway enrichment profiles and identifies the top *n* principal components that explain at least 85% of the variance in the data to input to downstream trajectory construction steps.

2.3 Network construction using filtered mutual information

Tempora employs the mutual information (MI) rank approach implemented in ARACNe[161] to calculate MI between all cluster pairs present in the data. Afterwards, the data-processing inequality is applied to remove the edge with the lowest MI in each triple to eliminate indirect

interactions between clusters. The result is an undirected network where nodes are clusters and edges represent MI strength relationships between clusters.

2.4 Direction identification

Tempora makes use of time information to determine the edge directions in the constructed network. Tempora assigns each time point an ordinal value corresponding to its position in the user-defined sequence of time points and calculate the temporal scores of each cluster by weighing the composition of cells from each timepoint. Specifically, the time score T_k of cluster k consisting of p_i percent cells at timepoint i (0 <i<N) for N timepoints is calculated as the sum of the proportions of cells in the cluster at each time point:

$$T_k = \sum_{i=1}^{N} p_i \cdot i$$

Tempora assigns directions to all edges in the network so that edges originate from clusters with low time scores (early) to clusters with high time scores (late). For edges that connect clusters with similar temporal scores (difference less than 1% of the lowest score), Tempora does not assign directions as these edges may represent small, elastic transitions in cell states over a short time.

2.5 Identification of time-dependent pathways

Tempora identifies pathways that vary over time by fitting a generalized additive model (GAM) on the pathway enrichment scores of each pathway across all clusters/time and using ANOVA to compare the fitted model with the null model of uniform pathway enrichment over time. Pathways with adjusted p-values below a user-defined threshold, with a default value of 0.01, are reported as varying over time. The model fitting and statistical testing are done using the *mcgv* package in R.

3 Validation

3.1 Validation on the human skeletal myoblast dataset

I evaluated Tempora's performance on the human skeletal muscle myoblast (HSMM) data, which includes 271 cells collected at 0, 24, 48 and 72 hours after the switch of human myoblast culture from growth to differentiation media[163]. The muscle myoblast culture is known to contain

contaminating fibroblast cells, which originate from the same muscle biopsy used to establish the primary culture[163,165]. At the optimal clustering resolutions, five clusters were identified and annotated with markers of proliferation (*CDK1*), muscle differentiation (*MYOG*) and contaminating fibroblast cells (*SPHK1*) (Figure 4a-d). I used Tempora to construct a trajectory connecting these clusters and visualized the known marker genes on the trajectory (Figure 5). Tempora identifies a branching trajectory connecting these clusters, rooted at the myoblast cluster that contains mostly cells at 0 hours after the media switch. This cluster leads to three separate branches, including a branch connected to the fibroblast cluster, one connected to the myotube cluster, and the last one connected to the partially differentiated myotube cluster via an intermediate cluster (Figure 5a). This branching trajectory agrees with the known biology of muscle differentiation *in vitro*, in which myoblasts proliferate and exit the cell cycle before differentiating into myotubes[9,166]. The fibroblast cluster contains equal proportions of cells from all time points and uniquely expresses fibroblast markers (*SPHK1*). The equal numbers of cells from all time points in this cluster suggest that the contaminating cells were present in the earliest time point and persist in the culture over time, while its separation from the other two branches suggest that these cells not go through the differentiation process. Thus, Tempora has identified fibroblasts as a source of contamination in the myoblast culture, consistent with results from other trajectory inference methods[148,152] and from the literature[165,167]. Another branch in this trajectory connects the myoblast cluster to the cluster of myotubes, which contains *MYOG*-positive cells mostly at 48 and 72 hours. (Figure 5a). *MYOG* is a required transcription factor for the terminal differentiation of myoblasts into myotubes and is rapidly upregulated when myoblasts start to differentiate around day 2 *in vitro*[168,169]. Therefore, the appearance of *MYOG*-positive myotubes at 48 hours and their connection to the myoblasts cluster, as predicted by Tempora, are aligned with previous findings in the literature. Finally, the myoblast cluster is also connected to an intermediate cluster, which contains 75% cells from two early time points, expresses lower level of *CDK1* and has yet to express *MYOG* (Figure 5a). The low *CDK1* expression suggests that cells in this cluster has begun to exit the cell cycle to start differentiation, thus representing an intermediate state between proliferating myoblasts and differentiated muscles as consistent with our understanding of muscle differentiation[9,166]. This intermediate cluster leads to a partially differentiated cells, which contains mostly cells from later time points and expresses low level of the muscle-specific transcription factor *MYOG*. Since HSMM cultures have been noted to differentiate asynchronously and with less than 100% efficiency, cells in this partially

differentiated cluster represent the population slower to differentiate and/or failed to go through differentiation as observed in previous studies[168,170]. Tempora, thus, predicts a branching trajectory that matches the structure and gene/pathway expression patterns of the known lineage[9].

I used the pathway exploration feature of Tempora to further identify pathways whose enrichment changed over time as well as pathways with cluster-specific enrichment. Pathways enriched early in the differentiation process include the cell cycle, biosynthesis and chromatin remodeling (Figure 5b-d). Pathways upregulated later are associated with the formation of myotubes, which include morphogenesis, extracellular matrix assembly[171] and FGFR signaling, which regulate myogenic activity[172,173] (Figure 5b-d). The pathway exploration component of Tempora, thus, can be used downstream of trajectory construction to identify the cellular activities of specific cell states based on the active pathways in that state, as well as infer important signaling pathways at various stages of a developmental process.

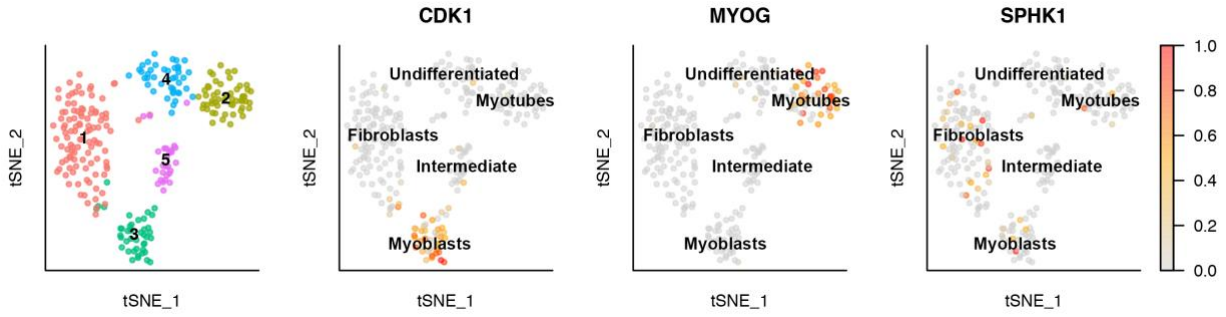

Figure 4. a. tSNE plot showing 271 cells in the HSMM dataset, colored by cluster number. **b-d.** Visualization of known marker genes for **b.** myoblasts, **c.** myotubes and **d.** fibroblasts on the HSMM dataset.

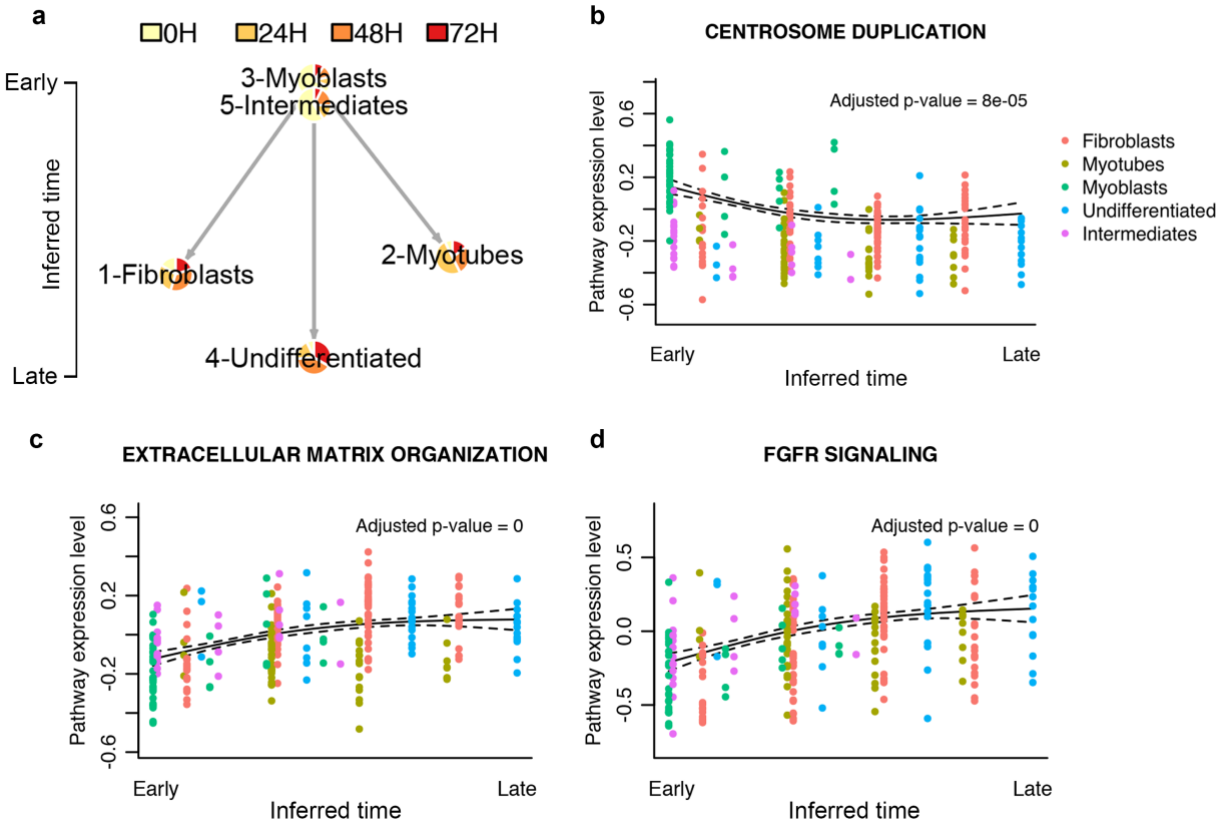

Figure 5. a. Tempora trajectory built on clusters in the HSMM dataset. **b-d.** Time-dependent pathways in the HSMM dataset as discovered by the Tempora pathway exploration feature.

3.2 Validation on the embryonic murine cerebral cortex

I next applied Tempora on the embryonic murine cerebral cortex development scRNAseq data, which contains approximately 6,000 neural cells collected at embryonic day 11.5 (E11.5), E13.5, E15.5 and E17.5[135] (Figure 6a). These cells cover a wide spectrum of neuronal development, from the early precursors (apical precursors (APs) and radial precursors (RPs)) to intermediate progenitors (IPs) and differentiated cortical neurons. After data integration and clustering, I annotated the seven resulting clusters using marker genes for APs (*Sox2*, *Pax6*, *Hes1*, *Mki67*), RPs (*Edrnb*, *Vim*, *Slc1a3*), IPs (*Eomes*, *Gadd45g*, *Mfap4*, *Sstr2*), newborn neurons (*Tbr1*, *Tubb3*, *Foxp2*, *Reln*) and neurons (*Tubb3*, *Bhlhe22*, *Satb2*, *Fezf2*, *Mef2c*, *Gria2*) (Fig 4b-f). This resulted in the annotation of two AP/RP clusters mostly comprising of cells at E11.5, which is consistent with the known emergence of RPs from APs at E11[135,174], as well as two IP clusters, one IP/young neuron cluster and two neuron clusters, all of which contain cells from multiple timepoints as expected from their gradual specification over time[135] (Figure 6b-g).

Tempora predicts three lineages, two rooted at the two AP/RP clusters and one rooted at an early IP cluster (Figure 7a). Each of the two AP/RP lineages has two branches: one terminating at an IP/young neuron cluster and another converging at a late neuron cluster. The lineage predicted by Tempora aligns with our understanding of AP/RP asymmetric division to generate IPs and neurons in early corticogenesis[135,174,175]. To better understand why there are two lineages arising from two AP/RP clusters instead of one AP/RP cluster transforming to another AP/RP cluster in a single lineage, I carried out a DEG analysis between the two clusters and identified cell cycle markers, such as *Mki67* and *Cdk1*, to be differentially expressed in one cluster over the other. This result suggests that the two AP/RP clusters differ based on their cell cycle state: one is actively proliferating and expressing cell cycle markers while the other is not (Figure 6d), consistent with the known decreased proliferation of APs as they transition to RPs[135,176]. The observation that both AP/RP clusters contain equal proportion of cells from all time points suggest that these two proliferative and non-proliferative AP/RP populations arise independently, instead of one transforming into the other. Similarly, the IP cluster that serves at the root of the third lineage contains many cells from the earliest time point and is thus unlikely to come from either of the AP/RP clusters, but may instead arise from earlier APs that are not captured in this time series data set. The transition between early IPs to young neurons to neurons predicted by Tempora is also consistent with our understanding of neurogenesis[135,174]. Tempora, thus, accurately identifies distinct lineages originating from different populations in the time-course murine cerebral cortex data.

I used the pathway exploration feature of Tempora to further analyze time-dependent pathways in the data. As known in the literature, growth and proliferation pathways are enriched early[176] (Figure 7b-d), while neuron-related pathways, such as synapse activity, dendritic morphogenesis and neurotransmitter synthesis, are enriched later[177] (Figure 7d-e). Tempora also identifies more subtle changes in signaling pathways over time, such as the early enrichment of Notch signaling and the later upregulation of Semaphorin-Plexin signaling, both of which are consistent with their known roles in early neurogenesis and neural circuit assembly respectively[178-180] (Figure 7b-e).

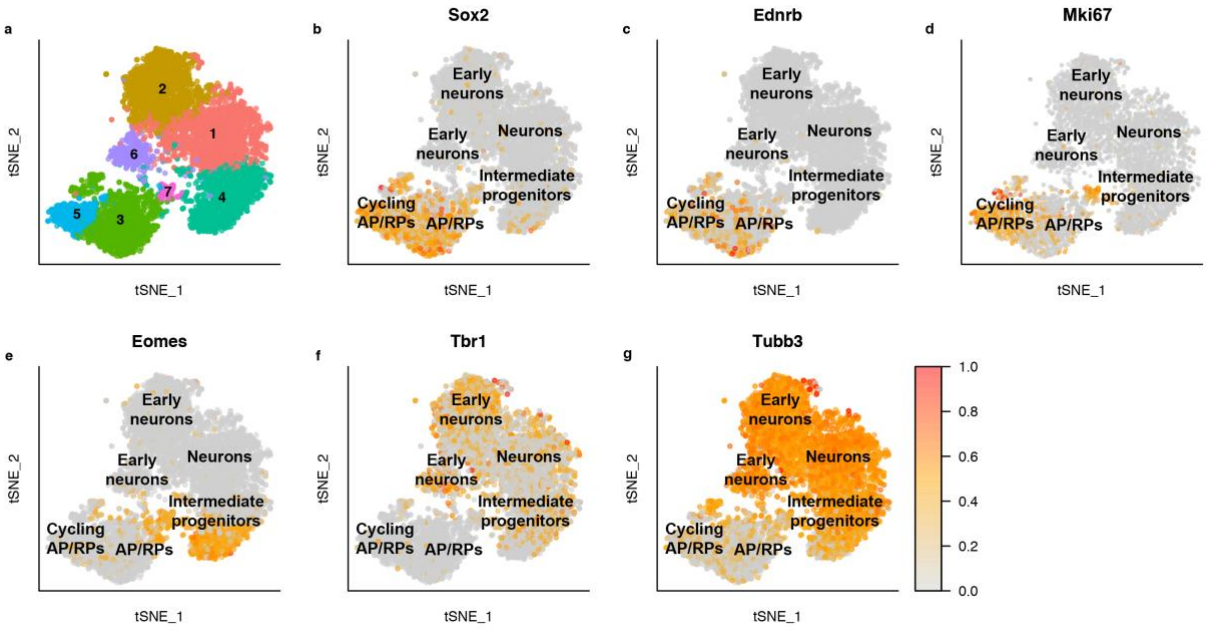

Figure 6. a. tSNE plot showing the ~6,000 neural cells captured in the murine cerebral cortex dataset, colored by cluster number. **b-f.** Visualization of marker genes for **b.** apical precursors (APs), **c.** radial precursors (RPs), **d.** cycling apical/radial precursors (AP/RPs), **e.** intermediate progenitors (IPs), **f.** early neurons and **g.** neurons in the murine cerebral cortex dataset.

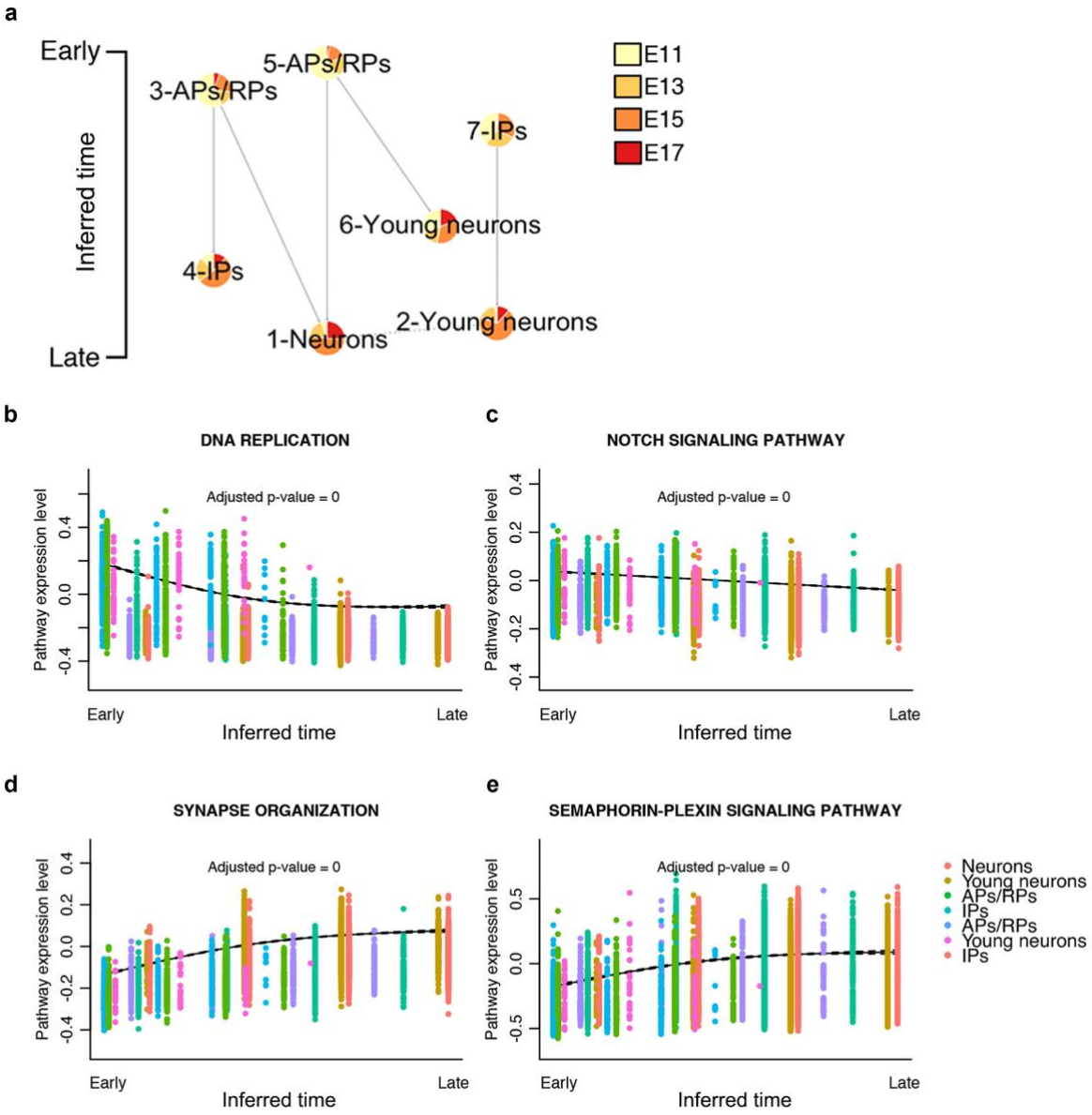

Figure 7. **a.** Tempora trajectory built on clusters in the murine cerebral cortex dataset. **b-e.** Time-dependent pathways in the murine cerebral cortex dataset as Tempora pathway exploration feature.

3.3 Performance evaluation

As the main motivation behind our method is to accurately predict the developmental trajectory underlying scRNAseq data, I used accuracy in recapitulating a gold standard of known lineages as the main criteria to evaluate Tempora's performance and compare it to other TI methods. For ease of comparison, I formalized all trajectories, both predicted and known, as graphs, with nodes representing cell types, and directed edges representing parent-child

relationships between connected nodes. The graph formalization of all trajectories allows me to use standard graph comparison methods to evaluate method performance.

The graph formalization of all trajectories allows me to use graph edit distance (GED), a measurement of mismatch between two graphs, as the first evaluation criteria. GED is calculated as the number of changes necessary (i.e. the number of necessary additions and removals of edges or nodes) to transform our inferred lineage to the known biological lineage. As Tempora inferred lineages include edge directions, the second evaluation criterion is the accuracy of the inferred directions using the F1 score.

3.3.1 Model trajectory construction

I curated the model trajectories for the in vitro differentiation of human myoblasts and murine cortical development through literature search[9,174,175,181] and described the lineage relationships between different cell types in the system using graphs. Each node in a model trajectory represents a distinct cell type as noted in the literature and described in the Cell Ontology[85], while the edges represent lineage connections (*develops_from* relationship in the Cell Ontology) between these cell types.

3.3.2 Mismatch score

I used the unweighted GED metric to measure the number of mismatches between the predicted and known trajectory, both formalized as undirected graphs to allow for comparisons with methods that do not predict edge directions[182]. GED is formally defined as the smallest total number of graph edit operations needed to transform one graph into another. In this context, the permitted operations included insertion and deletion of edges or vertices.

To calculate the mismatch score between a pair of graphs, I first labeled each cluster in the inferred trajectory with the cell type(s) it contains, based on expression of a set of well-known marker genes. The cell types used for labeling are terms from the Cell Ontology database. If multiple clusters contain one cell type, they are assigned the same label. In case one cluster contains multiple cell types as defined by the positive expression of multiple gene sets, I label the cluster based on the major cell type (more than 60% of cells in the cluster are positive for marker gene sets of this cell type) or label the cluster with both cell types if the proportion of cells expressing each set of marker genes is equal. I then calculated the number of differences in the cell

types of the predicted and known trajectories, as well as in the adjacency matrices of both trajectories. The sum of these two differences is the mismatch score for each pair of graph.

3.3.3 F1 score

To compare the accuracy of Tempora's time-based direction inference with the model trajectory, I calculated the F1 score on each predicted trajectory as follows:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{2TP}{2TP+FP+FN}$$

in which true positives (TP) are edges present in both the model and the predicted trajectory, false positive (FP) are edges in the predicted trajectory but not in the model, and false negatives (FN) are edges in the model but not in the predicted trajectory. An edge in the predicted graph is considered true positive only when its two vertices and direction match those of an edge in the model graph.

3.3.4 Performance evaluation on the HSMM dataset

Human myoblasts, after exiting the cell cycle, transition through intermediate states before differentiating into myotubes[9,166]. Since myoblasts have varied differentiating potentials and rates, a portion of them will become myotubes while the rest remain undifferentiated, i.e. they do not, or have yet to, express myogenic transcription factors such as MYOG, which leads to two possible branches from the intermediate state(s)[168-170]. The starting culture, however, is often contaminated with fibroblasts cells, which exert paracrine influence on the differentiation process but cannot differentiate into myotubes[167,172]. These contaminating cells, thus, form a branch separate from the main differentiation trajectory (Figure 8d).

Tempora's predicted trajectory (Figure 8a) is closely aligned with the model trajectory, except for the edge connecting the myotubes cluster to the myoblasts instead of to the intermediate state. This results in a mismatch score of 2, which means that only two edges need to be changed in Tempora's output to match the gold standard (Figure 8e). Since the myoblast and intermediate clusters are quite similar in their composition, it is possible that their pathway enrichment profiles share many features in common and thus the myoblast-myotube connection can be favored over the accurate intermediate-myotube connection only by a marginal difference in MI. Furthermore, Tempora achieves a high F1 score of 0.78 as it is able to infer the correct directions of most edges

in the trajectory, saved for the missing intermediate state to myotubes connection (Figure 8f). This result demonstrates that Tempora is able to infer a trajectory in the HSMM dataset that is mostly consistent with the gold standard, with minor mistakes that can be justified biologically.

3.3.5 Performance evaluation on the murine cerebral cortex dataset

Murine corticogenesis consists of transitions between well-characterized cell types. The apical precursors (APs), which delaminate from the neuroepithelium, dive asymmetrically to give rise to neurons and self-renew[174]. At around E11, APs transition to radial precursors (RPs), which continue the asymmetric division to generate neurons either directly or indirectly through IPs[174,175] (Figure 9d). Tempora's inferred trajectory of the murine cerebral cortex dataset achieved a low mismatch score and high F1 score. It predicts almost all possible transitions between different cell types in the systems, only missing the IPs to neuron connection, which results in a mismatch score of 1 (Figure 9e). Despite this mistake, Tempora achieves a perfect F1 score of 1 on the murine cerebral cortex dataset, demonstrating that it can accurately and robustly identify directed connections between cell types in a large data set with multiple branching trajectories (Figure 9f).

3.4 Comparison with other trajectory inference methods

3.4.1 Comparison methods

I compared Tempora's performance with Monocle 2[157] and TSCAN[152], two popular TI methods that have performed well in multiple studies[147,153]. Similar to previous performance evaluations, I formalized all predicted trajectories from Monocle 2 and TSCAN as graphs and compared them to the same reference trajectories discussed in 3.3.1 in order to calculate each method's mismatch score and F1 score. Since both Monocle 2 and TSCAN do not predict directions of the edges, I also considered the reference trajectory undirected for mismatch score calculation. To calculate the F1 score on undirected trajectories inferred by Monocle 2 and TSCAN, I first determined the origin of the trajectories based on high expression of a set of known marker genes (*CDK1*, *CCND5* for myoblasts in the HSMM dataset[9,166] and *Sox2*, *Pax5* for APs in the murine cerebral cortex dataset[135,176]). I then added directions to the inferred trajectories by directing all edges to go outward from the origin.

3.4.2 Monocle 2

I applied Monocle 2 on the HSMM and murine cerebral cortex datasets using the recommended pipeline[157]. Genes used for the pseudotime inference process were determined using the dpFeature procedure, in which cells were first clustered and highly differentially expressed genes across the clusters were selected for downstream analyses. The trajectory in the data was then inferred using reversed graph embedding, a machine learning algorithm that learns the tree-structured low-dimensional space on which the data points lie. To formalize a Monocle trajectory as a graph, I considered each state, or segment of the tree, as a vertex, and connected the vertices with appropriate edges to recapitulate Monocle's output.

3.4.3 TSCAN

I applied TSCAN on two datasets in this study using the Shiny GUI, which allowed me to make use of additional marker gene visualization features not available with command line[152]. As TSCAN clusters each dataset and constructs an MST on the clusters, its output lends itself nicely to the graph formalization. I optimized the number of clusters used in trajectory construction using TSCAN's built in optimization feature. I then retained the cluster-level MST that TSCAN outputted for each dataset and considered each cluster a vertex, while the segments of the MST are edges in the formalized graphs. I then determined the roots and directions of the graph as described above.

3.4.4 Comparison of performance on the HSMM dataset

When applied to the HSMM data, Tempora outperformed both Monocle 2 and TSCAN as evaluated by the mismatch and F1 scores described above (Figure 8). Monocle 2 predicted a trajectory with four states, annotated as myoblasts, partially differentiated, mesenchymal and differentiated myotubes respectively. The trajectory branches at the end of the myoblast state into the other three states without a clear intermediate state (Figure 8b). This lack of intermediate state leads to Monocle's worse mismatch score of 6. TSCAN's trajectory has a linear structure, with the myoblasts at the root (state 1, annotated on the TSCAN plot) progressing through an intermediate state (state 3) and terminating at a cluster with mixed differentiated and undifferentiated cells (state 2). TSCAN's trajectory is penalized because it neither separates the mesenchymal cells from the undifferentiated cells nor undifferentiated cells from differentiated myotubes at the terminal states, increasing its mismatch score to 8 (Figure 8c). Overall, even though all methods predict similar

branching trajectories, Tempora performs best in terms of mismatch score as it correctly identifies the expected cell states and their directions in the HSMM data (Figure 8e). Similarly, with a F1 score of 1, Tempora outperformed Monocle 2 (F1 of 0.3) and TSCAN (F1 of 0) (Figure 8f). These results suggest that Tempora is able to accurately infer a trajectory in the HSMM data that aligns well with the gold standard trajectory and outperforms two leading TI methods on the same dataset.

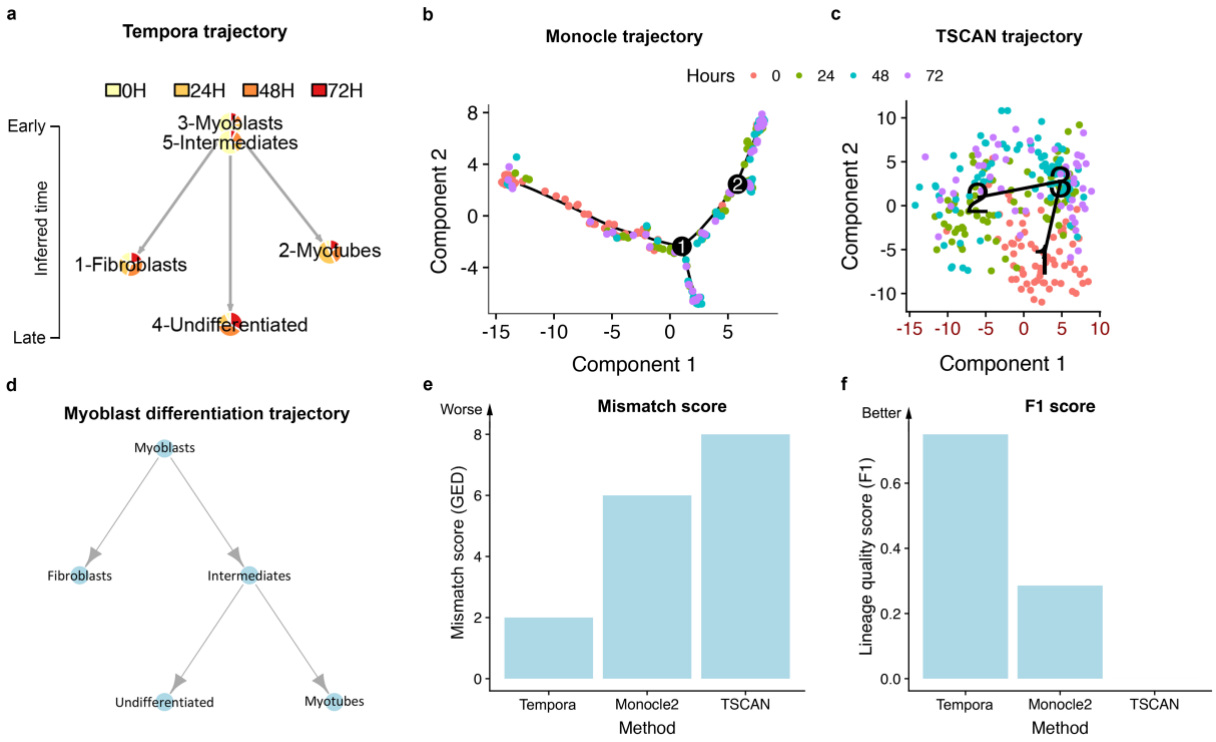

Figure 8. Performance evaluation on the HSMM dataset. **a-c.** Trajectories of the HSMM dataset inferred by **a.** Tempora, **b.** Monocle 2 and **c.** TSCAN. **d.** The model trajectory used to evaluate the accuracy of all inferred trajectories. **e.** Mismatch scores and **f.** F1 scores of trajectories from the three evaluated methods.

3.4.5 Comparison of performance on the murine cerebral cortex dataset

Tempora outperforms Monocle 2 and TSCAN on the murine cerebral cortex data set, which is larger and contains more transitions than the HSMM dataset. Monocle 2 infers a branched trajectory at two main branches: one from an APs/RPs branch to two IP branches (branchpoint 2) and one from the larger IP branch to two neuron branches (branchpoint 1) (Figure 9b). The early neurons are merged in both of the neuron branches instead of identified as a distinct state. Monocle 2's trajectory is thus penalized for this lack of young neuron state as well as its inability to predict the direct differentiation from APs/RPs to neurons, thus achieving a mismatch score of 5. TSCAN

predicts a linear trajectory that connects APs/RPs to IPs, then to two neuron clusters (Figure 9c). TSCAN is penalized because it forces an erroneous connection between two neuron clusters, and similar to Monocle 2, it does not recognize a separate young neuron state and the direct differentiation link between APs/RPs to neurons. This results in a higher mismatch score of 7. With a mismatch score of 1, Tempora's performance exceeds that of Monocle 2 and TSCAN by at least five times (Figure 9e). To calculate F1 scores on trajectories from these methods, I used *Sox2* expression to infer that both Monocle 2 and TSCAN trajectories were rooted where all E11.5 cells are collected, and determined that all edges are going outward from this root. The inferred directions of both trajectories are consistent with the gold standard trajectory. With a F1 score of 1, Tempora significantly outperforms Monocle 2 (F1 of 0.5) and TSCAN (F1 of 0.25) on the murine cerebral cortex dataset (Figure 9f).

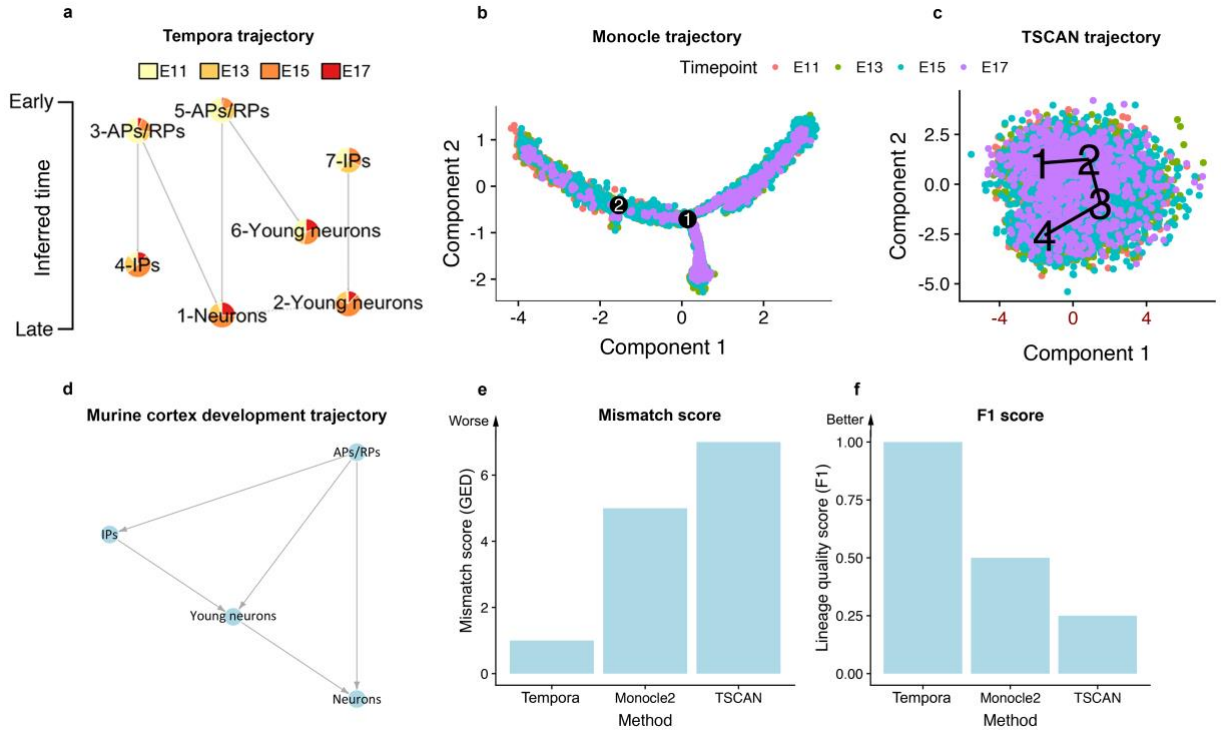

Figure 9. Performance evaluation on the murine cerebral cortex dataset. **a-c.** Trajectories of the murine cerebral cortex dataset inferred by **a.** Tempora, **b.** Monocle 2 and **c.** TSCAN. **d.** The model trajectory used to evaluate the accuracy of all inferred trajectories. **e.** Mismatch scores and **f.** F1 scores of trajectories from the three evaluated methods.

3.4.6 Comparison of Tempora performance with and without pathway enrichment analysis

To understand the impact of pathway enrichment information on lineage construction compared to gene expression inputs typically used by other methods, I compared trajectories in the HSMM and murine cerebral cortex datasets using Tempora with and without the pathway enrichment analysis (PEA) step. Removing the PEA step resulted in poorer performance, as evident in up to 4-fold increase in mismatch score and 3-fold decrease in F1 scores (Figure 10).

Upon closer examination of the trajectories, I observed that gene-input trajectories contain more edges between clusters with similar temporal scores compared to pathway-input trajectories, whose edges often connect clusters from different time points. I propose that this trend can be explained by the high similarity in gene expression profiles of clusters that are closer in developmental time, a fundamental assumption made by TI methods that rely on distance metrics to order cells. To test this hypothesis and better understand the discrepancies in inter-cluster gene vs. pathway enrichment profile similarity, I calculated the Pearson's correlation between the gene and pathway enrichment profiles of all pairs of clusters in each dataset. I found striking differences in the dynamic range of correlation observed: while correlations between gene expression profiles are uniformly strong and positive across all pairs of clusters, correlations between pathway enrichment profiles are negative for clusters of different cell types (neurons vs. APs, myoblasts vs. fibroblasts) and positive for clusters of the same cell types (neurons vs. neurons) (Figure 11a-b). These differences suggest that the highly similar gene expression profiles across clusters make them less informative than pathway enrichment profiles in capturing changes along a lineage, which results in the poorer performance of Tempora without the PEA step.

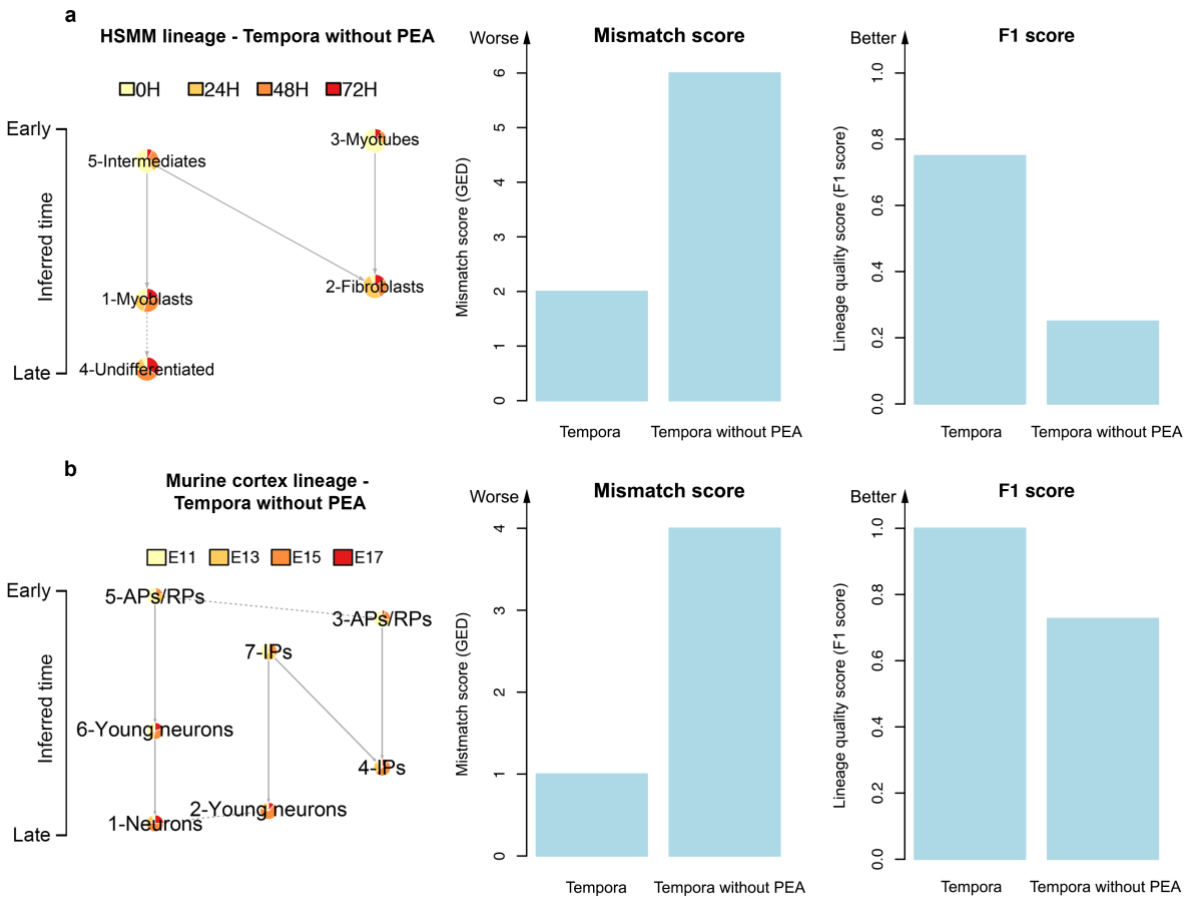

Figure 10. Performance of Tempora on **a.** HSMM and **b.** murine cerebral cortex dataset with and without pathway enrichment analysis (PEA).

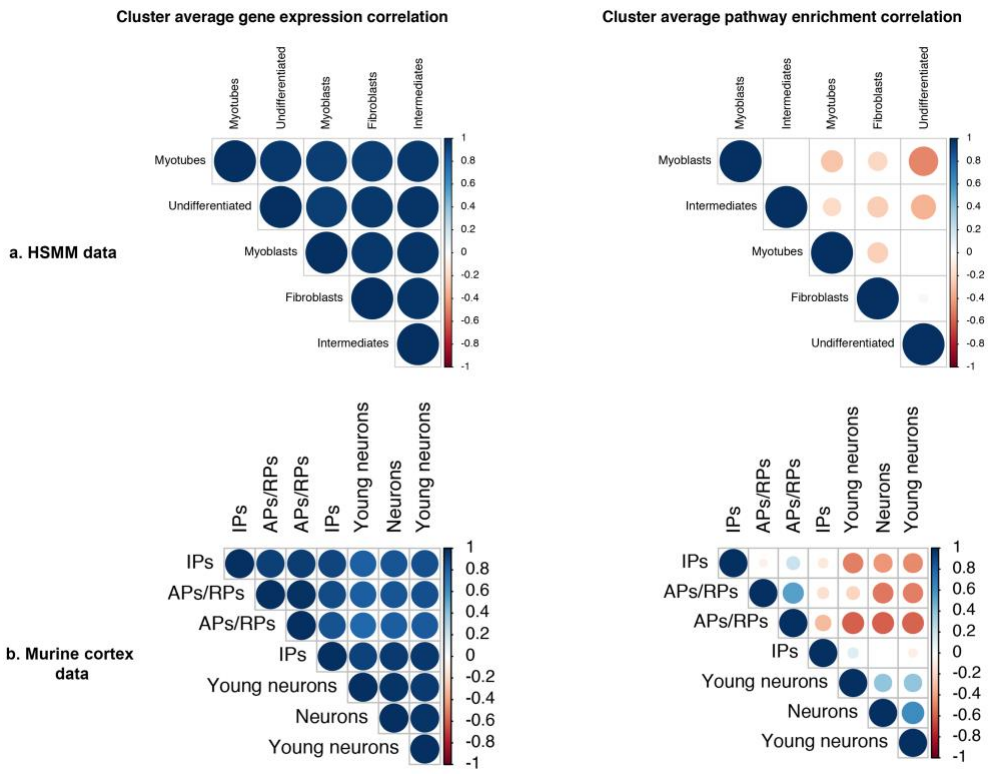

Figure 11. Correlation plots showing cluster-average gene expression and pathway enrichment profiles in **a.** HSMM and **b.** murine cerebral cortex data.

3.4.7 Comparison of Tempora performance with and without scRNA-seq data alignment

An important part of Tempora's pipeline, designed specifically to analyze time-series scRNAseq data, is batch effect assessment and correction, as time-series data are often collected and sequenced in batches, thus easily subjected to technical variations between experimental runs. I implemented two existing tools for this purpose: kBET to assess batch effect and Harmony to correct for any detected effect. Without such correction, Tempora's performance decreased

slightly on both gold standards (

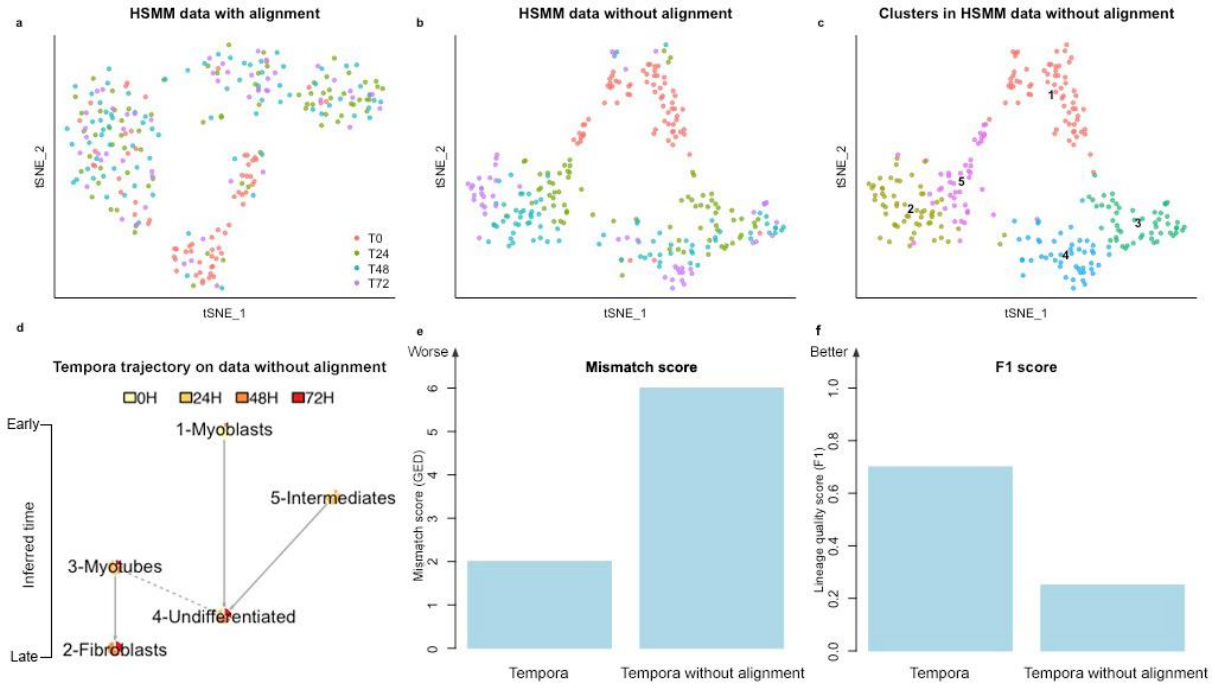

Figure 12, Figure 13). This is likely due to the suboptimal clustering driven by batch, which results in less accurate inference of trajectories based on these clusters. Therefore, when kBET indicates that there is batch effect, it is recommended that Harmony be run to ensure that clusters represent true cell types that are present across time points instead of groups with distinct technical variances. However, since kBET's runtime can be significant for large data sets, batch effect can also be qualitatively assessed by initially clustering all cells in a time-course experiment and evaluate whether clusters are clearly separated by batches.

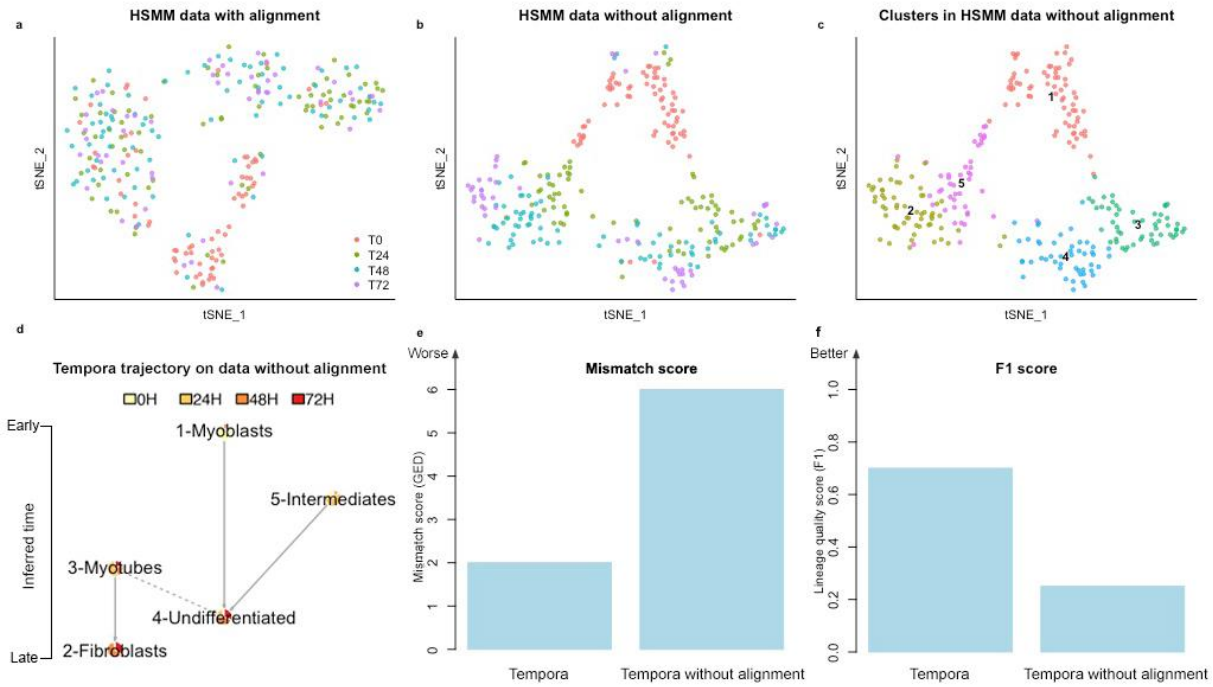

Figure 12. **a-b.** tSNE plots of HSMM data **a.** with and **b.** without Harmony alignment, with cells colored by timepoints. **c.** tSNE plot of clusters in HSMM data without alignment. **d.** Tempora trajectory and **e.** performance evaluation of Tempora on HSMM data without alignment.

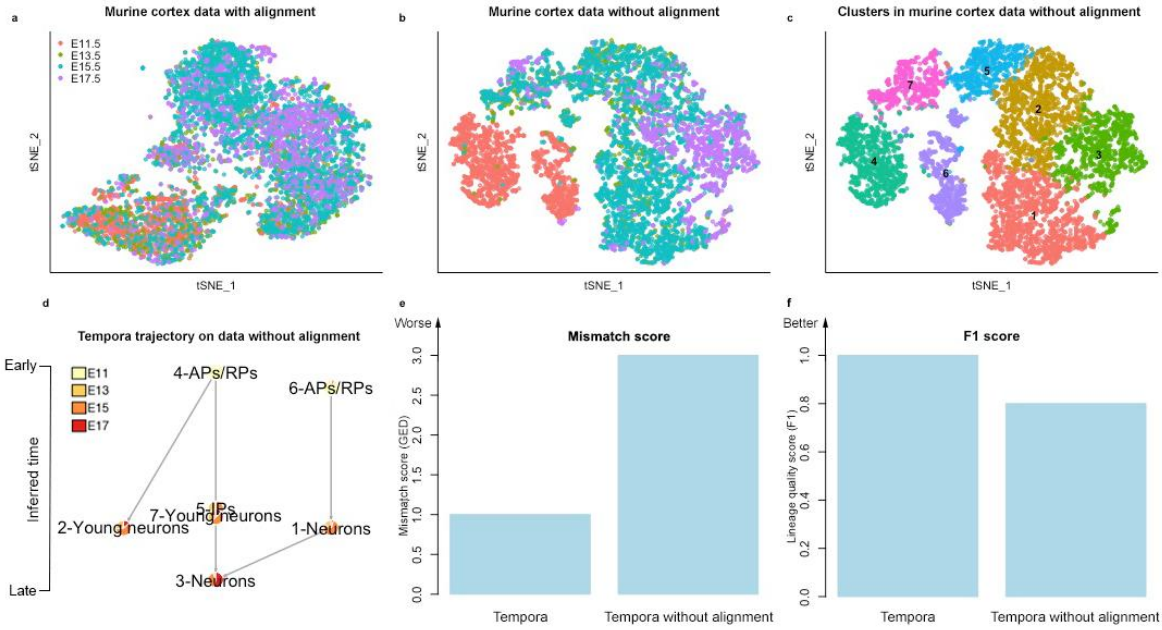

Figure 13. a-b. tSNE plots of murine cerebral cortex data **a.** with and **b.** without Harmony alignment, with cells colored by timepoints**. c.** tSNE plot of clusters in murine cerebral cortex data without alignment. **d.** Tempora trajectory and **e.** performance evaluation of Tempora on murine cerebral cortex data without alignment.

Chapter 3
Conclusion

I have described and validated Tempora, a novel pathway-based lineage analysis method for time-series scRNAseq data. Tempora uses an information theoretic approach to build a trajectory at the cluster level based on the clusters' pathway enrichment profiles, effectively connecting cell types and states across multiple time points. Taking advantage of the available time information, Tempora can infer the directions of all connections in a trajectory that go from early to late clusters. Validation on two time series scRNAseq datasets with known developmental trajectories (one on the *in vitro* differentiation of human skeletal muscle myoblasts and the other on an *in vivo* sample of early development of murine cerebral cortex) demonstrate that Tempora can accurately predict the lineages in time series data containing cell populations spanning all developmental stages. I showed that Tempora outperformed other state of the art TI methods on the same datasets when evaluated against a gold standard benchmark of known trajectories, using metrics for graph prediction accuracy including mismatch score and F1 score. Furthermore, downstream analyses using the pathway exploration feature of Tempora have identified signaling pathways known to play regulatory roles during the process under study, demonstrating the method's ability to recapitulate and discover important signals during development processes. Even though Tempora is only validated on two datasets, these datasets were selected based on the extrinsic availability of gold standard trajectories for easy comparison and not on any intrinsic features of the data or the systems under study. Tempora, thus, should be generalizable to scRNAseq datasets on other developmental systems beyond on the skeletal muscle and the brain.

Despite Tempora's proven performance, some components of the algorithm can become potential drawbacks. Cell clustering plays an integral role in Tempora's algorithm, which makes the resulting trajectory more stable and interpretable compared to using cells directly, but this comes with its confounding issues. Tempora assumes that the user has input an optimized clustering solution for their biological question into the method for trajectory construction. If the clustering is not optimal, the output trajectory may not be useful in answering the users' biological question of interest as it may present a view of a lineage that is either too general or too detailed. Over-clustering, when the resulting clusters are arbitrarily split without unique marker genes, can lead to parallel edges originating from oversplit clusters and terminating at another cluster, ostensibly suggesting a multiple-parent lineage (Figure 14). Meanwhile, under-clustering can

result in simplified lineages (Figure 14). Under-clustering of certain clusters can also lead to certain cell types appearing at earlier and later time points but absent from the intermediate timepoints, potentially because they have been clustered with other cell types at the intermediate states. As previously discussed in Chapter 1, these challenges are inevitable when clustering high-dimensional data, but researchers can ameliorate these difficulties by iterating through multiple clustering resolutions and determining an optimal one based on biological knowledge and a consistent set of user-defined rules and interests[93,94].

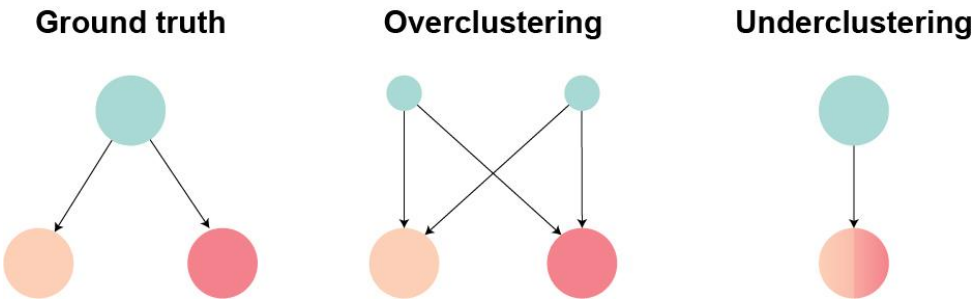

Figure 14. The effects of sub-optimal clustering resolution choice on trajectory inference. Overclustering (middle) can lead to complex lineages with converging connections, while underclustering (right) can lead to oversimplified lineages.

I reformulated the problem of building a trajectory between clusters as that of learning a graphical model that describes the statistical dependencies among these clusters. This formulation enabled us to take advantage of a family of structure learning algorithms to learn a model underlying the data under study. Tempora uses a filtered mutual information network (built using ARACNe[161]) to build the trajectory between cell clusters in a dataset because of the algorithm's proven robustness and low computational complexity, which scales well with large scRNAseq datasets[161]. However, many other algorithms in the same family, including constraint-based and search-based algorithms, can theoretically solve the same problem laid out here[183]. Future extensions to this work can include the evaluation and optimization of other structure learning algorithms in the Tempora framework, especially those developed to work with temporal data.

The increasing popularity and complexity of time-series scRNAseq to investigate dynamic biological processes, including development and differentiation, present both opportunities and challenges. The larger cell numbers and types captured in time-course experiments allow researchers to discover rare cell types and study cell transitions with higher resolution, yet the non-synchronous and uncorrelated nature of the populations across time points present a computational

challenge to characterize their trajectories[184]. Using time information to supervise the trajectory inference process, as Tempora does, enables accurate identification of cell types consisting of cells from different time points as well as the lineage connections between them. When combined with other methods to infer population and transcriptional dynamics to analyze time-series scRNAseq data[185,186], Tempora can generate powerful insights into different dynamic processes and their biological regulation.

References

1 Koshland, D. E. The Seven Pillars of Life. *Science* **295**, 2215-2216, doi:10.1126/science.1068489 (2002).

2 Perbal, L. The case of the gene: Postgenomics between modernity and postmodernity. *EMBO reports* **16**, 777-781, doi:10.15252/embr.201540179 (2015).

3 Van Regenmortel, M. H. V. Reductionism and complexity in molecular biology. *EMBO reports* **5**, 1016-1020, doi:10.1038/sj.embor.7400284 (2004).

4 Robinson, J. D. Aims and Achievements of the Reductionist Approach in Biochemistry/Molecular Biology/Cell Biology: A Response to Kincaid. *Philosophy of Science* **59**, 465-470 (1992).

5 Zhu, X., Gerstein, M. & Snyder, M. Getting connected: analysis and principles of biological networks. *Genes & Development* **21**, 1010-1024, doi:10.1101/gad.1528707 (2007).

6 Sam, S. A., Teel, J., Tegge, A. N., Bharadwaj, A. & Murali, T. M. XTalkDB: a database of signaling pathway crosstalk. *Nucleic Acids Research* **45**, D432-D439, doi:10.1093/nar/gkw1037 (2016).

7 Kirschner, M. W. The Meaning of Systems Biology. *Cell* **121**, 503-504, doi:10.1016/j.cell.2005.05.005 (2005).

8 Alberts, B. *et al. Chapter 21: Development of Multicellular Organisms. Molecular biology of the cell*. 4th edn, (Wiley Online Library, 2003).

9 Chal, J. & Pourquié, O. Making muscle: skeletal myogenesis in vivo and in vitro. *Development* **144**, 2104-2122, doi:10.1242/dev.151035 (2017).

10 Endo, T. Molecular mechanisms of skeletal muscle development, regeneration, and osteogenic conversion. *Bone* **80**, 2-13, doi:https://doi.org/10.1016/j.bone.2015.02.028 (2015).

11 Buckingham, M. *et al.* The formation of skeletal muscle: from somite to limb. *J Anat* **202**, 59-68, doi:10.1046/j.1469-7580.2003.00139.x (2003).

12 Li, X. *et al.* Temporal patterning of Drosophila medulla neuroblasts controls neural fates. *Nature* **498**, 456-462 (2013).

13 Erclik, T. *et al.* Integration of temporal and spatial patterning generates neural diversity. *Nature* **541**, 365-370, doi:10.1038/nature20794 (2017).

14 Lee, W. J. *et al.* An Integrative Developmental Genomics and Systems Biology Approach to Identify an In Vivo Sox Trio-Mediated Gene Regulatory Network in Murine Embryos. *BioMed Research International* **2017**, 16, doi:10.1155/2017/8932583 (2017).

15 Ihry, R. J. *et al.* Genome-Scale CRISPR Screens Identify Human Pluripotency-Specific Genes. *Cell Reports* **27**, 616-630.e616, doi:10.1016/j.celrep.2019.03.043 (2019).

16 Du, Z., Santella, A., He, F., Tiongson, M. & Bao, Z. De Novo Inference of Systems-Level Mechanistic Models of Development from Live-Imaging-Based Phenotype Analysis. *Cell* **156**, 359-372, doi:10.1016/j.cell.2013.11.046 (2014).

17 Sharma, S. & Petsalaki, E. Application of CRISPR-Cas9 Based Genome-Wide Screening Approaches to Study Cellular Signalling Mechanisms. *Int J Mol Sci* **19**, 933, doi:10.3390/ijms19040933 (2018).

18 Saunders, T. E. & Ingham, P. W. Open questions: how to get developmental biology into shape? *BMC Biology* **17**, 17, doi:10.1186/s12915-019-0636-6 (2019).

19 Takei, Y., Shah, S., Harvey, S., Qi, L. S. & Cai, L. Multiplexed Dynamic Imaging of Genomic Loci by Combined CRISPR Imaging and DNA Sequential FISH. *Biophysical Journal* **112**, 1773-1776, doi:https://doi.org/10.1016/j.bpj.2017.03.024 (2017).

20 Rompolas, P. *et al.* Live imaging of stem cell and progeny behaviour in physiological hair-follicle regeneration. *Nature* **487**, 496, doi:10.1038/nature11218 (2012).

21 Bar-Joseph, Z., Gitter, A. & Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics* **13**, 552, doi:10.1038/nrg3244 (2012).

22 Griffiths, J. A., Scialdone, A. & Marioni, J. C. Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular systems biology* **14**, e8046 (2018).

23 Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).

24 Stranneheim, H. & Lundeberg, J. Stepping stones in DNA sequencing. *Biotechnol J* **7**, 1063-1073, doi:10.1002/biot.201200153 (2012).

25 Liu, L. *et al.* Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* **2012**, 11, doi:10.1155/2012/251364 (2012).

26 Qi, J., Luo, H. & Hao, B. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic acids research* **32**, W45-W47 (2004).

27 Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* **101**, 5-22 (2017).

28 Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353 (2012).

29 Wang, K. *et al.* Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature genetics* **46**, 573 (2014).

30 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57-63, doi:10.1038/nrg2484 (2009).

31 Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harbor protocols* **2015**, 951-969, doi:10.1101/pdb.top084970 (2015).

32 Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**, 13, doi:10.1186/s13059-016-0881-8 (2016).

33 Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods* **8**, 469 (2011).

34 Oshlack, A., Robinson, M. D. & Young, M. D. From RNA-seq reads to differential expression results. *Genome biology* **11**, 220 (2010).

35 Consortium, E. P. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799 (2007).

36 Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology* **19**, A68 (2015).

37 International Cancer Genome, C. *et al.* International network of cancer genome projects. *Nature* **464**, 993-998, doi:10.1038/nature08987 (2010).

38 Clough, E. & Barrett, T. The Gene Expression Omnibus Database. *Methods Mol Biol* **1418**, 93-110, doi:10.1007/978-1-4939-3578-9_5 (2016).

39 Harrison, P. W. *et al.* The European Nucleotide Archive in 2018. *Nucleic Acids Research* **47**, D84-D88, doi:10.1093/nar/gky1078 (2018).

40 Gillies, A. R. & Lieber, R. L. Structure and function of the skeletal muscle extracellular matrix. *Muscle Nerve* **44**, 318-331, doi:10.1002/mus.22094 (2011).

41 Schiaffino, S. & Reggiani, C. Fiber Types in Mammalian Skeletal Muscles. *Physiological Reviews* **91**, 1447-1531, doi:10.1152/physrev.00031.2010 (2011).

42 Altschuler, S. J. & Wu, L. F. Cellular Heterogeneity: Do Differences Make a Difference? *Cell* **141**, 559-563, doi:10.1016/j.cell.2010.04.033 (2010).

43 Ballios, B. G., Clarke, L., Coles, B. L. K., Shoichet, M. S. & Van Der Kooy, D. The adult retinal stem cell is a rare cell in the ciliary epithelium whose progeny can differentiate into photoreceptors. *Biology Open* **1**, 237-246, doi:10.1242/bio.2012027 (2012).

44 Wagers, A. J. & Weissman, I. L. Plasticity of Adult Stem Cells. *Cell* **116**, 639-648, doi:10.1016/S0092-8674(04)00208-9 (2004).

45 Heymann, D. & Téllez-Gabriel, M. in *Single Cell Biomedicine* (eds Jianqin Gu & Xiangdong Wang) 45-58 (Springer Singapore, 2018).

46 Brady, G., Barbara, M. & Iscove, N. N. Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. *Methods Mol Cell Biol* **2**, 17-25 (1990).

47 Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377, doi:10.1038/nmeth.1315 (2009).

48 Kolodziejczyk, Aleksandra A., Kim, J. K., Svensson, V., Marioni, John C. & Teichmann, Sarah A. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell* **58**, 610-620, doi:https://doi.org/10.1016/j.molcel.2015.04.005 (2015).

49 Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **343**, 776-779, doi:10.1126/science.1247651 (2014).

50 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols* **9**, 171 (2014).

51 Gong, H., Do, D. & Ramakrishnan, R. in *Gene Expression Analysis: Methods and Protocols* (eds Nalini Raghavachari & Natàlia Garcia-Reyero) 193-207 (Springer New York, 2018).

52 Macosko, Evan Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

53 Zilionis, R. *et al.* Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols* **12**, 44, doi:10.1038/nprot.2016.154 (2016).

54 Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049, doi:10.1038/ncomms14049 (2017).

55 Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine* **50**, 96, doi:10.1038/s12276-018-0071-8 (2018).

56 Picelli, S. Single-cell RNA-sequencing: The future of genome biology is now. *RNA Biol* **14**, 637-650, doi:10.1080/15476286.2016.1201618 (2017).

57 Klein, Allon M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187-1201, doi:https://doi.org/10.1016/j.cell.2015.04.044 (2015).

58 Ziegenhain, C. *et al.* Comparative analysis of single-cell RNA sequencing methods. *Molecular cell* **65**, 631-643. e634 (2017).

59 Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9**, 72, doi:10.1038/nmeth.1778 (2011).

60 Baran-Gale, J., Chandra, T. & Kirschner, K. Experimental design for single-cell RNA sequencing. *Briefings in Functional Genomics* **17**, 233-239, doi:10.1093/bfgp/elx035 (2017).

61 10X Genomics. Chromium Single Cell 3' Reagent Kits v2 User Guide. (2018).

62 Saliba, A.-E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research* **42**, 8845-8860, doi:10.1093/nar/gku555 (2014).

63 10X Genomics. *Q&A: What fraction of mRNA transcripts are captured per cell?* , 2019).

64 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411, doi:10.1038/nbt.4096 (2018).

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

66 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 15 (2018).

67 Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A Systematic Evaluation of Single Cell RNA-Seq Analysis Pipelines: Library preparation and normalisation methods have the biggest impact on the performance of scRNA-seq studies. *bioRxiv*, 583013, doi:10.1101/583013 (2019).

68 AlJanahi, A. A., Danielsen, M. & Dunbar, C. E. An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Mol Ther Methods Clin Dev* **10**, 189-196, doi:10.1016/j.omtm.2018.07.003 (2018).

69 MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nature Communications* **9**, 4383, doi:10.1038/s41467-018-06318-7 (2018).

70 Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome biology* **17**, 29-29, doi:10.1186/s13059-016-0888-1 (2016).

71 Wen, L. & Tang, F. Boosting the power of single-cell analysis. *Nature Biotechnology* **36**, 408, doi:10.1038/nbt.4131 (2018).

72 Goh, W. W. B., Wang, W. & Wong, L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology* **35**, 498-507, doi:https://doi.org/10.1016/j.tibtech.2017.02.012 (2017).

73 Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology* **36**, 421, doi:10.1038/nbt.4091 (2018).

74 Korsunsky, I. *et al.* Fast, sensitive, and accurate integration of single cell data with Harmony. *bioRxiv*, 461954, doi:10.1101/461954 (2018).

75 Chen, G., Ning, B. & Shi, T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Frontiers in Genetics* **10**, doi:10.3389/fgene.2019.00317 (2019).

76 Jolliffe, I. *Principal component analysis*. (Springer, 2011).

77 Amir, E.-a. D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology* **31**, 545 (2013).

78 Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9**, 2579-2605 (2008).

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

80 LeCun, Y., Cortes, C. & Burges, C. J. C. (2018).

81 Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature methods* **14**, 565-571, doi:10.1038/nmeth.4292 (2017).

82 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).

83 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**, R25, doi:10.1186/gb-2010-11-3-r25 (2010).

84 L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology* **17**, 75, doi:10.1186/s13059-016-0947-7 (2016).

85 Bard, J., Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome biology* **6**, R21-R21, doi:10.1186/gb-2005-6-2-r21 (2005).

86 Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Research* **25**, 1491-1498, doi:10.1101/gr.190595.115 (2015).

87 Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* **20**, 273-282, doi:10.1038/s41576-018-0088-9 (2019).

88 Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods* **14**, 483, doi:10.1038/nmeth.4236 (2017).

89 Grün, D. *et al.* De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell* **19**, 266-277, doi:https://doi.org/10.1016/j.stem.2016.05.010 (2016).

90 Chen, J., Schlitzer, A., Chakarov, S., Ginhoux, F. & Poidinger, M. Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nature communications* **7**, 11988 (2016).

91 Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142 (2015).

92 Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics (Oxford, England)* **31**, 1974-1980, doi:10.1093/bioinformatics/btv088 (2015).

93 Innes, B. & Bader, G. scClustViz - Single-cell RNAseq cluster assessment and visualization *F1000Research* **7**, doi:10.12688/f1000research.16198.2 (2019).

94 Schwartz, G. W. *et al.* TooManyCells identifies and visualizes relationships of single-cell clades. *bioRxiv*, 519660, doi:10.1101/519660 (2019).

95 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

96 Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature Methods* **11**, 740, doi:10.1038/nmeth.2967 (2014).

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

98 Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* **20**, 40, doi:10.1186/s12859-019-2599-6 (2019).

99 Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods* **15**, 255, doi:10.1038/nmeth.4612 (2018).

100 Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols* **14**, 482-517, doi:10.1038/s41596-018-0103-9 (2019).

101 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25 (2000).

102 Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic acids research* **42**, D472-D477 (2013).

103 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30 (2000).

104 García-Campos, M. A., Espinal-Enríquez, J. & Hernández-Lemus, E. Pathway Analysis: State of the Art. *Front Physiol* **6**, 383-383, doi:10.3389/fphys.2015.00383 (2015).

105 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545-15550 (2005).

106 Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics* **14**, 7 (2013).

107 Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *Plos One* **5**, e13984, doi:10.1371/journal.pone.0013984 (2010).

108 Kucera, M., Isserlin, R., Arkhangorodsky, A. & Bader, G. AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations [version 1; referees: 2 approved]. *F1000Research* **5**, doi:10.12688/f1000research.9090.1 (2016).

109 Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* **9**, 75, doi:10.1186/s13073-017-0467-4 (2017).

110 Regev, A. *et al.* The Human Cell Atlas. *bioRxiv*, 121202, doi:10.1101/121202 (2017).
111 Ji, Q. *et al.* Single-cell RNA-seq analysis reveals the progression of human osteoarthritis. *Annals of the Rheumatic Diseases* **78**, 100-110, doi:10.1136/annrheumdis-2017-212863 (2019).
112 Zanini, F. *et al.* Virus-inclusive single-cell RNA sequencing reveals the molecular signature of progression to severe dengue. *Proceedings of the National Academy of Sciences* **115**, E12363-E12369, doi:10.1073/pnas.1813819115 (2018).
113 Steuerman, Y. *et al.* Dissection of influenza infection in vivo by single-cell RNA sequencing. *Cell systems* **6**, 679-691. e674 (2018).
114 Wallrapp, A. *et al.* The neuropeptide NMU amplifies ILC2-driven allergic lung inflammation. *Nature* **549**, 351 (2017).
115 Kim, C. *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **173**, 879-893.e813, doi:10.1016/j.cell.2018.03.041 (2018).
116 Sharma, A. *et al.* Longitudinal single-cell RNA sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy. *Nature Communications* **9**, 4931, doi:10.1038/s41467-018-07261-3 (2018).
117 Kumar, P., Tan, Y. & Cahan, P. Understanding development and stem cells using single cell-based analyses of gene expression. *Development* **144**, 17, doi:10.1242/dev.133058 (2017).
118 Chu, L.-F. *et al.* Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology* **17**, 173, doi:10.1186/s13059-016-1033-x (2016).
119 Tang, F. *et al.* Tracing the Derivation of Embryonic Stem Cells from the Inner Cell Mass by Single-Cell RNA-Seq Analysis. *Cell Stem Cell* **6**, 468-478, doi:https://doi.org/10.1016/j.stem.2010.03.015 (2010).
120 Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663-1677, doi:https://doi.org/10.1016/j.cell.2015.11.013 (2015).
121 Tsang, J. C. H. *et al.* Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells. *Genome Biology* **16**, 178, doi:10.1186/s13059-015-0739-5 (2015).
122 Macaulay, I. C. *et al.* Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. *Cell Reports* **14**, 966-977, doi:https://doi.org/10.1016/j.celrep.2015.12.082 (2016).
123 Zhou, F. *et al.* Tracing haematopoietic stem cell formation at single-cell resolution. *Nature* **533**, 487 (2016).
124 Hochane, M. *et al.* Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development. *PLOS Biology* **17**, e3000152, doi:10.1371/journal.pbio.3000152 (2019).
125 Johnson, M. B. *et al.* Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex. *Nature Neuroscience* **18**, 637, doi:10.1038/nn.3980 (2015).
126 Ofengeim, D., Giagtzoglou, N., Huh, D., Zou, C. & Yuan, J. Single-Cell RNA Sequencing: Unraveling the Brain One Cell at a Time. *Trends in Molecular Medicine* **23**, 563-576, doi:https://doi.org/10.1016/j.molmed.2017.04.006 (2017).
127 Shin, J. *et al.* Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell* **17**, 360-372, doi:https://doi.org/10.1016/j.stem.2015.07.013 (2015).
128 Camp, J. G. *et al.* Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proceedings of the National Academy of Sciences* **112**, 15672-15677, doi:10.1073/pnas.1520760112 (2015).
129 Combes, A. N. *et al.* High throughput single cell RNA-seq of developing mouse kidney and human kidney organoids reveals a roadmap for recreating the kidney. *bioRxiv*, 235499, doi:10.1101/235499 (2017).
130 Su, X. *et al.* Single-cell RNA-Seq analysis reveals dynamic trajectories during mouse liver development. *BMC Genomics* **18**, 946, doi:10.1186/s12864-017-4342-x (2017).
131 Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251, doi:10.1038/nature14966 (2015).
132 Mead, B. E. *et al.* Harnessing single-cell genomics to improve the physiological fidelity of organoid-derived cell types. *BMC Biology* **16**, 62, doi:10.1186/s12915-018-0527-2 (2018).
133 Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences* **115**, E2467-E2476, doi:10.1073/pnas.1714723115 (2018).
134 Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371, doi:10.1038/nature13173 (2014).
135 Yuzwa, S. A. *et al.* Developmental emergence of adult neural stem cells as revealed by single-cell transcriptional profiling. *Cell reports* **21**, 3970-3986 (2017).

136 Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524, doi:10.1038/nature25980 (2018).

137 Vladoiu, M. C. *et al.* Childhood cerebellar tumours mirror conserved fetal transcriptional programs. *Nature*, doi:10.1038/s41586-019-1158-7 (2019).

138 Lu, C.-J. *et al.* Single-cell analyses identify distinct and intermediate states of zebrafish pancreatic islet development. *Journal of Molecular Cell Biology*, doi:10.1093/jmcb/mjy064 (2018).

139 Wang, P. *et al.* Dissecting the Global Dynamic Molecular Profiles of Human Fetal Kidney Development by Single-Cell RNA Sequencing. *Cell Reports* **24**, 3554-3567.e3553, doi:https://doi.org/10.1016/j.celrep.2018.08.056 (2018).

140 Li, L. *et al.* Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions. *Cell Stem Cell* **20**, 858-873.e854, doi:https://doi.org/10.1016/j.stem.2017.03.007 (2017).

141 Cui, Y. *et al.* Single-Cell Transcriptome Analysis Maps the Developmental Track of the Human Heart. *Cell reports* **26**, 1934-1950. e1935 (2019).

142 Lescroart, F. *et al.* Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science* **359**, 1177-1181, doi:10.1126/science.aao4174 (2018).

143 Fan, X. *et al.* Single Cell and Open Chromatin Analysis Reveals Molecular Origin of Epidermal Cells of the Skin. *Developmental Cell* **47**, 21-37.e25, doi:https://doi.org/10.1016/j.devcel.2018.08.010 (2018).

144 Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981, doi:10.1126/science.aar4362 (2018).

145 Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131, doi:10.1126/science.aar3131 (2018).

146 DeLaughter, D. M. *et al.* Single-Cell Resolution of Temporal Gene Expression during Heart Development. *Developmental Cell* **39**, 480-490, doi:https://doi.org/10.1016/j.devcel.2016.10.001 (2016).

147 Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, doi:10.1038/s41587-019-0071-9 (2019).

148 Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32**, 381, doi:10.1038/nbt.2859 (2014).

149 Athanasiadis, E. I. *et al.* Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nature Communications* **8**, 2045, doi:10.1038/s41467-017-02305-6 (2017).

150 Kee, N. *et al.* Single-Cell Analysis Reveals a Close Relationship between Differentiating Dopamine and Subthalamic Nucleus Neuronal Lineages. *Cell Stem Cell* **20**, 29-40, doi:https://doi.org/10.1016/j.stem.2016.10.003 (2017).

151 Reid, A. J. *et al.* Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *eLife* **7**, e33105, doi:10.7554/eLife.33105 (2018).

152 Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research* **44**, e117-e117, doi:10.1093/nar/gkw430 (2016).

153 Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics* **19**, 477-477, doi:10.1186/s12864-018-4772-0 (2018).

154 Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989-2998, doi:10.1093/bioinformatics/btv325 (2015).

155 Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology* **37**, 451-460, doi:10.1038/s41587-019-0068-4 (2019).

156 Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology* **34**, 637, doi:10.1038/nbt.3569 (2016).

157 Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods* **14**, 979, doi:10.1038/nmeth.4402 (2017).

158 Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology* **20**, 59, doi:10.1186/s13059-019-1663-x (2019).

159 Marco, E. *et al.* Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences* **111**, E5643, doi:10.1073/pnas.1408993111 (2014).

160 Herring, C. A. *et al.* Unsupervised Trajectory Analysis of Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut. *Cell Systems* **6**, 37-51.e39, doi:10.1016/j.cels.2017.10.012 (2018).

161 Margolin, A. A. *et al.* ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* **7**, S7, doi:10.1186/1471-2105-7-s1-s7 (2006).

162 Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nature Methods* **16**, 43-49, doi:10.1038/s41592-018-0254-1 (2019).

163 Trapnell, C. HSMMSingleCell (R package, 2018).

164 Vivar, J. C., Pemu, P., McPherson, R. & Ghosh, S. Redundancy control in pathway databases (ReCiPa): an application for improving gene-set enrichment analysis in Omics studies and "Big data" biology. *OMICS* **17**, 414-422, doi:10.1089/omi.2012.0083 (2013).

165 Smith, J. & Schofield, P. N. The Effects of Fibroblast Growth Factors in Long-Term Primary Culture of Dystrophic (MDX) Mouse Muscle Myoblasts. *Experimental Cell Research* **210**, 86-93, doi:https://doi.org/10.1006/excr.1994.1013 (1994).

166 Bentzinger, C. F., Wang, Y. X. & Rudnicki, M. A. Building Muscle: Molecular Regulation of Myogenesis. *Cold Spring Harbor Perspectives in Biology* **4**, doi:10.1101/cshperspect.a008342 (2012).

167 Hannon, K., Kudla, A. J., McAvoy, M. J., Clase, K. L. & Olwin, B. B. Differentially expressed fibroblast growth factors regulate skeletal muscle development through autocrine and paracrine mechanisms. *The Journal of cell biology* **132**, 1151-1159 (1996).

168 Owens, J., Moreira, K. & Bain, G. Characterization of primary human skeletal muscle cells from multiple commercial sources. *In Vitro Cell Dev Biol Anim* **49**, 695-705, doi:10.1007/s11626-013-9655-8 (2013).

169 Venuti, J. M., Morris, J. H., Vivian, J. L., Olson, E. N. & Klein, W. H. Myogenin is required for late but not early aspects of myogenesis during mouse development. *The Journal of cell biology* **128**, 563-576 (1995).

170 Jiwlawat, N., Lynch, E., Jeffrey, J., Van Dyke, J. M. & Suzuki, M. Current Progress and Challenges for Skeletal Muscle Differentiation from Human Pluripotent Stem Cells Using Transgene-Free Approaches. *Stem Cells Int* **2018**, 6241681-6241681, doi:10.1155/2018/6241681 (2018).

171 Melo, F., Carey, D. J. & Brandan, E. Extracellular matrix is required for skeletal muscle differentiation but not myogenin expression. *Journal of Cellular Biochemistry* **62**, 227-239, doi:10.1002/(sici)1097-4644(199608)62:2<227::aid-jcb11>3.0.co;2-i (1996).

172 Scata, K. A., Bernard, D. W., Fox, J. & Swain, J. L. FGF Receptor Availability Regulates Skeletal Myogenesis. *Experimental Cell Research* **250**, 10-21, doi:https://doi.org/10.1006/excr.1999.4506 (1999).

173 Parakati, R. & DiMario, J. X. Repression of myoblast proliferation and fibroblast growth factor receptor 1 promoter activity by KLF10 protein. *J Biol Chem* **288**, 13876-13884, doi:10.1074/jbc.M113.457648 (2013).

174 Dwyer, N. D. *et al.* Neural Stem Cells to Cerebral Cortex: Emerging Mechanisms Regulating Progenitor Behavior and Productivity. *The Journal of Neuroscience* **36**, 11394-11401, doi:10.1523/jneurosci.2359-16.2016 (2016).

175 Martynoga, B., Drechsel, D. & Guillemot, F. Molecular control of neurogenesis: a view from the mammalian cerebral cortex. *Cold Spring Harbor perspectives in biology* **4**, a008359, doi:10.1101/cshperspect.a008359.

176 Homem, C. C. F., Repic, M. & Knoblich, J. A. Proliferation control in neural stem and progenitor cells. *Nature reviews. Neuroscience* **16**, 647-659, doi:10.1038/nrn4021 (2015).

177 He, Z. & Yu, Q. Identification and characterization of functional modules reflecting transcriptome transition during human neuron maturation. *BMC genomics* **19**, 262-262, doi:10.1186/s12864-018-4649-2 (2018).

178 Imayoshi, I., Sakamoto, M., Yamaguchi, M., Mori, K. & Kageyama, R. Essential Roles of Notch Signaling in Maintenance of Neural Stem Cells in Developing and Adult Brains. *The Journal of Neuroscience* **30**, 3489-3498, doi:10.1523/jneurosci.4987-09.2010 (2010).

179 Yoshida, Y. Semaphorin Signaling in Vertebrate Neural Circuit Assembly. *Frontiers in Molecular Neuroscience* **5**, doi:10.3389/fnmol.2012.00071 (2012).

180 Jongbloets, B. C. & Pasterkamp, R. J. Semaphorin signalling during development. *Development* **141**, 3292-3297, doi:10.1242/dev.105544 (2014).

181 van der Ven, P. F. M. *et al.* Differentiation of human skeletal muscle cells in culture: maturation as indicated by titin and desmin striation. *Cell Tissue Res.* **270**, 189-198, doi:10.1007/BF00381893 (1992).

182 Gao, X., Xiao, B., Tao, D. & Li, X. A survey of graph edit distance. *Pattern Analysis and Applications* **13**, 113-129, doi:10.1007/s10044-008-0141-y (2010).

183 Drton, M. & Maathuis, M. H. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application* **4**, 365-393 (2017).

184 Ding, J. *et al.* Reconstructing differentiation networks and their regulation from time series single-cell expression data. *Genome research* **28**, 383-395, doi:10.1101/gr.225979.117.

185 Fischer, D. S. *et al.* Inferring population dynamics from single-cell RNA-sequencing time series data. *Nature Biotechnology* **37**, 461-468, doi:10.1038/s41587-019-0088-0 (2019).

186 La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494-498, doi:10.1038/s41586-018-0414-6 (2018).

Time-series single-cell RNA sequencing (scRNAseq) can capture heterogeneity in cell states and transitions during dynamic biological processes, such as development and differentiation. Many trajectory inference methods have been developed to order cells by their progression through a dynamic process and infer the cells' movement trajectory. These methods, however, do not consider time information when ordering cells and are designed to work on snapshot scRNAseq data. In this thesis, I present a novel method, called Tempora, that uses pathway expression profiles and experiment time point information to infer the lineage relationships among different cell populations captured in time-series scRNAseq experiments. Tempora accurately inferred developmental lineages and important time-dependent signaling pathways in human skeletal myoblast differentiation and murine cerebral cortex development time-series scRNAseq data. These results demonstrate the power of using time information, when available, to supervise trajectory inference, as well as suggests that pathway expression profiles are an informative and less noisy alternative to gene expression profiles in representing individual cells for scRNA-seq based analysis.