# DISSECTING REGIONAL HETEROGENEITY AND MODELING TRANSCRIPTIONAL CASCADES IN BRAIN ORGANOIDS

**Dissertation**
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
*(Dr. rer. nat.)*

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von

**Daniel Rosebrock**

Berlin, 2023

Erstgutachter: **Prof. Dr. Martin Vingron**
Zweitgutachter: **Prof. Dr. Jan Philipp Junker**

Tag der Disputation: 23.02.2023

All life is an experiment. The more experiments you make the better.

— Ralph Waldo Emerson

# PREFACE

## PUBLICATIONS

The work discussed in this thesis grew from a collaboration between the labs of Yechiel Elkabetz, Peter Arndt, and Martin Vingron. The initial goals of the project were to compare the ability of a brain organoid derivation protocol developed in the Elkabetz lab (Triple-i) to other state of the art protocols to give rise to cortical neural stem cells and their progeny. This work culminated in a publication in the journal Nature Cell Biology (Rosebrock et al., 2022). Part of the results from the publication are summarized in Chapter 4. Chapter 5 contains a method we developed, in part out of necessity and in part of curiosity, while analyzing a pilot scRNA-Seq experiment for the above publication, which addresses a technological artifact that left us scratching our heads for many months. Chapter 6 contains yet unpublished work, addressing a question that arose after a series of observations of the spatial distribution of marker gene expression across different scRNA-Seq datasets in low-dimensional embeddings of the data. We expect to submit this part of the thesis in a future publication.

## ACKNOWLEDGEMENTS

I would like to thank the IMPRS for Computational Biology and Scientific Computing (IMPRS-CBSC) graduate program of the Max Planck Institute for Molecular Genetics and the Department of Mathematics and Computer Science at Freie Universität Berlin, which provided the opportunity to pursue my doctoral studies in Berlin. The wonderful community of students, and infrastructure including nanocourses and workshops, provided a supportive and engaging atmosphere. I would like to extend a special thank you to the PhD coordinator at the MPIMG, Kirsten Kelleher, and her successor, Anne-Dominique Gindrat, as well as our department administrator, Martina Lorse, for their unlimited support in all administrative and organizational matters surrounding my studies.

I would like to thank my supervisor Martin Vingron, who gave me the freedom to ask my own questions and explore my own interests. His relentless optimism, perceptive insights, and invaluable life lessons were inspirational throughout my entire studies. Thank you for your support and guidance. Furthermore, the incredibly supportive and collaborative atmosphere in the department made coming in to the office something to look forward to. I am very grateful to everyone from the department for creating such a wonderful environment.

I would like to thank my co-advisor, Peter Arndt, who supported me in all of my endeavours throughout my studies, and whose ideas made the final chapter of this thesis possible. Part of the reason for coming to the MPIMG in the first place was due to a chance encounter in Boston —

thank you for persuading me to apply to the program. And thank you for your enduring patience and insights throughout this work, and especially for providing essential caffeine supplies at the office!

I would like to thank my co-advisor, Yechiel Elkabetz, who gave me the unique opportunity to join his lab and become deeply involved in all of the exciting projects in the group, as well as the freedom to explore whatever problems I found interesting. I am very grateful for his countless hours of explaining neural stem cell biology and neurodevelopment, and instilling in me a fascination for this endlessly captivating and complicated topic. Thank you for your constant support and guidance. From the lab, I would especially like to thank Amèlia Aragonés Hernández, Naresh Mutukula, Sneha Arora and Anastasios Balaskas for thoughtful discussions, collaborations, and personal friendships.

Furthermore, I wish to express my gratitude to Jan Philipp Junker for his time in reviewing this thesis, and who provided many insightful comments during my TAC meetings.

I would also like to thank my parents for their never-ending support, no matter the circumstances. I am eternally grateful to both of you! A special thank you to my sister, Laina, who I could always turn to for advice in times of uncertainty. Finally, thank you to my wonderful wife, Anna, who supported me in all the ups and downs throughout this journey, and our beautiful daughter, Lise, whose smiles and giggles have been an unlimited source of joy in my life.

# CONTENTS

1 INTRODUCTION 1

2 BIOLOGICAL BACKGROUND 3

2.1 Regulation of gene expression . . . 3

2.1.1 Transcription . . . 4

2.1.2 Transcriptional regulation . . . 6

2.2 Nervous system development . . . 8

2.2.1 Early embryogenesis . . . 9

2.2.2 Neural induction and patterning . . . 10

2.2.3 Neocortex development . . . 13

2.2.4 In vitro neural induction with brain organoids . . . 15

2.3 Experimental techniques to measure the transcriptome . . . 16

2.3.1 RNA sequencing . . . 17

2.3.2 Single-cell RNA-Sequencing . . . 18

3 COMPUTATIONAL BACKGROUND 21

3.1 Learning probabilistic models . . . 21

3.1.1 Maximum likelihood estimation . . . 21

3.1.2 Bayesian parameter estimation . . . 23

3.1.3 Markov chain Monte Carlo . . . 26

3.2 Dimension reduction . . . 34

3.2.1 Principal component analysis . . . 35

3.2.2 Diffusion maps . . . 35

3.2.3 Uniform manifold approximation and projection . . . 36

3.3 Clustering . . . 37

3.3.1 Graph-based clustering approaches . . . 38

4 DISSECTING REGIONAL HETEROGENEITY IN BRAIN ORGANOIDS 39

4.1 Comparing regional specification from bulk RNA-Seq . . . 39

4.2 Measuring regional heterogeneity from scRNA-Seq . . . 41

4.3 Summary . . . 44

5 ADDRESSING TECHNOLOGICAL ARTIFACTS IN SCRNA-SEQ 45

5.1 Detecting cell-containing droplets in scRNA-Seq data . . . 45

5.2 Artifactual modes in scRNA-Seq data . . . 47

5.3 Detecting empty droplets . . . 48

5.3.1 Simulating empty droplets . . . 49

5.3.2 Classification of cell barcodes as background using co-clustering . 51

5.3.3 Genotyping individual cells to detect empty droplets . . . 54

5.3.4 Comparison with other methods . . . 57

5.3.5 Application to inDrops scRNA-Seq of day 50 brain organoids . . . 60

5.3.6 Application to 10X scRNA-Seq of day 50 brain organoids . . . 63

5.4 Summary . . . 67

6 MEASURING TRANSCRIPTIONAL CASCADES IN SCRNA-SEQ 69

6.1 Modeling dynamics of gene expression along pseudotemporal trajectories 70
6.1.1 Establishing a likelihood model . . . . . . . . . . . . . . . . . . . . 73
6.1.2 Estimation of global dispersion parameter . . . . . . . . . . . . . . 73
6.1.3 Mapping from log-normalized to count space . . . . . . . . . . . . 74
6.1.4 Model inference using MCMC . . . . . . . . . . . . . . . . . . . . . 75
6.1.5 Model Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . 79
6.1.6 MCMC Diagnostics . . . . . . . . . . . . . . . . . . . . . . . . . . 82
6.1.7 Estimating inflection points . . . . . . . . . . . . . . . . . . . . . . 85
6.2 Measuring transcriptional cascades and regulatory interactions . . . . . . 86
6.2.1 Transcriptional cascades in cortical cells of brain organoids . . . . 87
6.2.2 Constructing regulatory interactions in cortical cells of brain organoids 89
6.2.3 Comparing regulatory interactions in day 30 and 50 brain organoids 91
6.2.4 Dissecting Notch signaling in day 30 and 50 brain organoids . . . . 94
6.3 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 95
7 DISCUSSION AND CONCLUSION 97

ABBREVIATIONS 101
LIST OF FIGURES 103
LIST OF TABLES 105

A APPENDIX 107
A.1 Experimental Procedures and Data Processing . . . . . . . . . . . . . . . . 107
A.2 Supplemental Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . 109

BIBLIOGRAPHY 113
ABSTRACT 127
DECLARATION 129

# 1 INTRODUCTION

The formation of the nervous system is a highly elaborate developmental process involving the segregation of cells into distinct regions via external signaling cues and the differentiation of neural stem cells within these regions into various progenies, including neurons and glial cells, and gives rise to arguably the most complex structure in the known universe, the human cortex. Access to human prenatal brain samples is limited due to the difficulties in acquiring such samples. Therefore, the generation of robust in vitro model systems to study the formation and development of the human brain is critical. Brain organoids exhibit the remarkable ability to self-organize into 3-dimensional structures which mimic in vivo brain development. However, different derivation protocols give rise to heterogeneous populations with respect to regional specification, and differ in their differentiation capacities across cell lines. The utility of a given protocol is largely reflected in its ability to generate consistent cell populations across a variety of cell lines, further enabling a systematic comparison of brain development under healthy and disease conditions, for example from patient-derived cell lines.

In order to assess the cellular heterogeneity in these tissues, single-cell RNA-Sequencing (scRNA-Seq) has become the method of choice, enabling the unprecedented ability to survey the global transcriptomic profile of individual cells at scale. Over the past decade, scRNA-Seq has revolutionized our understanding of a variety of biological processes, from developmental biology to cellular reprogramming and cancer biology. This technology has enabled the identification of previously uncharacterized cell types, and novel marker genes defining different cell populations. However, while this technology has become an indispensable tool for many researchers in understanding cellular heterogeneity, a careful analysis of the data is essential to identify and remove potential technical artifacts, and enrich for meaningful biological signal. Once these issues are addressed, scRNA-Seq can furthermore be used to characterize cell types, and infer the underlying gene regulatory programs in cells as they differentiate to form different organs, such as the brain.

## Thesis outline

Following this chapter, a more detailed introduction into the biological background is provided in Chapter 2. We introduce fundamental concepts in molecular biology, starting from transcription and transcriptional regulation, discuss brain development in vivo, focusing on the signaling pathways and genes that specify regional patterning and neurogenesis, and provide a historic overview of modeling neurodevelopment in vitro. Finally, we describe how next-generation sequencing technologies work, in particular bulk RNA-Seq and scRNA-Seq. Chapter 3 provides an introduction into probabilistic models and parameter estimation. We also introduce mathematical concepts that are essential for the analysis of scRNA-Seq datasets, including dimensionality reduction techniques and clustering. In Chapter 4, we discuss how to dissect regional enrichments across different brain organoid protocols using bulk RNA-Seq and scRNA-Seq. In Chapter 5, we highlight sources of technological artifacts that arise in scRNA-Seq datasets, and describe a method we developed to remove one such artifact. Finally, in Chapter 6, we present a method to recover gene transcriptional and regulatory dynamics along developmental trajectories from scRNA-Seq data, and apply this method to developing brain organoids at various stages of development. Conclusions and a discussion on future directions, as well as potential experiments to validate the findings in Chapter 6, can be found in Chapter 7.

# 2 BIOLOGICAL BACKGROUND

Living organisms are comprised of cells, the functional units of all known life forms. Different cells in the human body have specialized functions to perform particular tasks. For example, nerve cells transmit electrical signals in the nervous system and send motor commands to muscles in the body, while white blood cells form part of the body's immune system to fight infections and other diseases. All of the cells in the human body originated from a single cell, the zygote, a fertilized egg resulting from the union of a human egg and sperm. The zygote contains all of the instructions, or genetic material, needed for specifying when and where all of the cells are generated within the human body during embryonic development. This process is governed by a strict regulation of the production of different cellular components within the cell and its respective progeny. In this chapter, we will explore these fundamental components and their roles in shaping cellular identity, with a specific focus on nervous system development.

## 2.1 REGULATION OF GENE EXPRESSION

The blueprint for the generation and maintenance of all living organisms is encoded by deoxyribonucleic acid (DNA). DNA carries the genetic instructions for the development and function of individual cells, and serves as the primary hereditary unit through which organisms pass on their traits to their offspring. It is comprised of nucleotides, where each nucleotide consists of a sugar (deoxyribose), a phosphate group, and one of four bases - cytosine (C), guanine (G), adenine (A) and thymine (T). The nucleotides are then joined together to form a DNA strand by their respective phosphate groups between the third (3') and fifth (5') carbon atoms of adjacent sugar rings, giving the DNA strand a specific orientation with either a downstream (5' to 3') or upstream (3' to 5') direction. DNA does not typically exist as a single-stranded molecule but rather a double-stranded molecule, with the individual strands of DNA being bound together according to base pairing rules, with A complementary to T and G complementary to C, to form a double helix. Variation in the composition of the nucleotides in the DNA sequence of individual organisms enables the diversification within and evolution of different species, as well as genetic diseases and the formation of aberrant cell types.

The functional unit of DNA is the gene, which is fundamentally a sequence of nucleotides in the DNA that is copied into ribonucleic acid (RNA), during the process of transcription. Similar to DNA, RNA is comprised of four nucleotides, with T replaced by uracil (U), joined together by a sugar-phosphate backbone containing ribose instead of deoxyribose. Unlike DNA, which is double-stranded, RNA is a single-stranded molecule. RNA can be synthesized into proteins, comprised of a sequence of amino acids, by ribosomes in a process known as translation, in the case of coding RNAs, or remain as RNA in the case of non-coding RNAs. This process of transcription followed by translation of individual genes is known as gene expression, and is summarized by the central dogma of molecular biology, illustrated in Figure 2.1. The central dogma was originally formulated by Francis Crick in the year 1958, who played a crucial role in deciphering the helical structure of DNA (Crick, 1958). The systematic regulation of gene expression gives strict control over the quantity of different RNAs and proteins in a cell, thereby enabling cellular differentiation and morphogenesis, as well as the adaptability of a cell to respond to changes in its environment.

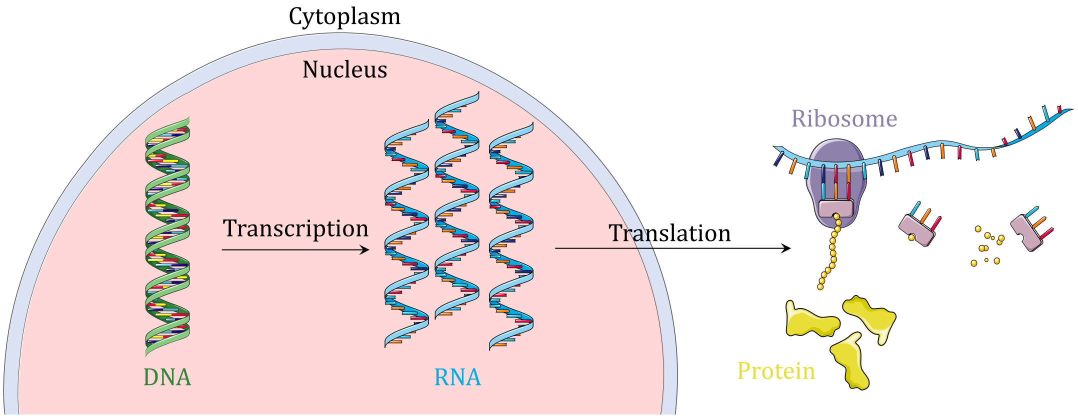

**Figure 2.1: Central dogma of molecular biology.** DNA is stored in the nucleus of the cell and is copied into RNA during the process of transcription, which also occurs in the nucleus. RNA is then transported to the cytoplasm, where ribosomes translate RNA into proteins. (Original illustrations taken from SMART Servier Medical Art [https://smart.servier.com/])

### 2.1.1 Transcription

There are many different types of RNAs in a cell. Messenger RNA (mRNA) is RNA that will be translated into proteins, and serves as the bridge between DNA and proteins, as illustrated in Figure 2.1. mRNA encodes for proteins via codons, or sequences of three ribonucleotides, each of which codes for a specific amino acid, with the exception of the stop codon, which terminates protein synthesis. However, not all RNA is translated into proteins. Recent studies have estimated around 21,000 protein-coding genes and 22,000 non-coding genes in the human genome (GTEx Consortium, 2017). Examples of non-coding RNA include transferRNAs (tRNAs) and ribosomal RNAs (rRNAs), which are essential for the process of translation by the ribosomes, as well as microRNAs, small interfering RNAs (siRNAs), and long non-coding RNAs (lncRNAs), which are involved in the regulation of gene expression, among others.

In eukaryotes (cellular organisms with a cell nucleus), mRNA is synthesized complementary to a DNA template by the the enzyme RNA polymerase, which traverses the DNA template strand in the 3' to 5' direction, opening the DNA double helix as it moves along, and synthesizes mRNA in a 5' to 3’ direction. While eukaryotes have three types of RNA polymerases, prokaryotes (cells without a nucleus) have only one. RNA Polymerase II (RNAP II) is responsible for transcription of some classes of non-coding RNAs and mRNA, which is the main focus in this section. Transcription initiation happens at gene promoters, *cis*-regulatory elements consisting of short sequences of DNA located directly upstream of the gene (in the 5’ direction) to which proteins called transcription factors bind and recruit RNAP II, which also binds to the DNA in the promoter. Transcription factors bind to specific transcription factor binding sites (TFBS), short sequences of DNA in the promoter, which can be recognized by the DNA-binding domain of the transcription factor. Transcription factors can either activate gene expression via recruitment of RNAP II and other coactivators, or repress gene expression by blocking the attachment of RNAP II to the promoter, and thus play a pivotal role in orchestrating gene expression. Transcription then begins at the transcription start site, located directly downstream of the promoter (in the 3’ direction), in the process of elongation, whereby the RNA strand is synthesized.

Following transcription of mRNA from the DNA, a number of maturation steps takes place during RNA processing to ensure the correct sequence of ribonucleotides is translated into protein, and to stabilize the mRNA molecule. These steps are highlighted in Figure 2.2. The initial mRNA molecule, or precursor mRNA (pre-mRNA), typically contains introns, regions of the mRNA which do not code for amino acid sequences and are removed prior to translation in a process known as splicing. The final mRNA product, or mature mRNA, then consists of exons, regions that will encode the protein, as well as untranslated regions on the 5’ end and 3’ end of the mRNA that are not translated, referred to as 5’ untranslated region (5’ UTR) and 3’ untranslated region (3’ UTR) respectively. Furthermore, a 5’ cap consisting of a methylated guanine is added to the 5’ end of the pre-mRNA to ensure its protection from degradation by ribonuclease (RNase), and is critical for recognition of the mRNA by the ribosome. At the 3’ end of the pre-mRNA, a poly-A tail is attached in a process known as polyadenylation, whereby 100 to 250 A’s are added to the 3’ end of the RNA molecule. Similar to the 5’ cap, the 3’ poly-A tail acts to protect the mRNA molecule from degradation by exonucleases and facilitates export of the mRNA molecule from the nucleus to the cytoplasm, where translation occurs.

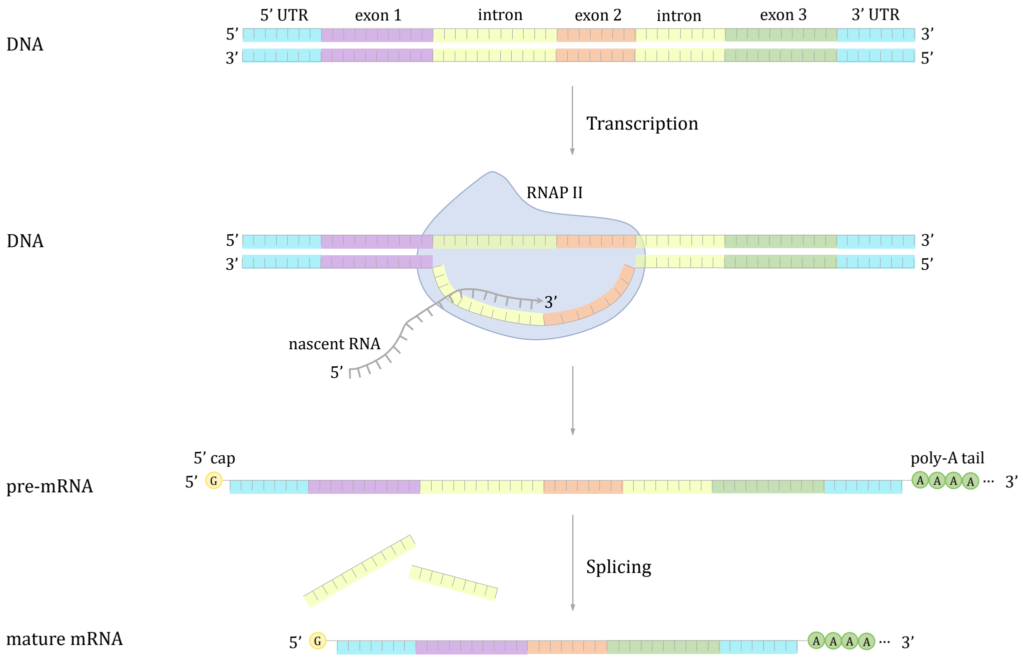

**Figure 2.2: Transcription in eukaryotic cells.** RNA is synthesized from a DNA template by the enzyme RNAP II in a 5' to 3' direction, with the 5' end transcribed first and 3' end transcribed last. Only one of the two DNA strands serves as the template for transcription. During elongation, nascent RNA, or newly synthesized RNA, is bound to the DNA template strand by RNAP II. The addition of a 5' cap and 3' poly-A tail provide stability to the mRNA molecule and facilitate export of the mature mRNA from the nucleus to the cytoplasm. Splicing out of introns takes place in the nucleus and can occur either during (cotranscriptionally) or immediately after transcription.

### 2.1.2 Transcriptional regulation

Less than 2% of the human genome represents coding DNA, or the stretches of DNA in genes which will ultimately be transcribed and translated into proteins. The remaining portions of the human genome contain intronic sequences, repetitive sequences, segmental duplications (duplications of large segments of genomic DNA ranging in size from 1 to 400 kilobases that were duplicated and reinserted into the genome at a different location), retrotransposons (genetic components that copy and paste themselves into different genomic locations by converting RNA to DNA via reverse transcription) including long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), long terminal repeat (LTR) retrotransposons, and DNA transposons, among other sequences (International Human Genome Sequencing Consortium et al., 2001). For decades, the function of the non-coding regions of the genome was poorly understood, and these non-coding regions were given the name “junk” DNA, originally coined by geneticist Susumu Ohno (1972). However, over the past decades, with projects such as ENCODE (The ENCODE Project Consortium, 2012), scientists have begun to shed light on the regulatory function of these regions of the genome.

Similar to promoters, enhancers are *cis*-regulatory elements that influence the transcription of genes on the same DNA molecule. Unlike promoters, enhancers can be located up to 1 megabase away from their target gene either upstream or downstream from the transcription start site (Pennacchio et al., 2013). While an individual gene has a single promoter upstream of the transcription start site, multiple enhancers can regulate the expression of a single gene, and multiple genes can be regulated by a single enhancer. Enhancers are known to be involved in cell fate and tissue development and control gene expression in a cell type-specific manner (Bonn et al., 2012; Taminato et al., 2016). For example, deletion of a single enhancer for the gene BIN1 harboring an Alzheimer's disease risk variant (a mutation in the DNA associated with the disease) resulted in highly reduced expression of the gene in microglia, but not in neurons or astrocytes, thus demonstrating the cell-type specific role of gene regulation for the individual enhancer (Nott et al., 2019). Enhancers are activated or repressed in a similar fashion to promoters, namely by transcription factors that bind to specific TFBSs in the enhancer region.

The 3D architecture of the DNA also plays a pivotal role in the regulation of transcription, with the creation of loops bringing enhancers into the proximity of their target promoters (Ptashne, 1986). In the nucleus of eukaryotic cells, the massive amount of DNA (3.2 billion base pairs across 46 chromosomes in humans, nearly 2 meters long when linearly stretched) is packaged into a highly compact form, with 147 base pairs of DNA wrapped around a histone octamer consisting of two copies of each core histone protein (H2A, H2B, H3 and H4) to form nucleosomes (Richmond et al., 2003). In addition, nucleosomes are further packaged into chromatin (the complex of DNA and protein in the nucleus) fibers and folded into loops. Segments of DNA are further segregated into spatially insulated regions known as topologically associating domains (TAD) (Dixon et al., 2012). These structures are formed through interactions between CCCTC-Binding factor (CTCF), a protein involved in regulating the three-dimensional chromatin structure, and cohesin ring-shaped complexes at the TAD boundaries (Pombo et al., 2015). The formation of TADs facilitates transcriptional regulation by restricting the interactions between distal enhancers and their respective target promoters to within a TAD, with genes often being coregulated during cell differentiation when they are located in the same TAD (Ramírez et al., 2018). Chromatin can also be segregated into lamina-associating domains (LAD) which are transcriptionally silent, in which chromatin attaches to the nuclear lamina, a network-like structure located at the inner membrane of the nucleus (Guelen et al., 2008).

Furthermore, gene expression can be regulated via modifications of the histone proteins around which the DNA is wrapped. Post-translational modifications of histones, such as phosphorylation, ubiquitination, and the addition of acetyl or methyl groups, alter the physical properties of the chromatin, thereby promoting a more transcriptionally active or transcriptionally repressed state. For example, the methylated histones H3K27me3 and H3K9me2/3 facilitate the formation of heterochromatin, a highly compacted form of DNA which is mostly inaccessible for the transcriptional machinery, and is typically found in transcriptionally inactive regions of the genome (Bannister et al., 2011). These groups are added by chromatin regulators such as EZH2, which are typically recruited by transcription factors whose role is to repress expression of the target gene. On the other hand, the acetylaion of H3K27, H3K27ac, and methylated H3K4me3, are associated with a higher activation of transcription, with these histone modifications being typically found near the transcription start site and at active promoters and enhancers (Calo et al., 2013). These modifications disrupt histone-DNA interactions and facilitate the formation of euchromatin, an open chromatin state where the DNA is not wrapped around nucleosomes, enabling the binding of transcriptional machinery to the DNA. By altering the histone modifications and chromatin state of different genomic regions, cells are able to influence the activation or repression of specific genes in a cell-type specific manner, enabling cell-type specific gene expression patterns necessary to maintain cellular identity during cell-lineage differentiation and other cellular processes (Clouaire et al., 2018; Heintzman et al., 2009; Roh et al., 2006).

Not only can histone modifications influence gene expression, but also methylation of the DNA itself can contribute to different levels of gene expression. Methylation of the DNA can occur at A or C bases, but typically occurs on a C nucleotide at CpG dinucleotides in the mammalian genome (Ehrlich et al., 1982). A common mechanism for gene silencing involves the methylation of CpG islands, regions in the genome of at least 200 base pair (bp) with a high frequency of CpG dinucleotides, located in the gene's promoter. In fact, approximately 70% of genes in the human genome are associated with a CpG island in the promoter region, indicating CpG methylation is a widespread mechanism for gene silencing (Saxonov et al., 2006). DNA methylation of gene promoter CpG dinucleotides silences gene expression by hampering the ability of transcription factors to bind to the DNA sequence, ultimately resulting in the formation of heterochromatin (Rose et al., 2014). DNA methylation is maintained across cell divisions by the enzyme DNMT1. However, the de novo DNA methyltransferases DNMT3A and DNMT3B are able to transfer methyl groups to CpG sites dynamically, while TET proteins can initiate demethylation of methylated CpG sites, enabling dynamic DNA methylation changes. After fertilization, a major wave of global demethylation occurs across the genome, however, already at the post-implantation stage of embryonic development, DNA methylation is relatively stable (Guo et al., 2014; Smith et al., 2014). Furthermore, as cells differentiate, they acquire unique DNA methylation patterns which regulate cell-type specific gene expression (Deaton et al., 2011; Ziller et al., 2013).

Together, the histone modifications and DNA methylation patterns represent the epigenome of a cell. These processes play an essential role in modifying the chromatin structure, enabling the compaction of DNA in the nucleus. However, the ultimate effect of epigenetics is the regulation of gene expression, altering when and how genes are expressed. In the context of cellular differentiation, transcriptional regulation is a tightly controlled and dynamic process, with different gene expression programs characterizing the formation and identity of the diversity of cells in the human body. In the following section, we will explore the development of the nervous system, with an emphasis on its molecular underpinnings, as well as recent advances in modeling this complex process in in vitro systems. Following this, we will review experimental techniques to measure the transcriptomes of bulk samples and individual cells, and how recent advances in these technologies are shaping our understanding of developmental biology.

## 2.2 NERVOUS SYSTEM DEVELOPMENT

In vertebrates, the nervous system is comprised of two components, the central nervous system (CNS) consisting of the brain and spinal cord, and the peripheral nervous system (PNS) consisting of nerves (cable-like bundles of nerve fibers known as axons) that connect the CNS to every other part of the body. The nervous system contains neurons, specialized cells which are able to receive and transmit electrical signals, along with glia, which provide structural and metabolic support to the neurons. As with all the other cell types and organs in the body, the nervous system developed from a single cell, the zygote, and underlying the process of cell fate determination and differentiation is the precise regulation of gene expression. This process can be nicely visualized in the seminal work by Conrad Waddington (1957), illustrated in Figure 2.3, in which Waddington related cellular differentiation to the process of a marble rolling down a hill, with gene expression patterns shaping the various peaks and valleys along the hill, thereby regulating cellular development and differentiation trajectories.

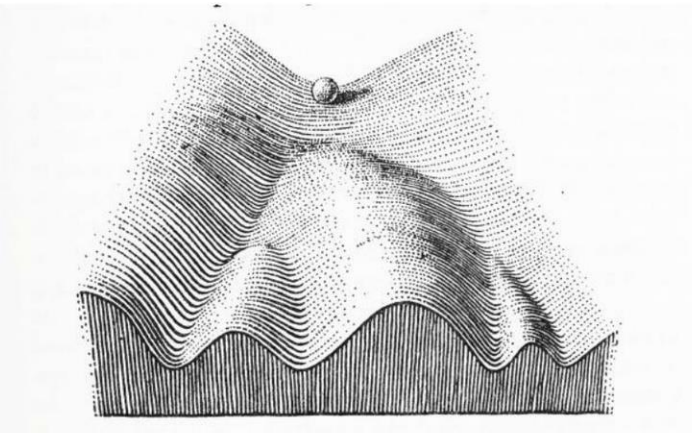

**Figure 2.3: Waddington's epigenetic landscape.** The depiction of a marble sitting on top of a hill marked by uneven slopes and valleys serves as a metaphor for cellular differentiation. The hill represents cellular differentiation, from a stem cell sitting on the top and various differentiated cell types on the bottom. Different gene expression programs form the various boundaries along the hill, enforcing the cell to differentiate into specific lineages upon successive bifurcations as it rolls down. Reprinted from (Waddington, 1957).

### 2.2.1 Early embryogenesis

Embryonic development is initiated by the fertilization of the oocyte by a sperm cell, giving rise to the zygote. In mammals, the zygote then undergoes a series of symmetrical cell divisions, in which these stem cells gives rise to two identical daughter cells with the same differentiation potential as the parent cell at each cell division, and eventually gives rise to the blastocyst, consisting of the inner cell mass (ICM) and the outer layer trophectoderm. The trophectoderm consists of the first cells to differentiate from the zygote, and during development gives rise to the embryonic portion of the placenta, an extraembryonic tissue which provides nutrients to the developing fetus (Wu et al., 2010). Post implementation of the blastocyst into the endometrium of the uterus, the ICM cells differentiate into the three germ layers known as the ectoderm (giving rise to the skin and nervous system), mesoderm (giving rise to the skeletal system, muscle tissue, kidney, heart, and hematopoietic system among other cell types), and the endoderm (forming the gastrointestinal and respiratory tracts, liver, pancreas and endocrine glands) in a highly coordinate process known as gastrulation (Gilbert, 2009). The initial stages of gastrulation are highlighted in Figure 2.4. The cells comprising the ICM are also known as pluripotent stem cells (PSC) due to their ability to give rise to any type of cell in the body.

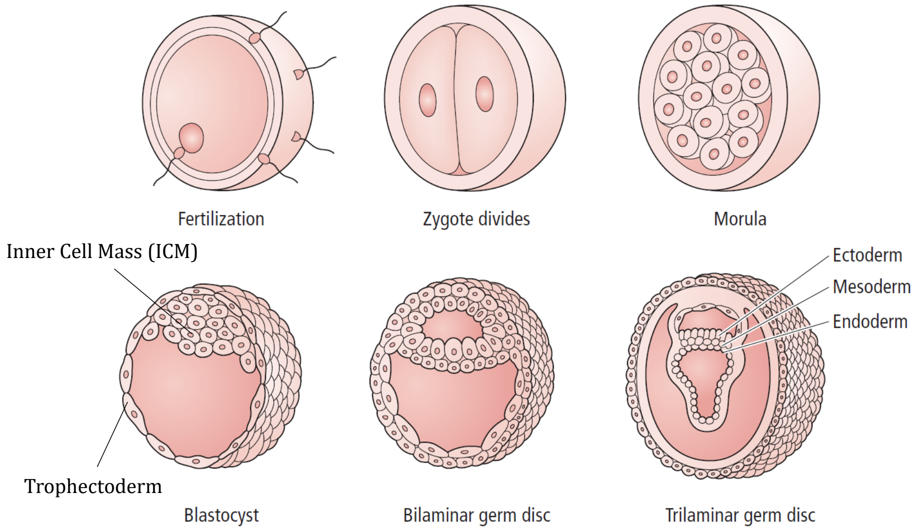

**Figure 2.4: Early stages of mammalian embryogenesis.** Following fertilization, the zygote gives rise to the blastocyst, consisting of the trophectoderm (giving rise to the embryonic portion of the placenta) and the ICM, which differentiates into all three germ layers, eventually giving rise to all cell types in the body. Adapted from (Patestas et al., 2016).

## 2.2.2 Neural induction and patterning

The nervous system is derived from the ectoderm within the blastocyst. Following the specification of the three germ layers, the ectoderm begins to thicken as a result of increased mitotic activity to form the neural plate. The formation of the neural plate is induced by the notochord, a long rod-like body which is derived from the mesoderm and defines the longitudinal axis of the embryo (Haines et al., 2018). This increased activity alongside alterations in cell morphology cause the neural plate to fold into a structure called the neural groove, flanked by neural folds on either side. The neural folds then fuse together with each other in the midline, which leads to the closure of the neural tube that then detaches from the ectoderm, completing internalization of the nervous system in a process known as neurulation. The overlaying ectoderm will later on give rise to the epidermis, the outermost layer of the skin. As the edges of the neural fold meet, a narrow strip of cells becomes separated to form the neural crest. These cells migrate throughout the entire body to form most of the PNS including sensory neurons (neurons that carry sensory input to the CNS), motor neurons (neurons that send signals from the CNS to muscles in the body to stimulate movement), and several non-neural cell types including melanocytes (pigment-containing cells of the epidermis), Schwann cells (glial cells that provide support and maintenance of PNS neurons), and skeletal and connective tissue of the head, among others. The neural tube then gives rise to the entire CNS consisting of the brain and spinal cord. The process of neurulation is depicted in Figure 2.5.

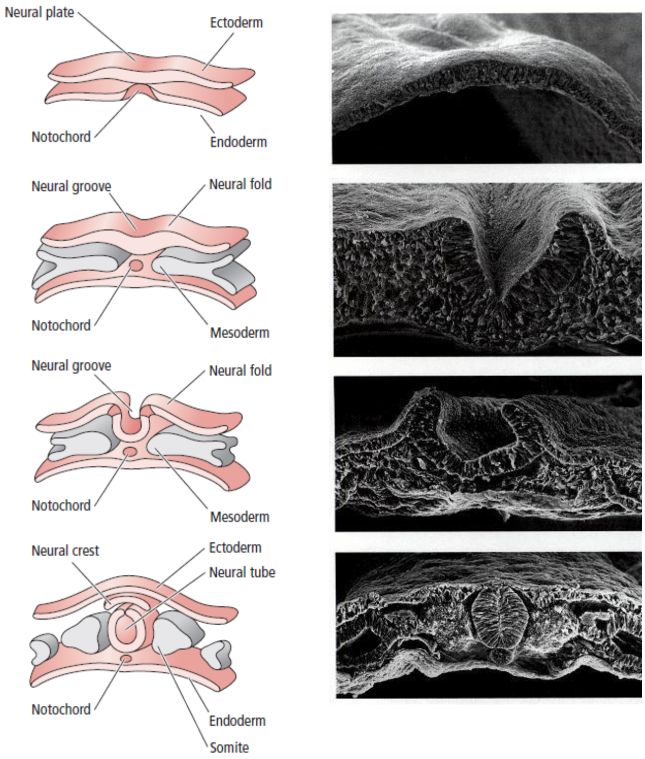

**Figure 2.5: Neurulation in mammalian embryogenesis.** Neurulation is the process whereby the nervous system is formed and internalized into the body. This process is induced by the notochord, which initiates formation of the neural plate, followed by the neural folds, which fuse together to form the neural tube, giving rise to the CNS, and neural crest, giving rise to the PNS among other cell types. Drawings on the left adapted from (Patestas et al., 2016). Scanning electron micrograph images of chick embryo on the right adapated from (Gilbert, 2009).

The early neural tube is a tubular structure composed of a single layer of neuroepithelial cells, also termed neural stem cells (NSC), the stem cells of the CNS. The neural tube is then subdivided into three major vesicles, the prosencephalon (forebrain), mesencephalon (midbrain) and rhombencephalon (hindbrain), with the caudal (posterior) region becoming the future spinal cord. As these vesicles continue to develop, they differentiate into secondary vesicles, with the prosencephalon sub-dividing into the telencephalon (future cerebral cortex, as well as subcortical structures including the hippocampus, amygdala, basal ganglia and olfactory bulb) and diencephalon (future thalamus, hypothamalus and optic vesicles). The rhombencephalon is also subdivided into two regions, the metencephalon (future pons and cerebellum) and myelencephalon (future medulla oblangata). This process is depicted in Figure 2.6.

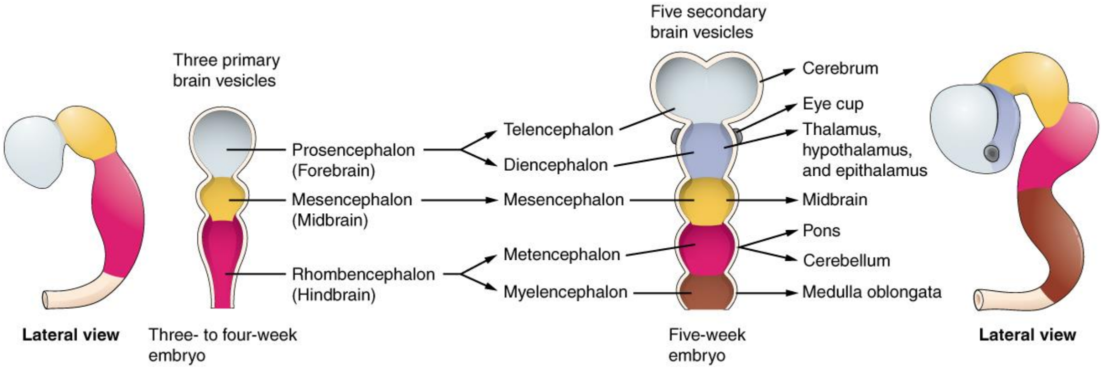

**Figure 2.6: Formation of primary and secondary brain vesicles.** The embryonic CNS develops from the neural tube into primary and secondary vesicles which form the future subregions of the brain, with the future spinal cord formed in the caudal region. Adapted from Textbook OpenStax Anatomy and Physiology.

The early development of the CNS is defined by regional patterning, the acquisition of distinct identities of the various regions of the neural tube according to spatial positions, which is controlled by signaling gradients across the anterior-posterior and dorsal-ventral axes (Lumsden et al., 1996). During this process, signaling molecules (also referred to as growth factors or ligands) are released from signaling centers along the neural tube. These molecules then bind to the receptor molecules on the cell surface of the target cell to initiate a sequence of intra-cellular events, ultimately resulting in the transcriptional regulation of a single gene or multiple genes in the target cell. For example, WNT signaling plays an important role in neural induction and anterior-posterior axis patterning, and begins when a Wnt protein binds to a Frizzled family receptor and several coreceptors such as lipoprotein receptor-related protein (LRP)-5/6 (Logan et al., 2004). During development, WNT1 is expressed in the midbrain-hindbrain regions (McMahon et al., 1992). After binding of WNT1 to the Frizzled and LRP5/6 ligand, $\beta$-catenin (CTNNB1) accumulates in the cytoplasm and eventually localizes to the nucleus, where it acts as a coactivator of transcription factors belonging to the TCF/LEF family (Logan et al., 2004). One of the downstream targets of the TCF/LEF family is GBX2, which is a key player in the formation of the midbrain-hindbrain region and contains a TCF/LEF binding site in its promoter (Li et al., 2009).

Another signaling molecule that is essential for proper neural tissue formation is Noggin, which is secreted by the notochord during the formation of the neural tube (Marcelino et al., 2001). Noggin is an inhibitor of several bone morphogenetic proteins (BMP), which are a group of growth factors that bind to bone morphogenetic protein receptors (BMPR) on the cell surface. Once activated, signal transduction through the activated BMPRs results in phosphorylation and subsequent activation of members of the SMAD protein family, which then form complexes with other SMAD proteins and translocate to the nuclues, where they then bind to DNA to regulate gene expression of various target genes (Hill, 2016). Noggin specifically binds tightly to the growth factors BMP4 and BMP7, thereby preventing the binding of these molecules to their receptors, blocking the downstream activation of SMAD complexes (McMahon et al., 1998). The BMP signaling pathway belongs to the transforming growth factor beta (TGF-$\beta$) superfamily. The TGF-$\beta$ ligand binds to the TGF-$\beta$ receptor on the cell surface of the target cell (which can also be activated by Activin/Nodal) to initiate SMAD activation similarly to BMP signals, and is a major inducer of mesoderm and endoderm lineages during embryonic development (Vallier et al., 2004). Similar to Noggin, the signaling molecule Sonic hedgehog protein (SHH) is also secreted by the notochord, and plays a key role in dorsal-ventral axis patterning by forming a gradient, with the ventral neural tube receiving high levels of SHH and dorsal neural tube receiving low levels of SHH (Blaess et al., 2006). The SHH gradient then induces the formation of the ventral structures of the CNS such as the basal ganglia and motor neurons by inducing the expression of the transcription factors OLIG2 and NKX2-2 among others (Ribes et al., 2009). Finally, the formation of the anterior regions of the CNS including the forebrain are believed to rely heavily on the anterior visceral endoderm (AVE), which migrates to the future anterior region of the embryo and secretes molecules such as Lefty1 and Cerebrus, as well as Dkk1, which act as inhibitors of TGF-$\beta$ and WNT respectively, and thus help to establish anterior neural fates (Andoniadou et al., 2013).

### 2.2.3 Neocortex development

The largest and most complex structure of the mammalian brain is the neocortex. The neocortex evolved most recently during mammalian evolution, and is responsible for higher-order brain functions such as sensory perception, cognition, spatial reasoning, generation of motor commands and language (Florio et al., 2014). It is part of the cerebral cortex and originates from the most anterior and dorsal part of the neural tube. During formation of the neocortex, the neuroepithelial cells in the anterior neural tube orient themselves in a pseudo-stratified manner with a strong apico-basal polarity, with apical processes connecting them to the lumen of the neural tube forming the ventricular zone (VZ) and basal projections connecting them to the basal lamina, and are thus often referred to as apical radial glia (aRG) (Ferent et al., 2020; Kriegstein et al., 2003). During initial stages of cortical development, neuroepithelial cells undergo a series of symmetric cell divisions resulting in two identical daughter stem cells, thereby greatly expanding the progenitor pool. This phase is followed by asymmetric cell divisions, in which aRG give rise to basal progenitors, also known as intermediate progenitors (IP). These IP cells detach from the VZ and migrate basally along the basal projections of the aRG, which serve as scaffolds for the migrating IPs, and reside in an area of the cortex referred to as the subventricular zone (SVZ). aRG also give rise to outer radial glia (oRG), which lose their apical projections, and are believed to play a central role in the gyrification and expansion of the human and primate cortex (Hansen et al., 2010). Upon subsequent cell divisions, IPs eventually differentiate into neurons, and further migrate basally to a region of the neocortex known as the cortical plate (CP). Initially, an outer layer of neurons (Layer I) is formed consisting of Cajal-Retzius neurons, which assist in proper CP formation, as well as a layer of subplate neurons, which are essential in establishing the correct wiring and maturation of the cerebral cortex (Greig et al., 2013). In successive waves of neurogenesis, different neuronal subtypes are generated and organize into layers in a process known as cortical lamination, with the deepest layer (Layer VI) formed first, and the uppermost layer (Layer II) formed last. This process is depicted in Figure 2.7.

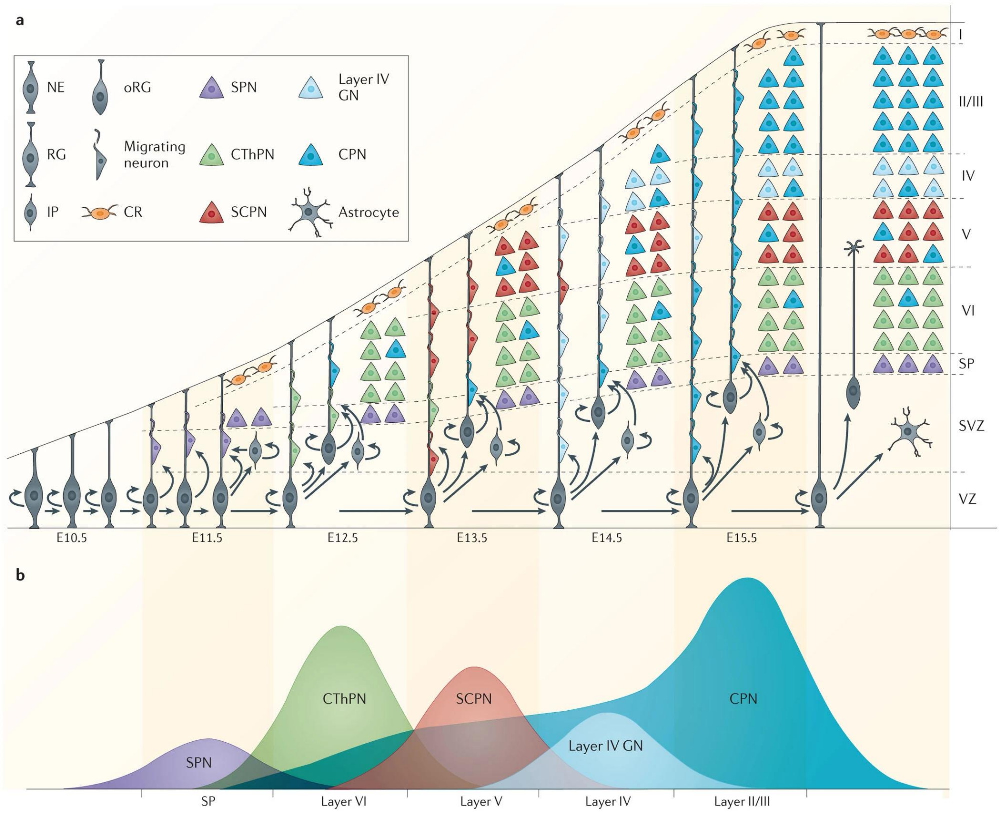

**Figure 2.7: Neocortical expansion and lamination in mouse neocortex. a.** Schematic representation of neocortical development in the mouse. Initially, neuroepithelial cells divide symmetrically, increasing the initial progenitor pool. Cortical neurons are then generated from apical radial glia (aRG) via intermediate progenitors (IP), which occupy the SVZ and migrate basally towards the cortical plate, giving rise to distinct neuronal layers in the neocortex. This is followed by a wave of gliogenesis, during which radial glia in the neocortex produce glial cells including astrocytes and oligondendrocytes. **b.** Schematic representation of the sequential waves of production of cortical projection neurons during coritcal neurogensis. NE - neuroepithelial cell; CR - Cajal-Retzius neuron; SPN - subplate neuron; CThPN - corticothalamic projection neuron; SCPN - subcerebral projection neuron; GN - granular neuron; CPN - callosal projection neuron. Reprinted from (Greig et al., 2013).

The neuronal layers of the neocortex perform specific functions, in particular in the wiring of the cortex, connecting it to other brain regions and connecting different regions within the cortex to each other. Layer VI neurons project their axons to the thalamus (corticothalamic projection neurons), while layer V neurons project their axons to subcerebral structures (subcerebral projection neurons) including the midbrain, pons, and spinal cord, among other regions (Greig et al., 2013). Upper layer neurons (Layers II/III, IV) mainly consist of callosal projection neurons, which project axons across the corpus callosum, connecting the two hemispheres of the neocortex (Fame et al., 2011). Neurons in different cortical layers are also defined molecularly by the expression of various transcription factors, which ultimately modify the expression levels of genes essential for the formation of the various subtypes of neurons. For example, deep layer neurons (Layers V and VI) are characterized by the expression of transcription factors TBR1, CTIP2 and FEZF2, and upper layer neurons are characterized by the expression of transcription factors CUX1, CUX2, and SATB2 (Greig et al., 2013; Molyneaux et al., 2007). This highlights the pivotal role of transcription factors in determining cell function even within cellular subtypes.

Furthermore, transcription factors play an essential role during the differentiation of aRG to IPs and IPs to neurons. This differentiation process is governed by an underlying gene regulatory network involving the transcription factors PAX6, EOMES and TBR1 (Elsen et al., 2018). PAX6 is highly expressed in aRG and directly upregulates EOMES by binding to its promoter (Sansom et al., 2009). EOMES, a gene specifically expressed in IPs, then represses PAX6 and consequently activates the expression of the pro-neuronal gene TBR1, which then signals the transition into postmitotic neurons (Sessa et al., 2016). Measuring the regulatory interactions between the genes important for these cell-state transitions will be discussed in detail in Chapter 6.

### 2.2.4 In vitro neural induction with brain organoids

Studying the molecularly mechanisms underlying brain development and neurogenesis in humans is difficult due to the scarcity of available embryonic tissues. Most of the findings described in the previous sections were derived from studying developing mouse or chick embryos. Thus, there is a widespread need to accurately model early human brain development in vitro in order to better understand the molecular underpinnings of human brain development both in health and disease.

In vitro modeling of the nervous system begins with the culturing of human pluripotent stem cells. The first embryonic stem cell (ESC) line was generated in 1998 by isolating cells from the ICM of human embryos produced by in vitro fertilization (IVF) (Thomson et al., 1998). These cells were able to give rise to all three germ layers after being injected into severe combined immunodeficient mice, indicating a pluripotent state. Furthermore, in 2006, the groundbreaking work of Yamanaka paved an alternative path to deriving PSCs, whereby the first induced PSCs (iPSC) were derived by introducing four transcription factors - Myc, Oct3/4, Sox2 and Klf4 - to mouse fibroblasts, reprogramming them into a pluripotent state (Takahashi et al., 2006). In the following year, this technique was successfully applied to human adult fibroblasts as well (Takahashi et al., 2007). Due to their ability to give rise to every cell type in the human body, human ESCs or iPSCs provide a powerful source to generate neural stem cells and their progeny at different stages of development, enabling the study of the development of these otherwise inaccessible tissues.

Neural induction methods from human and mouse PSCs have evolved dramatically over the past two decades (Kelava et al., 2016). The first derivation of human neural rosettes, structures resembling neural tube formation, from embryoid bodies derived from ESCs was performed in 2001 in the presence of FGF2 (Zhang et al., 2001). These neural rosette precursors were able to differentiate into neurons and glia including astrocytes and oligodendrocytes. Following this, mouse ESCs were differentiated into neural precursors in the absence of serum, growth factors or other inductive signals, indicating the intrinsic ability of PSCs to give rise to the neural lineage (Ying et al., 2003). These methods were further improved upon to derive more regionally specified neural cell types including telencephalic identities using Nodal and WNT pathway antagonists in mouse ESCs (Watanabe et al., 2005), with the addition of BMP antagonists in human ESCs (Elkabetz et al., 2008). This method was again improved upon by adding TGF$-\beta$ inhibition, commonly referred to as the Dual-SMAD inhibition protocol (Chambers et al., 2009). However, these protocols were somewhat limited in their ability to derive the full complexity of in vivo cell types in the developing brain due to an intermediate plating step, limiting the ability of the cells to migrate and expand into 3D space. The first entirely 3D neural culture was accomplished in 2011 with the generation of self-organizing optic cups from human PSCs in a 3D floating culture (Eiraku et al., 2011). This was later improved upon by embedding cells in a supportive extracellular matrix called Matrigel to provide structural support for the self-organization of the 3D tissues in a system known as organoids. Using this 3D organoid culture system, a number of human brain organoid derivation protocols were developed, ranging from Inhibitor-free conditions (Lancaster et al., 2013) to Dual-SMAD inhibition (Paşca et al., 2015; Qian et al., 2016), and TGF$-\beta$/WNT inhibition (Bershteyn et al., 2017; Kadoshima et al., 2013; Velasco et al., 2019).

Brain organoids are complex structures that exhibit high levels of variability, with different protocols giving rise to widely heterogeneous populations across cell line and within organoids derived from the same cell line. Nonetheless, it can be difficult to attribute this heterogeneity to inherent biases in the PSC lines, or to biases in the derivation protocol itself. The utility of a given protocol is largely reflected in its ability to generate consistent cell populations across a variety of cell lines, further enabling a systematic comparison of brain development under healthy and disease conditions, for example from patient-derived iPSC lines. Furthermore, brain organoids provide an unprecedented opportunity to study brain development in a controlled environment without the need for obtaining human embryonic samples, and are thus invaluable to the study of human neurodevelopment. In Chapter 4, we will describe a comparative study performed as a part of this thesis to measure this heterogeneity using transcriptomics.

Transcriptomics is the study of the set of RNA transcripts produced by the genome. With this technology, it is possible to measure the gene expression profile of a sample in bulk or at the single cell level. This enables a detailed view into the complexity of cell types arising in a brain organoid, and the underlying molecular components defining these cell types and their development. In the following section, we will give an overview of how this technology works.

## 2.3 EXPERIMENTAL TECHNIQUES TO MEASURE THE TRANSCRIPTOME

The field of transcriptomics has evolved substantially over the past three decades. The first high-throughput transcriptomic technologies used microarrays, microscope slides with DNA probes located at defined positions on the slide (Schena et al., 1995). These probes hybridize to fluorescently-labeled complementary DNA (cDNA) reverse transcribed from mRNA molecules derived from a sample, and gene expression levels are measured from the intensity of the fluorescent signal using a microscope. One of the major drawbacks of microarrays is that the DNA probes need to be defined a priori, which requires prior knowledge about the genome and limits the number of genes that can be measured on a microarray slide. The advent of next-generation sequencing (NGS) revolutionized the field, enabling high-throughput sequencing of the entire transcriptome without the need to define the target sequences a priori. NGS also relies on measuring fluorescent signals using a microscope, however, instead of labeling entire DNA molecules with a fluorescent tag, individual nucleotides are labeled and the signal is measured for each nucleotide in a process known as "sequencing by synthesis". Typically, a complementary strand of DNA is synthesized from a template strand using fluorescently-labeled nucleotides on a flow cell, such that each nucleotide (A, C, G or T) has a different fluorescent label. Upon iterative rounds of DNA synthesis, an image is captured and the emitted fluorescent signal from each nucleotide is measured. This process is done in parallel across the flow cell to produce hundreds of millions of reads, or sequenced DNA fragments. In the following sections, we will describe how this technology has been applied to the field transcriptomics.

### 2.3.1 RNA sequencing

RNA sequencing (RNA-Seq) begins with the generation of a cDNA sequencing library, which is prepared in general as follows. RNA is isolated from a tissue sample, following which further isolation steps may be performed such as poly(A) selection involving filtering for mRNA with a 3' poly-A tail using poly-T oligomers typically attached to a magnetic bead (Mortazavi et al., 2008). Following RNA isolation, cDNA is synthesized from the RNA molecules, which is then fragmented into short sequences by sonication or enzymatically, ligated with adapter sequences, and finally amplified using DNA polymerases in a process known as polymerase chain reaction (PCR) (Wang et al., 2009). This process typically generates hundreds of millions of cDNA fragments, which are then sequenced using a next-generation sequencer in either single-end sequencing, in which one end of the cDNA is sequenced, or paired-end sequencing, in which both ends of the cDNA are sequenced. The sequenced reads are then aligned to a reference genome using alignment algorithms such as STAR (Dobin et al., 2013) or directly to an annotated reference transcriptome with Bowtie2 (Langmead et al., 2012) among other tools, after which gene expression levels are quantified based on the number of reads aligning to each annotated gene in the reference, which can be done using RSEM (Li et al., 2011) or HTSeq (Anders et al., 2015), among others. An outline of the RNA-Seq procedure is illustrated in Figure 2.8.

Bulk RNA-Seq of an entire tissue provides a detailed landscape of the transcriptome. Further downstream applications include measuring the expression levels of different isoforms of a given gene, novel splicing events and variant detection, among others. Furthermore, it enables the detection of differentially expressed genes across different tissues or conditions with tools such as DESeq2 (Love et al., 2014) or limma (Ritchie et al., 2015). However, as seen in the developing cortex, a tissue is comprised of individual cells which can have different cell identities with vastly heterogeneous gene expression profiles. When performing bulk RNA-Seq of an entire tissue, the transcriptomes of the individual cells inside of the tissue are merged together, and the gene expression information at the individual cell level is lost. Over the past decade, this limitation has

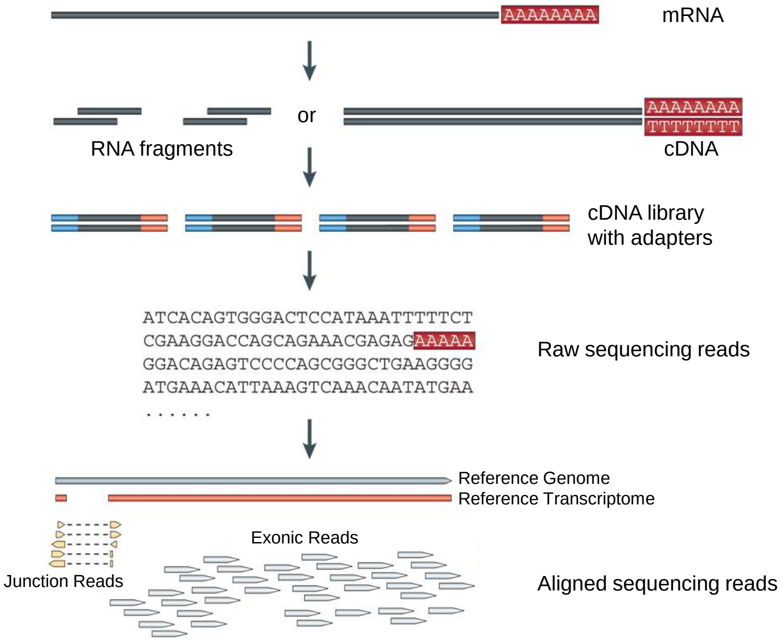

**Figure 2.8: Overview of RNA-Seq.** RNA-Seq begins with RNA isolation from a tissue. This is followed by fragmentation and generation of cDNA from the RNA template strand. Afterwards, adapters are ligated to the cDNA and amplified using PCR to generate a sequencing library. Sequencing with NGS produces short reads of typically 50-150bp in length which are then aligned to a reference genome or transcriptome. Adapted from (Wang et al., 2009).

been overcome with the advent of single-cell RNA-Sequencing, which enables the transcriptomic profiling of individual cells on a genomic scale.

## 2.3.2 Single-cell RNA-Sequencing

The first single-cell RNA-Sequencing (scRNA-Seq) paper was published in 2009 (Tang et al., 2009), and since then, there has been a dramatic increase in the number of experimental protocols to generate scRNA-Seq data including Smart-seq2 (Picelli et al., 2013), MARS-Seq (Jaitin et al., 2014), inDrops (Klein et al., 2015), Drop-Seq (Macosko et al., 2015), CEL-Seq2 (Hashimshony et al., 2016) and 10X Genomics (Zheng et al., 2017), among others. These technologies have been used to identify rare populations which would have otherwise been missed using bulk RNA-Seq (Aizarani et al., 2019; Montoro et al., 2018; Plasschaert et al., 2018), build single cell atlases of the transcriptomes of all cells in individual organs, as well as the entire organism, across different species (Regev et al., 2017; The Tabula Muris Consortium et al., 2018), and trace lineage and developmental relationships during diverse biological processes such as embryonic development (Blakeley et al., 2015), lung epithelium differentiation (Treutlein et al., 2014), and cancer (Tirosh et al., 2016). In this thesis, we use scRNA-Seq to measure regional and cellular heterogeneity in brain organoids and model gene expression dynamics in differentiating cortical neural stem cells in brain organoids, in particular using the inDrops and 10X Genomics protocols. Therefore, in this section, we give a general overview of this technology, in particular highlighting these two platforms.

The generation of a scRNA-Seq library requires keeping track of which cell each mRNA molecule was derived from. This is done by adding cell barcodes, short oligonucleotide sequences that are unique to individual cells, to every cDNA molecule derived from an mRNA transcript within each cell. These cell barcodes are then read during NGS, and mRNA transcripts are assigned to individual cells according to their respective cell barcode. To achieve this, scRNA-Seq platforms use a microfluidic device, whereby individual cells are encapsulated into droplets, along with beads (in the 10X Genomics and inDrops protocols, these beads are made of hydrogel) which contain primers, with every primer in each bead having the same cell barcode. These primers also contain a poly-T oligomer which is designed to bind to the poly-A tail of an mRNA molecule in order to specifically capture mRNA, a unique molecular identifier (UMI), which is a random oligonucleotide sequence used to remove PCR duplicates, and a PCR primer used to initiate PCR amplification. Beads and cells are then loaded onto the microfluidic device in separate channels, and encapsulated into water-in-oil droplets along with reagents to initiate lysis (breaking down of the cell membrane) and reverse transcription of RNA to cDNA. The beads and cells are introduced at low concentrations to reduce the chances of two beads or two cells entering an individual droplet. The cells are lysed, enabling the capture of released mRNA molecules via hybridization of the mRNA poly-A tails to the poly-T oligomers on each primer, and reverse transcription is carried out within individual droplets. Following reverse transcription, droplets are demulsified, and the single-cell barcoded cDNA material is amplified (using in vitro transcription followed by reverse transcription PCR (RT-PCR) in the case of inDrops, and using PCR in the case of 10X Genomics), fragmented, and sequencing adapters are ligated (i.e. P5 and P7 adapters for Illumina sequencing), followed by further rounds of amplification, and finally sequencing. An overview of the general procedure is highlighted in Figure 2.9a.

In both inDrops and 10X, paired-end sequencing is performed, such that one end (read 1) contains the cell barcode and UMI, and the other end (read 2) contains the mRNA transcript sequence. After sequencing, a bioinformatics pipeline is run to generate a count matrix containing the number of observed transcripts in each cell. Initially, read pairs containing the same cell barcode are aggregated in a process known as demultiplexing. The inDrops and 10X protocols generate cell barcodes which are not completely random, and read pairs containing a cell barcode not present in the manufacturers' whitelists are filtered. Errors may be present in the cell barcode due to sequencing errors or errors arising during DNA synthesis, and therefore a cell barcode correction step is run to correct for this, typically allowing for a 1-bp mismatch (Hamming distance 1) between the sequenced cell barcode and most similar cell barcode from the whitelist. The read containing the transcript sequence is then aligned to a reference genome/transcriptome and assigned to a gene. Following this, read pairs with the same cell barcode and same UMI that map to the same gene in the reference are merged so that they only contribute a single count towards that gene in the given cell, since these are assumed to originate from the same mRNA molecule, resulting in sequencing of PCR duplicates. This procedure of demultiplexing and UMI counting after sequencing is illustrated in Figure 2.9b.

This procedure results in a cell by gene count matrix, which is then further processed in downstream analyses. The experimental procedure to produce a scRNA-Seq dataset is complex,

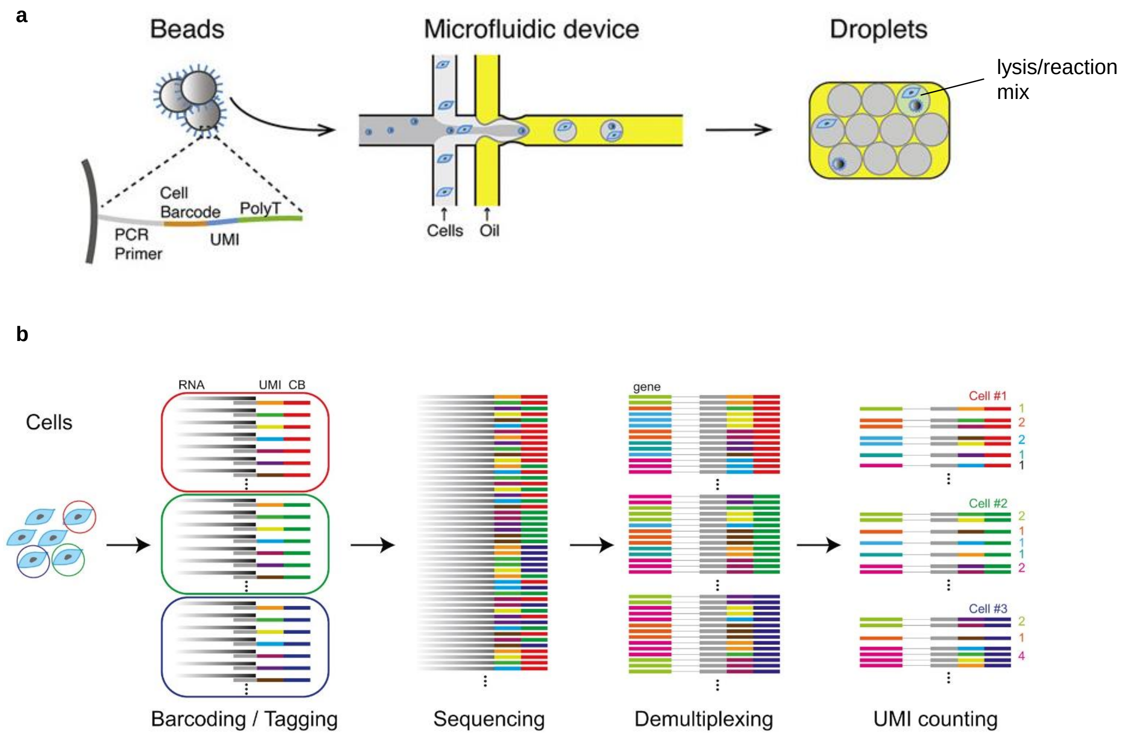

**Figure 2.9: Overview of scRNA-Seq. a.** Illustration of the protocol for generating a scRNA-Seq library. Cells are loaded onto a microfluidic device along with beads containing primers consisting of a poly-T oligomer, UMI, cell barcode and PCR primers. Beads and cells are encapsulated in oil-to-water droplets, in which mRNA molecules are captured, from which cDNA is synthesized and subsequently amplified, and finally sequenced using NGS. **b.** Illustration of the data processing workflow of scRNA-Seq data. After sequencing, read pairs with the same cell barcode are demultiplexed, and the number of transcripts sequenced in each cell is counted after collapsing reads pairs mapping to the same gene with the same UMI and cell barcode. Adapted from (Zhang et al., 2019).

and many technical artifacts may arise during library generation. In Chapter 5, we present a tool developed as a part of this thesis work to address one such error, which involves the detection and removal of empty droplets which contain mainly free-floating mRNA from the sample, or ambient mRNA. We also explore a method to measure transcriptional dynamics during the differentiation of cortical neural stem cells into neurons within brain organoids in Chapter 6, and using scRNA-Seq to measure the regional heterogeneity arising in brain organoids across different derivation protocols in Chapter 4. However, before we explore these applications of scRNA-Seq, we provide a comprehensive background into the mathematical concepts used in these methods, as well as those used in the analysis of scRNA-Seq data in general.

# 3 COMPUTATIONAL BACKGROUND

In this chapter, we explore some fundamental concepts which are used throughout this work, including dimension reduction techniques, clustering, and Bayesian inference. In particular, we highlight the utility of these concepts in the analysis of scRNA-Seq data. The content of this chapter mainly draws from the book of Christopher Bishop (2006).

## 3.1 LEARNING PROBABILISTIC MODELS

Probabilistic models incorporate randomness to predict outcomes of a certain event. In a biological context, or any physical context for that matter, we typically gather data and would like to make generalizations about the underlying process or phenomenon, under the assumption that the process is stochastic in nature, by fitting the data to a probabilistic model.

### 3.1.1 Maximum likelihood estimation

To begin, we first define $\mathcal{D} = \{x_1, x_2, ..., x_N\} = \{x_i\}_{i=1}^N$ as a set of $N$ independent identically distributed (i.i.d.) observations of a given stochastic process. For example, assume $x \in \{0, 1\}$ describes the outcome of flipping a coin, with $x = 1$ representing heads, and $x = 0$ representing tails. The flips are independent because the outcome of one flip does not influence the outcome of another flip, and identically distributed because each flip has the same probability of flipping a head. This process represents a Bernoulli process, and can be parameterized with the probability of flipping a head, $p(x = 1) = \theta$. The probability distribution over $x$ can be written as

$$\text{Bernoulli}(x|\theta) = \theta^x(1-\theta)^{1-x}. \tag{3.1}$$

This describes the probability distribution for a single trial, $x$. We can also work out the distribution of the number of heads, $m$, in a dataset of size $N$, as follows. This requires adding a normalization coefficient which sums up all the possible ways of obtaining $m$ heads, and is accomplished using the binomial distribution, defined as

$$\text{Binom}(m|N,\theta) = \binom{N}{m}\theta^m(1-\theta)^{N-m}, \tag{3.2}$$

where

$$\binom{N}{m} = \frac{N!}{(N-m)!m!} \tag{3.3}$$

is the number of ways of choosing $m$ objects out of $N$ total objects.

The likelihood function, $\mathcal{L}(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)$, is then defined as the probability of the observed data under the assumption that the observations are drawn from the underlying probability model, in the case above, that the observations are drawn independently from a Bernoulli process with $p(x = 1) = \theta$. The likelihood function for this example is

$$\mathcal{L}(\theta|\mathcal{D}) = p(\mathcal{D}|\theta) = \binom{N}{m}\theta^m(1-\theta)^{N-m}, \tag{3.4}$$

which is the Binomial evaluated at $\theta$ given the data $\mathcal{D}$. The likelihood function is not a probability distribution, but rather describes the relative likelihoods for different values of the parameter $\theta$ given the data. In most cases when there are many observations, the likelihood function will take on very small values. Therefore, it is often convenient to work with the log-likelihood function, $\ln(\mathcal{L}(\theta|\mathcal{D}))$, where ln represents the natural logarithm, which is a monotonic increasing function of $\mathcal{L}(\theta|\mathcal{D})$. In the example above, we assume $\theta$ is unknown a priori, and would like to solve for a reasonable estimate of $\theta$. One approach to do so is to use maximum likelihood estimation (MLE), which solves for the value of the parameter $\theta$, $\hat{\theta}_{\text{ML}}$, that maximizes the likelihood function $\mathcal{L}(\theta|\mathcal{D})$. This can be written as

$$\hat{\theta}_{\text{ML}} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta|\mathcal{D}). \tag{3.5}$$

In other words, the maximum likelihood estimate solves for the parameters of the statistical model under which the observed data is most probable. Since the log-likelihood function is a monotonic increasing function of the likelihood function, the value for $\theta$ which maximizes the log-likelihood function will also maximize the likelihood function. In some cases, solving for the maximum likelihood estimate can be done analytically by computing the derivative of the log-likelihood function with respect to $\theta$, and setting it to zero. For the coin-flipping example, this can be accomplished as follows,

$$\begin{aligned} \frac{\mathrm{d}\ln(\mathcal{L}(\theta|\mathcal{D}))}{\mathrm{d}\theta} &= \frac{\mathrm{d}}{\mathrm{d}\theta}\ln\left(\binom{N}{m}\theta^m(1-\theta)^{N-m}\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}\theta}\ln\binom{N}{m} + \frac{\mathrm{d}}{\mathrm{d}\theta}\ln\left(\theta^m\right) + \frac{\mathrm{d}}{\mathrm{d}\theta}\ln\left((1-\theta)^{N-m}\right) \\ &= \frac{m}{\theta} - \frac{N-m}{1-\theta}. \end{aligned} \tag{3.6}$$

Setting this equation to zero and solving for $\theta$ gives the following maximum likelihood estimate,

$$\hat{\theta}_{\text{ML}} = \frac{m}{N}, \tag{3.7}$$

which is also the sample mean, in this case the fraction of observed heads in the dataset.

While MLE is both intuitive and flexible, it has some major pitfalls. Notably, if the number of observations is small, then the maximum likelihood estimate has a tendency to be heavily biased. For example, if we observe only 3 coin flips in the above example and they are all heads, then the maximum likelihood estimate, $\hat{\theta}_{\text{ML}}$, will be 1, which is most likely not the case. This is an example of extreme overfitting, which occurs when a model is fit too closely to a particular dataset, and therefore becomes unreliable when making predictions about future observations. There may also be multiple local maxima across the parameter space. MLE will pick the solution which maximizes the likelihood globally, and in so doing, we lose information about the certainty of the estimate. However, uncertainty of the maximum likelihood estimate can be estimated by measuring a theoretical lower bound on the variance of the estimator, i.e. with a Cramér–Rao bound (Cramér, 1946). Furthermore, for more complex probabilistic models, the analytical derivation of the maximum likelihood estimate can be difficult, if not impossible. For these cases, it is necessary to use an optimization algorithm, such as gradient descent or stochastic gradient descent. Finally, MLE does not incorporate prior information about the distribution over $\theta$. In the next section, we discuss how to include prior information in parameter inference methods.

### 3.1.2 Bayesian parameter estimation

In the maximum likelihood approach, the observations $\mathcal{D} = \{x_i\}_{i=1}^N$ were assumed to be random variables, however, the model parameter, $\theta$, was estimated as a point estimate. In the Bayesian approach, the model parameters are also treated as random variables. In order to do this, we first introduce Bayes' theorem,

$$p(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})}. \tag{3.8}$$

This equation incorporates the evidence provided by the observed data to convert a prior probability (i.e. prior beliefs about what model parameters should be before observing the data) into a posterior probability (i.e. what the model parameters should be after observing the data). In the above equation, $p(\theta)$ encodes the prior distribution and $p(\theta|\mathcal{D})$ encodes the posterior distribution. As seen in Equation 3.4, $p(\mathcal{D}|\theta)$ encodes the likelihood, that is the probability of the observed data given the parameters. The denominator of Equation 3.8, $p(\mathcal{D})$, also referred to as the evidence, can also be expressed in terms of the likelihood function and prior distribution, by marginalizing (integrating) out $\theta$ as follows,

$$p(\mathcal{D}) = \int p(\theta)p(\mathcal{D}|\theta)d\theta. \tag{3.9}$$

This quantity can be viewed as a normalization constant ensuring that the posterior distribution is a proper distribution (i.e. sums to 1), and is typically ignored when measuring which values of $\theta$ are more probable, due to difficulties in computing this integral. Therefore, Bayes' theorem can also be written more succinctly, with $\propto$ meaning 'is proportional to', as

$$p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta). \tag{3.10}$$

The Bayesian approach involves calculating the posterior distribution, $p(\theta|\mathcal{D})$, however, in order to do this, we first need to specify a prior distribution, $p(\theta)$. For the coin-flipping example, one could simply choose a noninformative prior, which assumes as little prior information as possible about the distribution of $\theta$, such as a uniform distribution, where $p(\theta) = 1$ for $\theta \in [0, 1]$. Unfortunately, there is no single recipe for choosing an optimal prior distribution. In many cases, priors are chosen for mathematical convenience rather than as a reflection of any prior beliefs. However, a reasonable choice of prior distribution will accurately inform the estimation of the posterior, and in cases where prior information is not available, a noninformative prior can be used.

In the coin flipping case presented in Section 3.1.1, based on prior information that most coins are fair (i.e. $p(x_i = 1) = 0.5$), a prior distribution should make $\theta = 0.5$ more likely. The beta distribution is particular useful for this purpose, and is defined as

$$\text{Beta}(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \tag{3.11}$$

where $a > 0$, $b > 0$, $\theta \in [0, 1]$, and $\Gamma(x)$ is the gamma function and is defined by

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du. \tag{3.12}$$

The mean of the beta distribution is $\frac{a}{a+b}$, so the hyperparameters should be specified such that $a = b$, if they reflect a prior belief that the coin is fair. The beta distribution has the nice property of being a conjugate prior to the binomial, since it has the same functional form as the likelihood function of the binomial distribution defined in Equation 3.2, containing powers of $\theta$ and $(1-\theta)$. The parameters $a$ and $b$ are called hyperparameters, because they influence the distribution of the parameter $\theta$. The posterior distribution can now be calculated by multiplying the beta prior distribution in Equation 3.11 by the binomial likelihood function in Equation 3.2. Keeping only terms which depend on $\theta$ gives

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\theta)p(\mathcal{D}|\theta) \\ &\propto \theta^{a-1}(1-\theta)^{b-1}\theta^m(1-\theta)^{N-m} \\ &= \theta^{m+a-1}(1-\theta)^{N-m+b-1}. \end{aligned} \tag{3.13}$$

We can also derive a point estimate for $\theta$ from the posterior distribution using the maximum a-posterior (MAP) estimate, which chooses the parameters which are most likely given the posterior distribution. Solving for the MAP does not involve having to calculate the posterior distribution explicitly, and can be simplified using Bayes' theorem as follows,

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \underset{\theta}{\operatorname{argmax}}\, p(\theta|\mathcal{D}) \\ &= \underset{\theta}{\operatorname{argmax}}\, p(\theta)p(\mathcal{D}|\theta) \\ &= \underset{\theta}{\operatorname{argmax}}\, (\ln p(\theta) + \ln p(\mathcal{D}|\theta))\,. \end{aligned} \tag{3.14}$$

The MAP estimate can be solved similarly to the MLE, namely by taking the derivative of $\ln p(\theta) + \ln p(\mathcal{D}|\theta)$ with respect to $\theta$, and setting it equal to zero. When doing so, and solving for $\theta$, we get

$$\hat{\theta}_{\text{MAP}} = \frac{m+a-1}{N+b+a-2}. \tag{3.15}$$

Note the similarity in Equation 3.14 between MLE and MAP estimation. While MLE maximizes $\ln p(\mathcal{D}|\theta)$ over $\theta$, MAP includes the prior term $\ln p(\theta)$ in the maximization. If a uniform prior is used over $\theta$, then MAP and MLE are equivalent. However, MLE and MAP both treat $\hat{\theta}$ as a single fixed value, and are thus considered point estimators. On the other hand, Bayesian inference explicitly calculates the posterior probability distribution. The posterior distribution in Equation 3.13 has the form of a beta distribution with parameters $m + a$ and $N - m + b$. This highlights the utility of using conjugate priors, making it unnecessary to calculate the integral, or normalizing factor, in Equation 3.9. So, the posterior distribution will have the form

$$p(\theta|\mathcal{D}) = p(\theta|N, m, a, b) = \frac{\Gamma(N+m+a)}{\Gamma(m+a)\Gamma(N-m+b)}\theta^{m+a-1}(1-\theta)^{N-m+b-1}. \tag{3.16}$$

Explicitly solving for the posterior distribution provides much more information than MLE or MAP, which only provide point estimates. As an example, Figure 3.1 highlights two examples showing the prior distribution, likelihood function, and posterior distribution, and highlights the influence of the prior on the posterior. For a small set of observations on the left, the prior has a high influence on the posterior, while for a larger set of observations on the right, the posterior is minimally affected by the prior.

If we want to predict the next outcome of a trial, we need to evaluate the posterior predictive distribution of $x$ given the total observed data, $\mathcal{D}$. This can be accomplished as follows,

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1|\theta)p(\theta|\mathcal{D})d\theta = \int_0^1 \theta p(\theta|\mathcal{D})d\theta = \mathbb{E}[\theta|\mathcal{D}]. \tag{3.17}$$

This is simply the mean of the posterior distribution, which follows a beta distribution with parameters $m + a$ and $N - m + b$. Therefore,

$$p(x=1|\mathcal{D}) = \frac{m+a}{N+a+b}. \tag{3.18}$$

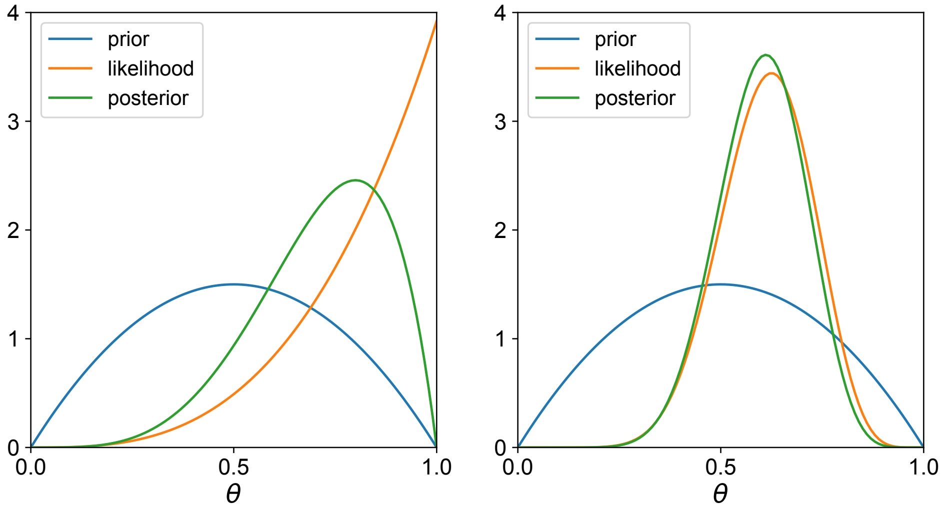

**Figure 3.1: Bayesian inference.** Plots of the prior, likelihood, and posterior for the coin flipping example. In both cases, a prior distribution of Beta$(2, 2)$ was used, which has a mean of $0.5$, reflecting a prior belief that coins are typically fair. For the left plot, the number of observations, $N$, was set to 3, and number of heads, $m$, was set to 3. For the right plot, $N = 16, m = 10$. Note that the likelihood function is normalized to integrate to 1 for visualization purposes.

As the number of observations increases to $\infty$, this term reduces to the the maximum likelihood estimate from Equation 3.7 to be $m/N$, as does the MAP estimate from Equation 3.15. This is a general property of Bayesian inference and MLE. However, full Bayesian inference provides the entire posterior probability distribution over the parameter space, which is critical when measuring how confident we are in a parameter estimate. Unfortunately, for the Bayesian approach, we need to compute an integral to marginalize out the model parameters to calculate $p(\mathcal{D})$ in Equation 3.9, which becomes intractable when the models are more complex and the number of parameters becomes large. One approach to do this uses a sampling based approach called Markov chain Monte Carlo (MCMC), which we will discuss in the following section.

### 3.1.3 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods have their origins in solving differential equations arising in physics (Metropolis et al., 1949). The general aim of MCMC methods are to approximate a probability distribution by generating random samples from it. In the following sections, we give a brief introduction to the theory behind MCMC, and various MCMC methods that have been developed.

### Monte Carlo techniques

Monte Carlo techniques use repeated random sampling from a target probability density, $p(x)$, to make statistical approximations about the distribution. They come in different flavors, with the most widespread being importance sampling and rejection sampling. Here, we briefly explain rejection sampling, as it plays an important role in MCMC algorithms. For further reading of sampling techniques, we refer the reader to Robert et al. (1999).

The rejection sampling framework allows sampling from complex distributions, $p(z)$, where direct sampling from $p(z)$ is difficult, but we are able to evaluate $p(z)$ for any given value of $z$ up to a normalizing constant $Z$. In the context of Bayesian inference, $p(z)$ represents the posterior distribution $p(\theta|\mathcal{D})$, and the normalizing constant $Z$ represents $p(\mathcal{D})$ from Equation 3.8. That is,

$$p(z) = \frac{1}{Z}\tilde{p}(z), \tag{3.19}$$

where $\tilde{p}(z)$ can be evaluated, but $Z$ is unknown. For rejection sampling, we need to specify a proposal distribution, $q(z)$, which is a simpler distribution from which we can draw random samples, and satisfies

$$Mq(z) \geq \tilde{p}(z), \tag{3.20}$$

where $1 < M < \infty$. The rejection sampler requires generating two random numbers, a number $z_0$ from the distribution $q(z)$, and a number $u_0$ from the uniform distribution over $[0, Mq(z_0)]$. If $u_0 > \tilde{p}(z_0)$, then the sample is rejected, otherwise it is accepted. The corresponding $z$ values of the accepted samples are then distributed according to $p(z)$, since they are uniformly distributed under the curve $\tilde{p}(z)$. The accept-reject random sampling procedure is highlighted in Figure 3.2.

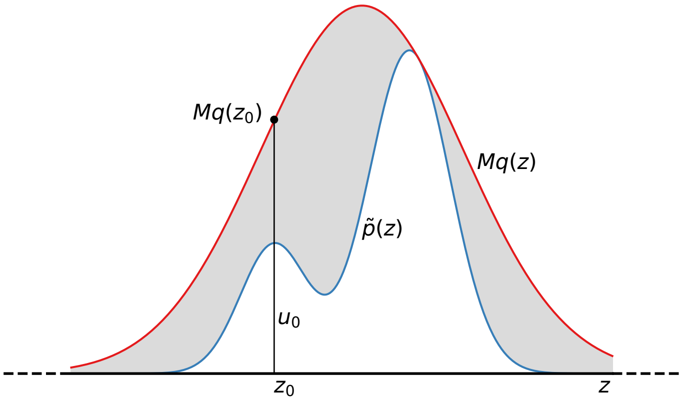

**Figure 3.2: Rejection sampling.** In rejection sampling, a sample $z_0$ is drawn from the distribution $q(z)$, and a sample $u_0$ is drawn from the uniform distribution $U[0, Mq(z_0)]$. If $u_0 > \tilde{p}(z_0)$ (gray area in the plot), reject $u_0$, otherwise accept $u_0$. The resulting accepted samples are distributed according to $p(z)$.

The values $z_0$ are accepted with probability $\tilde{p}(z)/Mq(z)$, and the probability that a sample at random will be accepted is,

$$p(\text{accept}) = \int \frac{\tilde{p}(z)}{Mq(z)}q(z)dz = \frac{1}{M}\int \tilde{p}(z)dz. \tag{3.21}$$

Thus, if $M$ is very large, the probability of accepting a random draw will be small, and many draws will essentially be wasted when computing $p(z)$. However, an efficient proposal distribution, $q(z)$, is difficult to find. MCMC methods use the idea of rejection sampling as a part of its algorithm, however, instead of having a fixed proposal distribution generating i.i.d. samples, the proposal distribution is updated at each iteration based on an underlying Markov chain. In the next section, we describe the basics of Markov chains, and move on to MCMC methods in the following section.

## Markov Chain

A Markov chain is a series of random variables $z^{(0)}, ..., z^{(M)}$ describing a sequence of possible events such that the probability of an event is only dependent on the immediate past, also referred to as the "memoryless property". That is, for $m \in 0, ..., M-1$,

$$p(z^{(m+1)}|z^{(0)}, ..., z^{(m)}) = p(z^{(m+1)}|z^{(m)}). \tag{3.22}$$

This condition is known as the Markov property. A Markov chain is then specified by two probability distributions, the initial probability distribution and transition probabilities. The initial probability distribution describes the initial variable $p(z^{(0)})$, and the transition probability, or transition kernel, describes the probabilities of transitioning from one state to another, namely $T_m(z^{(m)}, z^{(m+1)}) = p(z^{(m+1)}|z^{(m)})$. A Markov chain is homogeneous if the transition probability is the same for all $m$, that is, $T_m(z^{(m)}, z^{(m+1)}) = T(z^{(m)}, z^{(m+1)})$.

For the discrete case, when there are a discrete number of possible states, we can specify a transition matrix, $T$, where the $(i, j)^{\text{th}}$-entry in the matrix $T$ is simply the transition probability of state $i$ to state $j$, i.e. $T_{ij} = T(z_i^{(m)}, z_j^{(m+1)}) = p(z_j^{(m+1)}|z_i^{(m)})$. As an example, assume we have a Markov process with 3 states representing different weather conditions on a particular day, with states $(sunny, cloudy, rainy)$. Assume the transition matrix is specified as

$$T = \begin{pmatrix} 0.4 & 0.4 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}. \tag{3.23}$$

With this formulation, the marginal probability of $p(z^{(m+1)})$ can then be written as follows,

$$p(z^{(m+1)}) = p(z^{(m)})T, \tag{3.24}$$

where $p(z^{(m)})$ is a row vector containing the probability distribution of $z^{(m)}$, in this case the probability of the weather conditions on a given day. The transition matrix can also be visualized as a diagram, as shown in Figure 3.3.

If we initialize $p(z^{(0)}) = (0.5, 0.2, 0.3)$, then the probability distribution of the weather conditions for the next day is $p(z^{(1)}) = p(z^{(0)})T = (0.3, 0.44, 0.26)$, and the following day is $p(z^{(2)}) =$

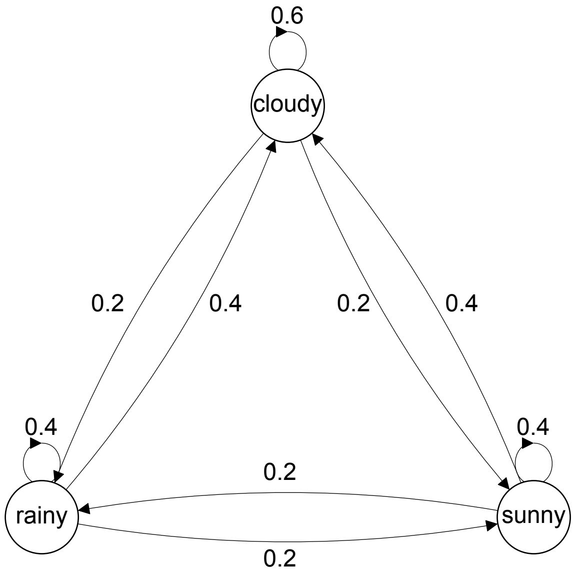

**Figure 3.3: Markov process.** Diagram of a Markov chain with 3 discrete states, $(sunny, cloudy, rainy)$. The numbers represent the transition probabilities from one state to another, with the direction indicated by the arrow. For example, if it is cloudy one day, the probability the next day will be cloudy is 0.6, and the probability the next day will be rainy or sunny are both 0.2.

$p(z^{(1)})T = p(z^{(0)})T^2 = (0.28, 0.488, 0.252)$. Upon further iterations, the probability distribution converges to the invariant, or stationary, distribution, $\pi = (0.25, 0.5, 0.25)$, which has the following property,

$$\pi = \pi T. \tag{3.25}$$

The row vector $\pi$ corresponds to the left eigenvector of $T$, with eigenvalue 1. For an invariant distribution, further steps in the Markov chain leave the distribution unchanged. In this example, no matter what the initial distribution $p(z^{(0)})$ is, the Markov chain will converge to $\pi$. This stability will play an important role in MCMC simulations, which we discuss later in this section. For a Markov chain to have a unique invariant distribution, the Markov chain must be irreducible (i.e. every state can eventually be reached from every other state) and aperiodic (i.e. the Markov chain does not get trapped in cycles) (Serfozo, 2009). Such Markov chains are said to be ergodic. For example, if the transition probabilities are defined by $p(sunny|rainy) = 1$, $p(rainy|cloudy) = 1$ and $p(cloudy|sunny) = 1$, then the chain can return to a given state only at multiples of 3, making it a periodic Markov chain.

A sufficient, but not necessary, condition to ensure a target probability distribution, $\pi$, is invariant under the Markov chain is to choose the transition probabilities such that they satisfy the property of detailed balance, that is

$$\pi_i T_{ij} = \pi_j T_{ji}. \tag{3.26}$$

A Markov chain satisfying detailed balance is called reversible, since reversing the dynamics leads to the same chain. For the case of continuous variables, the transition matrix $T$ becomes a probability density, and the detailed balance condition can be written as

$$\pi(x)T(x,x') = \pi(x')T(x',x), \tag{3.27}$$

where $T(x,x') = p(x'|x)$ and $\pi(x)$ is the invariant distribution, where $\pi(z) = \int \pi(x)T(x,z)dx$. The goal of MCMC is to use ergodic Markov chains to sample from a given distribution, with the target distribution defined as the invariant distribution of the Markov chain. That is, as $m \to \infty$, the distribution $p(z^{(m)})$ converges to the target distribution $\pi(z)$, irrespective of the choice of $p(z^{(0)})$. Most MCMC algorithms accomplish this by ensuring that detailed balance is satisfied.

## Metropolis–Hastings algorithm

In this section, we briefly describe one of the most popular MCMC methods, the Metropolis–Hastings algorithm (Hastings, 1970), in order to provide some intuition to how this class of algorithms works. We recall that the ultimate goal of Bayesian inference is to estimate the posterior distribution over a set of parameters for a given probabilistic model given a set of observations. Let $\pi(x)$ be the posterior distribution we want to calculate. We then want to establish a Markov chain with $\pi(x)$ as its stationary distribution. To do so, ensure detailed balance is satisfied in the Markov chain using Equation 3.27, that is,

$$\pi(x)p(x'|x) = \pi(x')p(x|x') \implies \frac{p(x'|x)}{p(x|x')} = \frac{\pi(x')}{\pi(x)}. \tag{3.28}$$

The Monte Carlo part of the algorithm is then established by defining a random sampling accept-reject procedure as follows,

$$p(x'|x) = q(x'|x)A(x'|x), \tag{3.29}$$

where $q(x'|x)$ is the proposal distribution and $A(x'|x)$ is the acceptance distribution. Plugging this into Equation 3.28 gives

$$\frac{A(x'|x)}{A(x|x')} = \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}. \tag{3.30}$$

The acceptance distribution is then defined to ensure detailed balance holds, by setting

$$A(x|x') = \min\left(1, \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}\right). \tag{3.31}$$

Normally, we can only calculate the posterior probability $\pi(x)$ up to a given normalization factor, as in Equation 3.10 (i.e. $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$), however, due to the terms $\pi(x')$ in the numerator and $\pi(x)$ in the denominator of Equation 3.31, this is sufficient for calculating the acceptance distribution, since the normalization factor cancels out. Thus, using this formulation of the acceptance distribution, an ergodic Markov chain is established with the stationary distribution of interest, and individual iterations in the MCMC algorithm can be considered as a random walk through the parameter space. The Metropolis–Hastings algorithm can be summarized as follows:

**Algorithm 1:** Metropolis–Hastings algorithm.

```
Initialize $x^{(1)}$.
for $i = 1, ..., N$ do
  Sample $u \sim U[0, 1]$.
  Sample $x^* \sim q(x^*|x^{(i)})$.
  if $u < A(x^*|x^{(i)}) = \min\left(1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})}\right)$ then
    $x^{(i+1)} \leftarrow x^*$,
  else
    $x^{(i+1)} \leftarrow x^{(i)}$.
  end if
end for
```

If $q(x|x')$ is symmetric, as in the case for a Gaussian or uniform proposal distribution, then $q(x^{(i)}|x^*) = q(x^*|x^{(i)})$, and these terms cancel out in the acceptance distribution. This then becomes the Metropolis algorithm (Metropolis et al., 1953). The ability of the Metropolis–Hastings algorithm to accurately estimate $p(x)$ strongly depends on the choice of proposal distribution. If the proposal distribution is too narrow (i.e. low variance), then the Markov chain can get stuck at a single mode of the distribution, while if the proposal distribution is too wide (i.e. high variance), then the proposals will often jump to regions of low probability, leading to a low percentage of accepted proposals, as highlighted in Figure 3.4. This problem becomes even more difficult to handle when the number of dimensions in the parameter space increases, especially when these parameters are correlated. This problem is often referred to as the “curse of dimensionality”. Furthermore, if the initialization of $x_0$ is far away from the stationary distribution, then it can take a large number of iterations before the Markov chain reaches a region in the parameter space around the stationary distribution. Thus, a number of initial iterations is often discarded, also called the “burn-in”. For a further overview of other MCMC algorithms designed to overcome these limitations, we refer the reader to Andrieu et al. (2003). Specifically, we will describe another approach using ensemble samplers in the following section, which does not require the explicit specification of a proposal distribution, and will be utilized in Chapter 6 of this thesis.

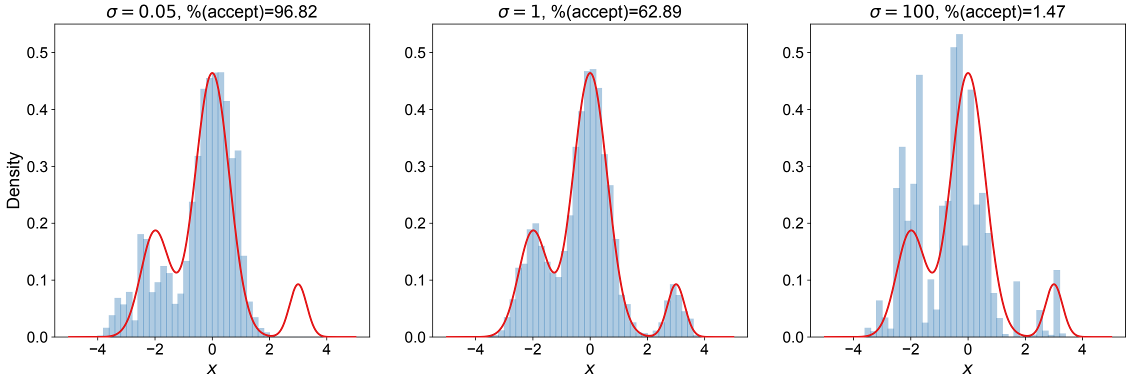

**Figure 3.4: Metropolis–Hastings examples.** The target distribution is shown in red and histograms of results using the Metropolis–Hastings algorithm are shown in blue. For all three examples, $x_0$ was initialized to 0, and $N = 10,000$ iterations was used. The left-most plot shows an example for a Gaussian proposal distribution with mean 0 and standard deviation 0.05, the middle plot with mean 0 and standard deviation 1, and the right-most plot with mean 0 and standard deviation 100. The percentage of accepted proposal jumps is displayed in the title. Note, the proposal distribution in these examples giving the best results uses a standard deviation of 1, however, an optimal proposal distribution is not generally known a priori.

## Ensemble samplers

Ensemble samplers describe a class of MCMC algorithms which run many random-walkers in parallel to automatically generate a proposal distribution. These methods make use of the set of independent walkers in the parameter space to modify the step-sizes and directions for new proposals, without specifying the proposal distribution a priori. The ensemble then expands or shrinks to fill the appropriate volume in each dimension of the parameter space to accurately sample the posterior distribution. Here, we describe an affine-invariant ensemble sampler (Goodman et al., 2010), which we utilize in Chapter 6 to model transcriptional dynamics in scRNA-Seq data.

The method from (Goodman et al., 2010) involves simultaneously running an ensemble of $K$ walkers, $S = \{X_k\}$. At each iteration, the proposal distribution for one walker, $k$, is based on the current positions of the $K - 1$ walkers in the complementary set, $S_{[k]} = \{X_j, j \neq k\}$. For a given walker $k$, a new position is proposed by a “stretch move”, defined by

$$X_k^* = X_j + Z\left(X_k^{(i)} - X_j\right), \tag{3.32}$$

where $X_k^{(i)}$ is the current position of walker $k$, $X_j$ is the current position of walker $j$ in $S_{[k]}$, and $Z$ is a random variable drawn from a distribution $g(z)$ defined by

$$g(z) = \begin{cases} \frac{1}{\sqrt{z}} & \text{if } z \in \left[\frac{1}{a}, a\right], \\ 0 & \text{otherwise,} \end{cases} \tag{3.33}$$

where $a > 1$ is an adjustable scale parameter, and typically set to 2. This choice of $g(z)$ ensures the proposal distribution is symmetric, and therefore will cancel out in the calculation of the acceptance distribution. An illustration of the move in Equation 3.32 is highlighted in Figure 3.5.

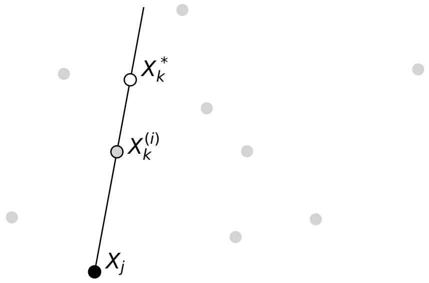

**Figure 3.5: Ensemble sampler stretch move.** A potential move is generated by stretching along the straight line connecting $X_k^{(i)}$ and $X_j$. Based on figure from (Goodman et al., 2010).

Finally, the proposal is accepted with probability

$$A\left(X_k^*|X_k^{(i)}\right) = \min\left(1, Z^{N-1}\frac{p(X_k^*)}{p(X_k^{(i)})}\right), \tag{3.34}$$

where the number of dimensions in the parameter space is $N$. This choice of acceptance distribution ensures detailed balance will be satisfied. This algorithm can be summarized as follows:

**Algorithm 2:** Single stretch move of affine-invariant ensemble sampler (Goodman et al., 2010).

```
for $k = 1, ..., K$ do
    Draw a walker $X_j$ at random from the complementary set of walkers, $S_{[k]}$.
    Sample $u \sim U[0, 1]$.
    Sample $z \sim g(z)$.
    $X_k^* \leftarrow X_j + z\left(X_k^{(i)} - X_j\right)$.
    if $u < A\left(X_k^*|X_k^{(i)}\right) = \min\left(1, z^{N-1}\frac{p(X_k^*)}{p(X_k^{(i)})}\right)$ then
        $X_k^{(i+1)} \leftarrow X_k^*$,
    else
        $X_k^{(i+1)} \leftarrow X_k^{(i)}$.
    end if
end for
```

This process can further be parallelized by splitting the walkers into two subsets at each iteration, ($S^{(0)} = \{X_k, k = 1, ..., K/2\}$) and ($S^{(1)} = \{X_k, k = K/2 + 1, ..., K\}$), and updating the walkers in one set based on the positions of the walkers in the other set.

The underlying motivation for using this affine invariant algorithm is that it will perform equally well under linear transformations of the parameter space, and is insensitive to covariances among parameters, whereas other MCMC methods will suffer from highly correlated parameters (Goodman et al., 2010). Furthermore, there is no need to specify a multi-dimensional proposal distribution when the number of parameters exceeds 1. Instead the only free parameter is the adjustable scale parameter, $a$, in Equation 3.33, which specifies the relative range of distances the stretch move can take.

Unfortunately, it is impossible to determine if an MCMC run has converged to the stationary distribution of the Markov chain, and we need to resort to heuristic measurements to measure convergence (Roy, 2020). Simply plotting the trace plots - parameter values as a function of iteration number - can be used to select “burn-in” lengths and qualitatively judge convergence (Hogg et al., 2018). Furthermore, the acceptance ratio is informative to determine the behaviour of the random walk process, with a high acceptance fraction indicating a step size that is too low, and a low acceptance fraction indicating a step size that is too high (see Figure 3.4 for an example). Finally, the integrated autocorrelation time, which conceptually measures how many iterations of the Markov process are needed for a single independent draw from the posterior distribution, can also be used to measure convergence. This is useful, because the MCMC process causes the draws to be correlated, meaning that the effective sample size is generally lower than the actual number of iterations. We will highlight these concepts in a concrete example in Section 6.1.6.

Variational inference provides another approach to estimate a distribution over the parameter space, whereby the posterior distribution is approximated by a family of distributions using some optimization technique. While variational inference approaches will not typically find the globally optimal solution, unlike MCMC, they are much faster and have more well-defined convergence criteria.

In the following sections, we describe mathematical concepts which are essential for the analysis of scRNA-Seq datasets, without going into as much detail of the individual algorithms, but providing more of a conceptual intuition for the methods instead.

## 3.2 DIMENSION REDUCTION

The starting point for the analysis of a scRNA-Seq dataset after read alignment, droplet quantification, and droplet filtering, consists of a cell-by-gene count matrix. Initial pre-processing of the cell barcodes and running various quality control measurements to determine high quality cell-containing droplets is essential (Luecken et al., 2019). We will go into further detail of this topic in Chapter 5. The count data is then normalized, with the most common normalization technique using a size factor proportional to the count depth per cell (L. Lun et al., 2016; Vallejos et al., 2017), followed by a log transformation. Following this, feature selection is performed, where the data is subset to the genes exhibiting the highest level of variability, often measured as a regularized variance to mean ratio, or dispersion estimate (Butler et al., 2018; Wolf et al., 2018). The idea here is that among the genes profiled with a scRNA-Seq protocol, for the human genome around 25,000 to 30,000 genes, only a small percentage are truly informative to the specific dataset at hand. These genes will have variable expression levels across the cells in the dataset, and subsetting to these genes for other downstream analyses will help to reduce run-time and computation. These are typically referred to as highly variable genes (HVG) (Brennecke et al., 2013). Following selection of the top 1,000 - 5,000 HVGs, dimension reduction is performed for noise reduction and data compression, as well as visualization purposes. In this section, we briefly describe three dimension reduction techniques - Principal component analysis (PCA), diffusion maps, and Uniform Manifold Approximation and Projection (UMAP) - which will be used throughout this work. For a more extensive review on dimension reduction techniques and their use in scRNA-Seq data, we refer the reader to Moon et al. (2018).

### 3.2.1 Principal component analysis

Principal component analysis (PCA) is a linear dimension reduction approach which transforms the data into a new coordinate system, such that the greatest variance in the data lies on the first coordinate, with the residual variance maximized in each further dimension (Hotelling, 1933; Pearson, 1901). Mathematically, we can define PCA as follows. Let $X$ be our data matrix containing $n$ rows (cells) and $p$ columns (genes) ($n \times p$ matrix) with column-wise zero mean (i.e. mean-centered for each gene). Then, PCA is a transformation of the data, $T$, corresponding to

$$T = XW, \tag{3.35}$$

where $T$ is a $n \times p$ matrix containing the transformed data, and $W$ is a $p \times p$ matrix whose columns are the eigenvectors of the covariance matrix, $X^T X$, ordered corresponding to the largest eigenvalues of $X^T X$. The columns of $W$ are also referred to as the principal components. The dimension reduction component comes into play when subsetting to the dominant $d < p$ principal components that are able to capture most of the variability in the data. While there are different methods to estimate an optimal $d$ including elbow heuristics or the permutation-based jackstraw method (Chung et al., 2015), typical values for $d$ range from 30 to 50 when using PCA on scRNA-Seq datasets (Wolf et al., 2018).

One of the pitfalls of PCA is that it does not capture geometrical structure of the data in few dimensions as well as non-linear dimension reduction methods. Nonetheless, it is typically used as a pre-processing step for other downstream methods.

### 3.2.2 Diffusion maps

A diffusion map is a non-linear dimension reduction technique that embeds the data in such a way that the Euclidean distance between points in the embedded space approximates the diffusion distance in the original space (Coifman et al., 2005). Conceptually, this diffusion distance can be thought of as an average length of all the paths connecting two points in the original space, and is related to the probability of travelling from one point to another in a fixed number of iterations using a random walk. The probability of travelling from one point to another is specified in terms of a kernel function, which is symmetric and preserves positivity, typically a Gaussian kernel, where points that are closer together in the original space have a higher probability of transitioning to one another. For points $(x, y)$ in the original space, the Gaussian kernel (also known as the heat kernel) is defined as

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{\epsilon}\right), \tag{3.36}$$

where $\epsilon > 0$ represents the kernel scale, with a smaller kernel scale capturing more local structures in the data, and a larger kernel scale capturing more global structures. The kernel function must be symmetric ($k(x, y) = k(y, x)$) and positivity preserving ($k(x, y) \geq 0$), which enables it to be interpreted as a scaled probability. From this, a kernel matrix, $K$, is created where $K_{i,j} = k(x_i, x_k)$, and a diffusion matrix, $P$, can then be constructed, where

$$P = D^{-1}K, \tag{3.37}$$

where $D$ is a diagonal matrix consisting of the row sums of $K$. The diffusion matrix can be interpreted as a transition matrix of an ergodic Markov chain defined on the data, as described in Section 3.1.3 by Equation 3.23, which follows detailed balance, and thus has a unique stationary distribution (Coifman et al., 2005). Finally, the diffusion distances can be expressed in terms of the eigenvectors $\{\psi_l\}_{l\geq 0}$ and eigenvalues $\{\lambda_l\}_{l\geq 0}$ of $P$, where $1 = \lambda_0 > |\lambda_1| \geq |\lambda_2| \geq ...$, with the diffusion map defined as

$$\Psi_t(x) = \left(\lambda_1^t\psi_1(x), \lambda_2^t\psi_2(x), ..., \lambda_k^t\psi_k(x)\right), \tag{3.38}$$

where the first $k$ eigenvectors are kept, with $k$ less than the dimensionality of the original data, and $t \in \mathbb{Z}^+$, which corresponds to the $t$'th power of the transition matrix, representing a random walk of length $t$ on the data. Note, the first eigenvalue and eigenvector are dropped because they represent the stationary state corresponding to eigenvalue $\lambda_0 = 1$. Diffusion maps are particularly useful for analyzing scRNA-Seq datasets which contain cells transitioning from one state to another, as in the case of differentiating cells, as the diffusion components represented in Equation 3.38 highlight transitions in the data. Diffusion maps were first applied to scRNA-Seq data as a low-rank approximation for visualization purposes in (Haghverdi et al., 2015), and further extended to estimate the diffusion pseudotime, a distance measure representing the distance over random walks of arbitrary length from a fixed root cell to all other cells in the dataset, in (Haghverdi et al., 2016). We will be utilizing this approach in Chapter 6 of this thesis, where we also discuss the concept of pseudotime in scRNA-Seq data in more depth.

### 3.2.3 Uniform manifold approximation and projection

The final dimension reduction technique we will describe is the graph-based approach Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). The UMAP approach has gained immense popularity for visualizing scRNA-Seq datasets due to its ability to preserve both local and global structures in the data in a 2-dimensional embedding, and its scalability to a large number of cells (Becht et al., 2019). Essentially, UMAP starts by building a high-dimensional graph representation of the data, where points are connected to each other if an extended radius surrounding each point is overlapping with the radius surrounding another point. Instead of using a fixed radius, this radius is based on the distance to each point's $k$-th nearest neighbor. Thus, points which are located in densely populated regions will have a small radius, and points located in sparsely populated regions will have a large radius. This enables the graph to accurately represent both the local and global topology of the dataset in a high dimension. The edges between points are then weighted according to their relatively scaled distances to their nearest neighbors to get a connection probability between points. This high-dimensional graph is then projected into a lower dimensional space in such a way that the high-weighted edges remain close to each other, and the low-weighted edges are far apart.

Throughout this thesis, we use UMAP purely for visualization purposes to highlight similarities between cells in a 2-dimensional embedding. Other graph-based approaches which have been widely used for visualization purposes of scRNA-Seq data include t-SNE (van der Maaten et al., 2008) and SPRING (Weinreb et al., 2018). In the final section of this chapter, we briefly discuss clustering approaches used in scRNA-Seq data analysis.

## 3.3 CLUSTERING

One of the ultimate goals of scRNA-Seq is to determine the identity of the individual cells in the sample. Ideally, it would be possible to annotate cells to a corresponding cell type individually, however, due to issues related to data sparsity, where the gene expression profile of an individual cell is limited by the count depth in that cell, leading to zero counts in genes which were actually expressed in the cell, it is necessary to group cells together with similar global gene expression profiles. This can help to amplify the gene expression signal in groups of cells. This grouping of cells based on similarity of expression profile can be accomplished with clustering approaches. The similarity of expression profiles between two cells can be measured by various distance metrics, which often use the dimension-reduced data, for example Euclidean distance on the top principal components. Some widely used clustering algorithms in scRNA-seq data analysis include $k$-means clustering (MacQueen, 1967), which determines cluster centroids and assigns cells to the nearest centroid, and hierarchical clustering approaches (Defays, 1977; Sibson, 1973), which arrange cells into a hierarchy based on relative similarity. However, the most common approaches for clustering scRNA-Seq data are graph-based community detection algorithms due to their scalability. In this section, we briefly describe the most common community detection algorithm, the Louvain method for community detection (Blondel et al., 2008), which we utilize throughout this work.

### 3.3.1 Graph-based clustering approaches

The Louvain method (Blondel et al., 2008) is a graph-based clustering approach. As input, it requires a graph (weighted or unweighted), and outputs a partitioning of the graph into groups of clusters. In the context of scRNA-Seq anlaysis, this graph is typically constructed using a $k$-nearest neighbors (kNN) approach, where the $k$ nearest neighbors for cell $i$ are the cells with the smallest distance from cell $i$, typically constructed using Euclidean distance on the top principal components. The Louvain method is a method to estimate communities, sets of densely interconnected nodes, with nodes belonging to separate communities being sparsely connected. This is accomplished by finding a partitioning of the graph which optimizes a quantity known as modularity (Newman et al., 2004), which is defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \tag{3.39}$$

where $A_{ij}$ contains the weight of the edge between node $i$ and node $j$, $k_i = \sum_j A_{ij}$ is the sum of weights of all edges connected to node $i$, $c_i$ is the community to which node $i$ belongs, $\delta(c_i, c_j)$ is the Kronecker delta function (i.e. equals 1 if $c_i = c_j$, otherwise equals 0), and $m = \frac{1}{2} \sum_{i,j} A_{ij}$ is half the sum of all edge weights in the graph. This quantity measures the sum of weights of intracommunity edges minus the expected sum of weights of intracommunity edges (summarized by the term $\frac{k_i k_j}{2m}$). However, optimizing this quantity alone may fail to identify well-defined small communities (Fortunato et al., 2007). The inclusion of a resolution parameter, $\gamma > 0$, allows the detection of communities at different modular scales (Reichardt et al., 2006), overcoming this limitation. With the inclusion of $\gamma$, modularity becomes

$$Q(\gamma) = \frac{1}{2m} \sum_{i,j} \left( A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(c_i, c_j). \tag{3.40}$$

The resolution parameter has the effect of scaling the expected sum of weights of intracommunity edges, with larger values of $\gamma$ increasing this value, enabling the detection of smaller communities. Finding a partitioning of the graph which maximizes the modularity quantity globally is NP-hard due to the explosion of possible partitions as the number of nodes increases. Therefore, the Louvain method uses a greedy optimization approach based on iterating over two phases. Initially, each node in the network is assigned to its own community. Then, for each node $i$, the gain of modularity is calculated based on placing $i$ in the community of each of its neighbors $j$, and node $i$ is placed in the community which exhibits the maximal gain. This phase stops when a local maximum is achieved. The second phase consists of building a new network whose nodes consist of the communities found in the first phase, where the weights between these nodes are calculated as the sum of the weights of all edges between nodes in the respective communities. After this, the first phase is re-applied, and this continues until no subsequent change in the communities is detected. Adaptations to the Louvain algorithm have been made recently in order to guarantee communities are connected, i.e. there is a path from any node to any other node in each community (Traag et al., 2019). For a comprehensive review on graph-based community detection methods, we refer the reader to Fortunato et al. (2016).

# 4 DISSECTING REGIONAL HETEROGENEITY IN BRAIN ORGANOIDS

In this chapter, we summarize part of the findings from (Rosebrock et al., 2022). This study was a comparative study measuring the ability of various brain organoid derivation protocols to selectively enrich for cortical fates. In this study, we compared three brain organoid derivation protocols - Inhibitor-free conditions (Lancaster et al., 2013), Dual-SMAD inhibition (Paşca et al., 2015; Qian et al., 2016), and Dual-SMAD/WNT inhibition (Elkabetz et al., 2022). By systematically comparing these methods side by side across multiple cell lines, and measuring differences at the transcriptomic level using both bulk RNA-Seq and scRNA-Seq, we showed that the combination of Dual SMAD and WNT inhibition is essential for establishing a robust cortical identity at various stages of organoid development. In the following sections, we describe how these technologies were used to compare the regional specification across the different brain organoid derivation protocols.

## 4.1 COMPARING REGIONAL SPECIFICATION FROM BULK RNA-SEQ

As described in Section 2.3.1, bulk RNA-Seq provides a landscape of the gene expression patterns at a cell-population level, enabling the detection of differentially expressed genes across different tissues. This provides a list of genes that are significantly up-regulated and significantly down-regulated when comparing one set of samples to another. However, one gene is not enough to determine whether an underlying biological process, pathway, or cell type is enriched. Therefore, one needs to measure the enrichment of gene sets, which are comprised of genes linked to a single process. In the context of brain region specification, a gene set can be used to define genes that are specifically expressed in a single brain region during embryonic development, and not expressed in others. To this end, the Allen Human Brain Atlas dataset (Kang et al., 2011), comprised of bulk RNA-Seq of prenatal brain structures (neocortex, hippocampus, amygdala, striatum, thalamus and cerebellum) across multiple developmental time points, as well as post-conception samples, provides a comprehensive reference dataset to deduce region-specific genes during brain development. Brain region specific marker genes were estimated by subsetting to bulk RNA-Seq data of week 12–21 embryonic brain tissue samples derived from the neocortex, hippocampus, thalamus and cerebellum, and genes were defined as region-specific if they exhibited a log2 fold change of at least two when compared with samples from all other regions. This provided a list of region-specific genes during human embryonic development, which could then be used for gene set enrichment analysis.

Brain organoids were grown under the three derivation protocols, Inhibitor-free, Dual-SMAD inhibition (Dual SMAD-i), and Dual-SMAD/WNT inhibition (Triple-i), from a human ESC line (H9) until 30 days of development, and then individual organoids were subjected to bulk RNA-Seq. When performing a PCA on the expression levels across the top 2,000 HVGs, a clear separation of samples corresponding to protocol can be observed, as seen in Figure 4.1.

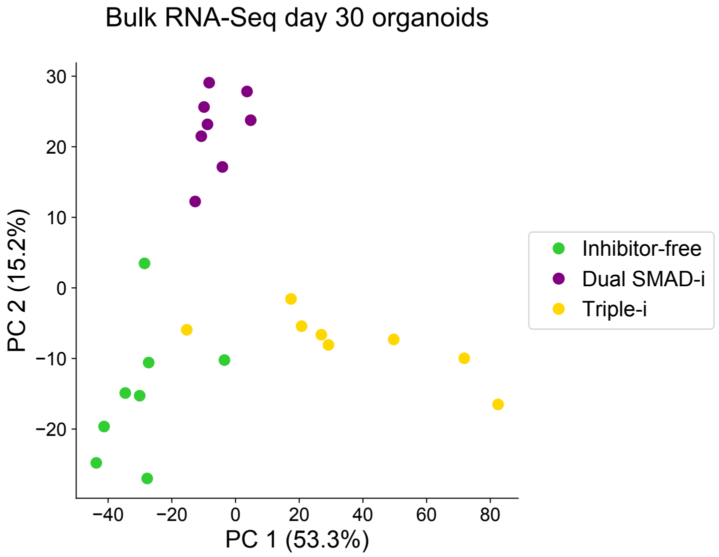

**Figure 4.1: PCA of bulk RNA-Seq day 30 brain organoids.** Expression levels for the genes in each sample were first estimated using fragments per kilobase of exon per million mapped fragments (FPKM) normalization. Then, the expression levels were normalized using a log2 normalization after adding a pseudocount of 1. The top 2,000 HVGs were estimated using the variance in log-normalized expression across all samples. The top 2 principal components are shown in the plot, and samples are colored according to derivation protocol. The percentages in the axes display the percent of variance explained by each principal component.

Following this, a differential gene expression analysis was performed using DESeq2 (Love et al., 2014) using the count data as input across three pairwise treatment comparisons, Triple-i versus Dual SMAD-i, Triple-i versus Inhibitor-free, and Dual SMAD-i versus Inhibitor-free, using the eight biological replicates of individual organoids from each protocol in each comparison. A gene set enrichment analysis was then performed to determine the significance of the enrichment of the regional specific gene sets derived from the Allen Brain Atlas in each of the three comparisons using the procedure described in (Subramanian et al., 2005) with a Benjamini–Hochberg multiple hypothesis correction (Benjamini et al., 1995) for each comparison. These results are highlighted in Figure 4.2.

These analyses highlight that Triple-i organoids significantly enriched for cortical markers when compare with both Dual SMAD-i and Inhibitor-free organoids, indicating that Triple-i promotes forebrain/cortical specification. In contrast, Dual SMAD-i organoids significantly enriched for thalamic and cerebellar markers when compared with both Dual SMAD-i and

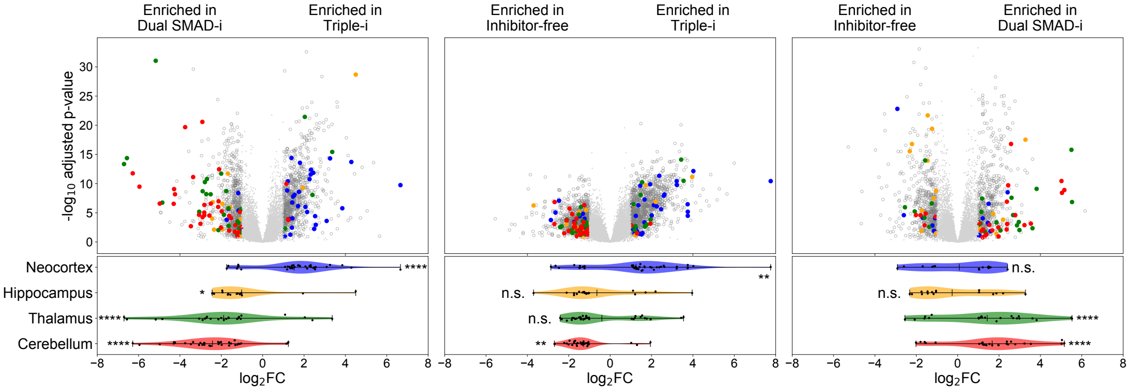

**Figure 4.2: Differential gene expression and gene set enrichment analysis of bulk RNA-Seq of day 30 brain organoids.** Volcano plots of day 30 Triple-i organoids compared with Dual SMAD-i organoids (left plot), Triple-i organoids compared with Inhibitor-free organoids (middle plot), and Dual SMAD-i organoids compared with Inhibitor-free organoids (right plot). DESeq2 (Love et al., 2014) was used to estimate the log2-transformed fold-change and adjusted p-values after Benjamini–Hochberg correction. Genes were assigned as significantly differentially expressed if they had an absolute log2-fold change of at least 1, and adjusted p-value less than 0.1. Region-specific genes from the Allen Human Brain Atlas that were significantly differentially expressed are highlighted in the volcano plots and in the violin plots below. Adjusted p-values from a gene set enrichment analysis statistical test after Benjamini–Hochberg correction for regional gene set enrichments are highlighted in the violin plots. *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001; and n.s., not significant.

Inhibitor-free organoids, indicating that Dual SMAD-i promotes posteriorization. Finally, Inhibitor-free organoids did not exhibit a consistent enrichment for any brain region.

While these results highlight the utility of bulk RNA-Seq to measure relative regional enrichments in different brain organoid protocols, it is impossible to measure the heterogeneity of cell types in an organoid with bulk RNA-Seq. To this end, scRNA-Seq can be utilized to measure the cellular compositions in a sample in order to gain a more detailed understanding of the cellular heterogeneity present in these brain organoids, as will be shown in the next section.

## 4.2 MEASURING REGIONAL HETEROGENEITY FROM SCRNA-SEQ

To further assess regional specification and cellular heterogeneity in brain organoids under these protocols, brain organoids were grown until day 50 of culture across four iPSC lines - FOK1, KUCG2, ZIP8K8 and ZIP13K5 - under Dual SMAD-i and Triple-i derivation protocols, and under two iPSC lines - ZIP8K8 and ZIP13K5 - under Inhibitor-free conditions. These organoids were then subjected to scRNA-Seq using the 10X protocol, with each scRNA-Seq experiment consisting of 4-5 pooled organoids for each cell line and derivation protocol. In order to gain a general overview of the dataset, the data was projected into a UMAP embedding, with input data consisting of the top 50 principal components of a PCA on the log-normalized expression levels after subsetting to the top 2,000 HVGs. This UMAP is shown in Figure 4.3. While the single cell transcriptomes exhibited a widespread overlap across all four iPSC lines, there was a clear segregation according to derivation protocol, with some populations containing a more heterogeneous mixture of cells from different protocols than others.

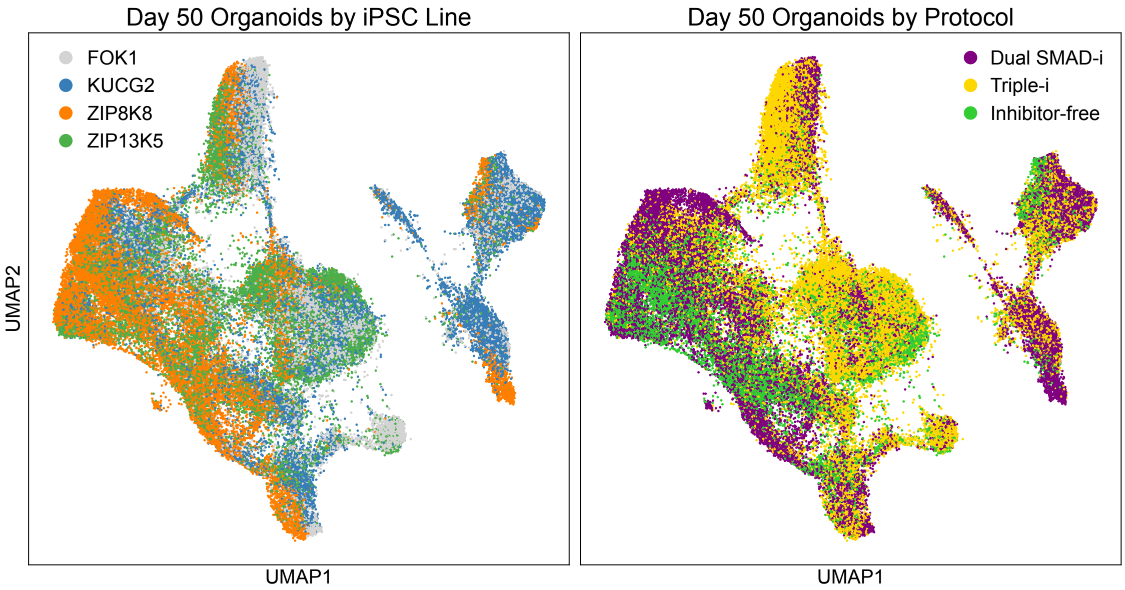

**Figure 4.3: scRNA-Seq UMAP embedding of day 50 organoids.** A UMAP embedding was generated from scRNA-Seq data of day 50 Triple-i organoids derived from FOK1, KUCG2, ZIP8K8 and ZIP13K5 iPSC lines, Dual SMAD-i organoids derived from FOK1, KUCG2, ZIP8K8 and ZIP13K5 iPSC lines, and Inhibitor-free organoids derived from ZIP8K8 and ZIP13K5 iPSC lines. In the left plot, the cells are colored according to the respective iPSC line across all derivation protocols, and in the right plot by derivation protocol across all cell lines.

To determine the cell types present in the dataset, cells were clustered and the relative expression levels of curated marker gene sets representing different cell states (i.e. dividing or non-dividing), cell types (i.e. neural stem cell, neuron, Schwann cell, mesenchyme, epithelial, choroid plexus), as well as brain regions (neocortex, optic vesicle (retinal cell types), medial pallium (hippocampus), diencephalon (thalamus), midbrain/hindbran, PNS) were used to annotate the clusters to a corresponding cell type. Clustering was performed with Louvain clustering and resolution parameter $\gamma = 4$, using an unweighted kNN graph with $k = 14$ as input, and using the top 50 principal components and Euclidean distance metric to build the kNN graph. The clustering and cell type annotations, as well as relative expression levels for the curated marker gene sets are shown in the Figure 4.4.

Finally, based on the clustering and cell type annotations, brain organoids generated under the Triple-i protocol exhibited consistent and robust cortical specification across all four cell lines (median 60% cortical cell types) accompanied by a repression of posterior (thalamus, midbrain/hindbrain, Schwann cells) and PNS fates (median 23% posterior/PNS cell types). In stark contrast, three of the four cell lines differentiated under Dual SMAD-i exhibited an overwhelming

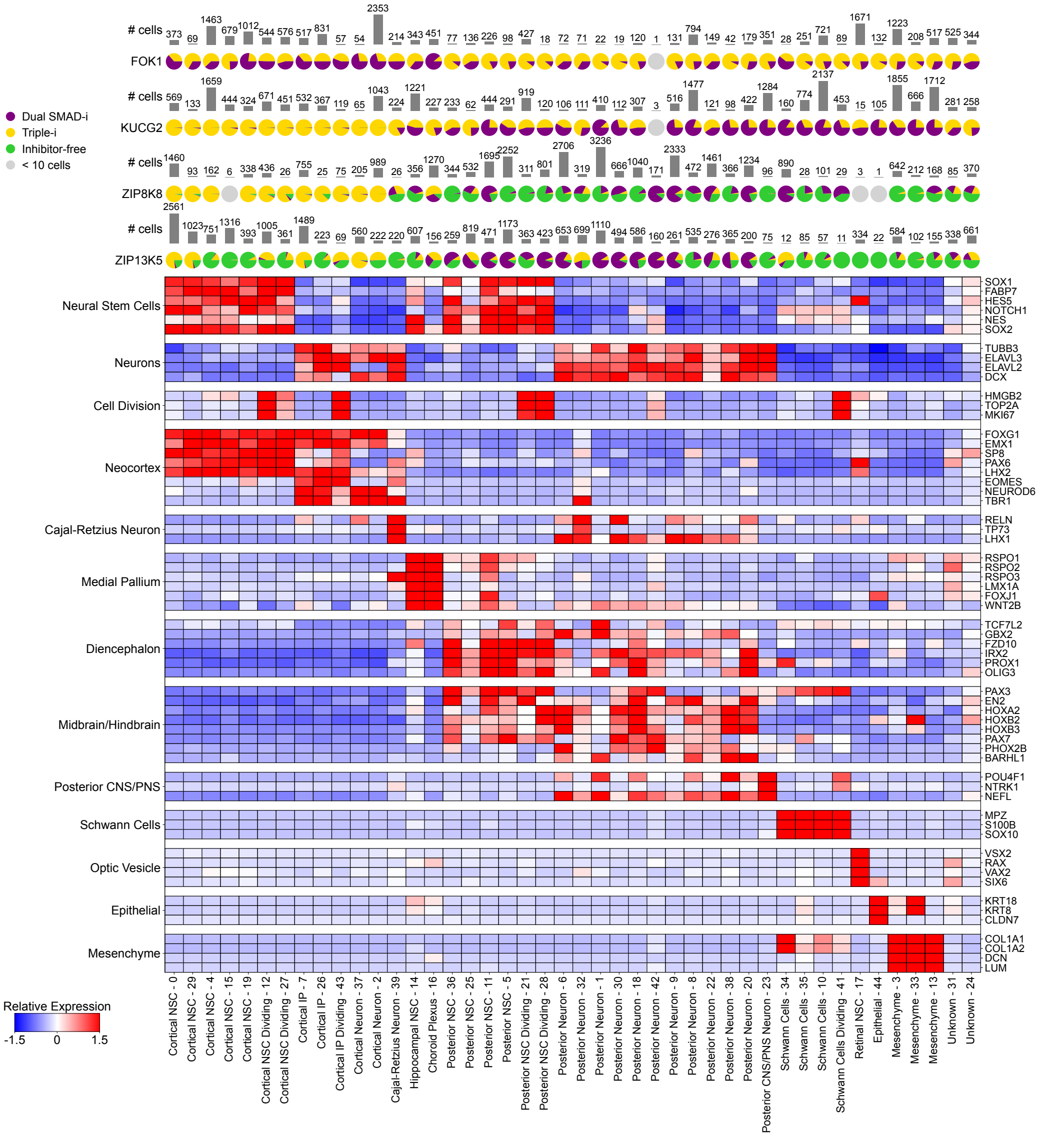

**Figure 4.4: scRNA-Seq relative expression of marker genes across clusters in day 50 organoids.** The heatmap displays the relative expression values after z-score normalization of the average log-normalized expression values for each gene across clusters after doublet removal for selected genes categorized according to cell state, cell type, and brain region. The percentage of cells from each derivation protocol separated by iPSC line comprising each cluster are provided (top; pie charts). The pie charts are coloured in grey if fewer than ten cells from that iPSC line were assigned to the given cluster. The bar plots (top) display the total number of cells in each iPSC line assigned to the given cell type.

posterior CNS/PNS identity (median 78% posterior/PNS cell types), whereas only one cell line (FOK1) contained a high level of cortical specification (64% cortical cell types). Inhibitor-free conditions also inconsistently gave rise to cortical populations, with one cell line (ZIP13K5) containing 52% cortical identity and the other (ZIP8K8) yielding merely 0.7% cortical identity. Furthermore, the proportions of cortical cell types within Triple-i organoids was highly conserved across iPSC lines, indicating that this derivation protocol generates consistent cortical populations regardless of intrinsic biases in the underlying iPSC line. The regional compositions and cortical cell compositions in Triple-i organoids are displayed in Figure 4.5.

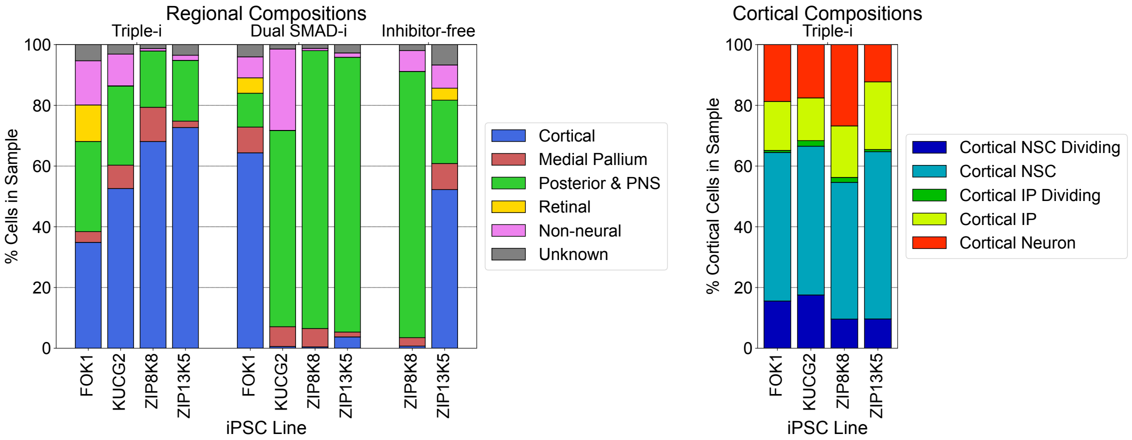

**Figure 4.5: Regional composition of day 50 organoids from scRNA-Seq.** The bar plot on the left displays the regional compositions of day 50 brain organoids derived from scRNA-Seq data separated by cell line and derivation protocol. The bar plot on the right displays the cortical compositions of the corresponding Triple-i organoids.

## 4.3 SUMMARY

Methods for deriving brain organoids are highly diverse and give rise to immensely heterogeneous populations with respect to cortical identity. Despite this fact, comparative studies measuring the level of cortical specification across a variety of cell lines of different brain organoid protocols are still exceptionally sparse. By systematically comparing methods side by side using bulk RNA-Seq and scRNA-Seq in this chapter, we characterized the regional biases present across derivation methods. This analysis revealed that organoids generated by Triple-i exhibit a highly consistent cortical identity independent of cell line, whereas organoids generated by Dual SMAD-i and Inhibitor-free conditions exhibited sporadic cortical specification across different cell lines, with a strong posterior bias in Dual SMAD-i organoids.

Furthermore, we highlighted the insights scRNA-Seq enables in deducing the cellular heterogeneity present in complex tissues, such as brain organoids. Nonetheless, we did not discuss the pre-processing steps involved in the analysis of scRNA-Seq data. In the following chapter, we will present a method developed as a part of this work to detect and remove artifactual cell types consisting mainly of free-floating ambient mRNA in the sample. We apply this method to a day 50 brain organoid inDrops scRNA-Seq dataset, as well as the day 50 brain organoid 10X scRNA-Seq dataset presented in this chapter.

# 5 ADDRESSING TECHNOLOGICAL ARTIFACTS IN SCRNA-SEQ

A typical bioinformatics program will require input of some form, and from this generate an output. In some cases, it is possible to verify that the input meets some previously specified measures of data quality before generating the output. However, in many instances, measures of data quality can be difficult to define, especially when the potential error modes of the input data are not fully understood. If the input data is however low quality or contains potential artifactual modes, then it is impossible to generate meaningful output without properly addressing the data quality. This is generally known as the "Garbage In, Garbage Out" principal, originally coined by George Fuechsel, an early IBM programmer and instructor. This chapter gives an overview of potential artifactual modes in scRNA-Seq data, which when not properly addressed, can lead to potentially confounding and incorrect results. In particular, we also present a method we developed as a part of this thesis in order to correct for one such error mode consisting of empty droplets.

## 5.1 DETECTING CELL-CONTAINING DROPLETS IN SCRNA-SEQ DATA

Once reads have been aligned and cell and UMI barcodes assigned to individual read pairs in a scRNA-Seq dataset, a threshold on the library size of individual cell barcodes is set empirically to distinguish cell-containing droplets from droplets which do not contain cells. Droplets which do not contain cells will typically have a lower number of UMI counts than a cell-containing droplet. This empirical threshold can be determined by plotting the library size per cell barcode in a log-log scale and placing a cut-off where the library size begins to drop dramatically, namely at the inflection point highlighted in Figure 5.1. The sharp increase in cell barcodes with a lower number of assigned UMI counts than the inflection point can more clearly be seen when plotting a distribution of the UMI counts per cell barcode on a $\log_{10}$ scale in Figure 5.2.

After filtering for cell-containing droplets based on UMI counts per cell barcode, further cutoffs can be used to filter cells with a high expression of mitochondrial genes, indicative of cells undergoing apoptosis, or cells with a ruptured cell membrane, in which cytoplasmic mRNA has leaked, leaving higher relative levels of mitochondrial RNA in the cell (Ilicic et al., 2016). These cell barcodes typically have a lower UMI count than the rest of the cells in the sample, and plotting UMI counts per cell barcode and mitochondrial content jointly can assist in determining more accurate cutoffs. Figure 5.3 displays an example for the above inDrops experiment of day 50 Triple-i organoids. All cell barcodes above the % mitochondrial content cutoff are filtered from

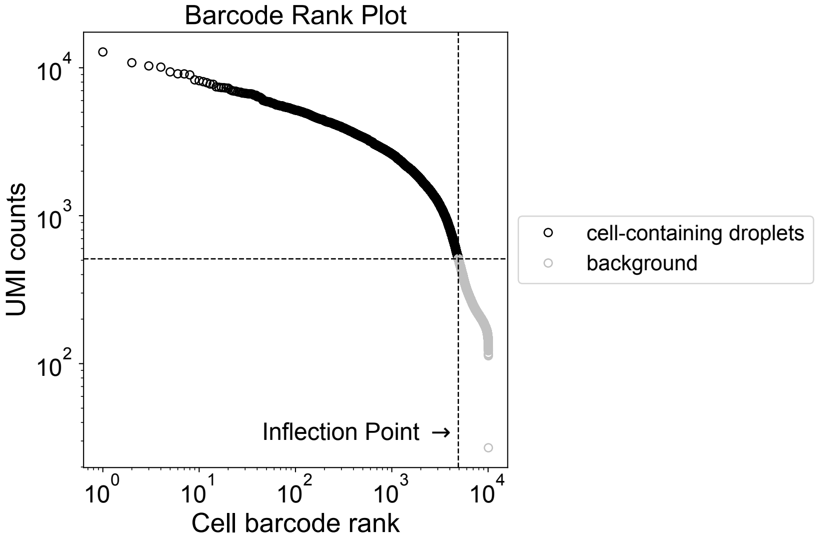

**Figure 5.1: Cell barcode rank plot from scRNA-Seq data generated using inDrops protocol.** To determine the empirical threshold for cell-containing droplets from those containing a background signal, total UMI counts per cell barcode are ordered and the rank is plotted in a $\log_{10}$ scale on the x-axis and and the total UMI counts in a $\log_{10}$ on the y-axis corresponding to each cell barcode. All cell barcodes with a UMI count higher than the inflection point are assumed to be cell-containing droplets.

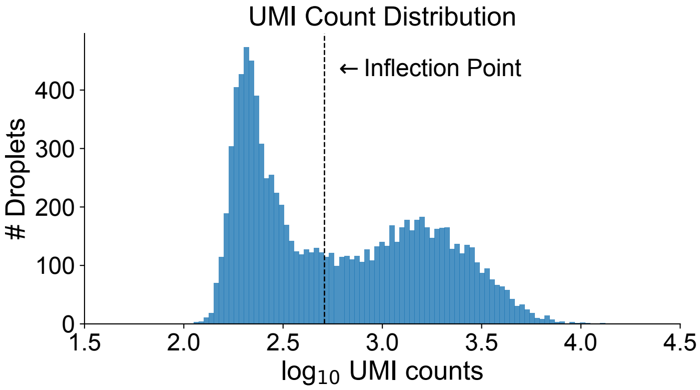

**Figure 5.2: Distribution of UMI counts per cell barcode.** When plotting a distribution of the UMI counts per cell barcode, a sharp increase in the number of UMI counts per cell barcode below the inflection point can be seen. The majority of these droplets most likely do not contain cells.

downstream analyses. However, mitochondrial genes are ubiquitously expressed in all human cells and the expression of these genes vary from cell to cell depending on cell stress, metabolic function, and other extrinsic factors (Galluzzi et al., 2012; Kotrys et al., 2019). Therefore, these cutoffs are sample specific and also typically chosen empirically.

Simply filtering cells based on the UMI counts alone is typically not enough to keep only high quality cells. Further filtering based on other error modes is needed to ensure all "garbage" has been removed from the input data.

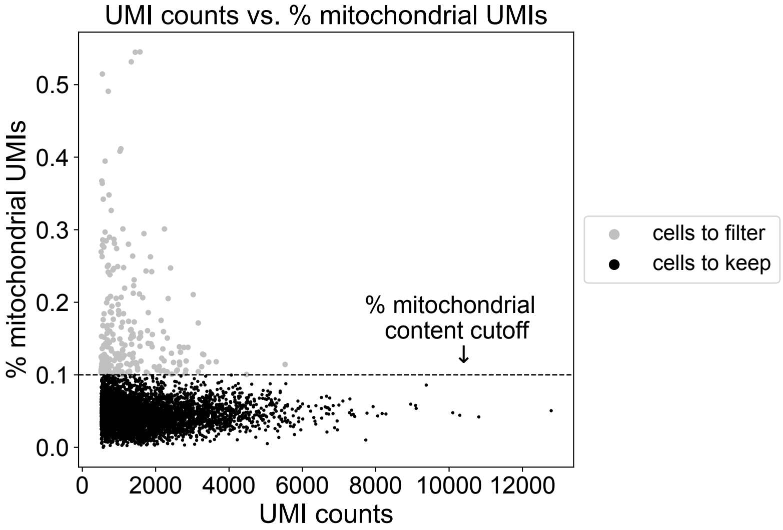

**Figure 5.3: UMI counts vs. mitochondrial content per cell barcode.** Plotting UMI counts per cell barcode on the x-axis and mitochondrial content (% mitochondrial UMI counts per cell barcode) on the y-axis, enables the identification of low-quality cells, which have a low UMI count and high mitochondrial content, indicative of cells which have lost cytoplasmic mRNA or are undergoing apoptosis.

## 5.2 ARTIFACTUAL MODES IN SCRNA-SEQ DATA

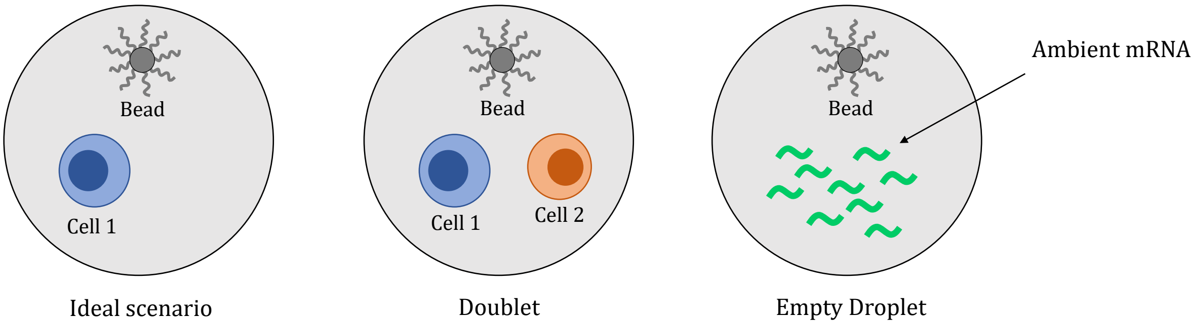

**Figure 5.4: Artifactual modes in scRNA-Seq data.** The left panel shows the ideal scenario when generating droplets for scRNA-Seq, namely where one droplet contains a single bead and a single cell. Two widely known error modes when forming droplets are the generation of doublets, in which one droplet contains two cells, as well as empty droplets, in which a droplet does not contain a cell. Ambient mRNA is typically present throughout the entire mixture and every droplet will contain ambient mRNA to some extent, however, when an empty droplet is formed, the only source which contributes mRNA to the droplet is free-floating ambient mRNA.

The various artifactual modes arising in scRNA-Seq datasets are displayed in the diagram in Figure 5.4. Doublets or multiplets occur when a droplet contains two or more cells respectively.

Cell barcodes with much higher UMI counts than the rest of the cell barcodes in the sample may be indicative of doublets or multiplets. Potential doublets can be filtered by applying a maximum cutoff to the UMI count and removing cell barcodes with a larger UMI count than this cutoff. More sophisticated approaches to detecting doublets exist, such as Scrublet (Wolock et al., 2019) and DoubletFinder (McGinnis et al., 2019). These methods work by generating simulated doublets from the gene expression profiles of randomly selected cell barcodes across two clusters, and deriving a similarity score for each cell in the dataset based on the proportion of simulated doublets in the cell's $k$-nearest neighbors.

Empty droplets are droplets which do not contain a cell. The measured mRNA in these droplets is comprised of ambient mRNA, or mRNA from lysed cells which is free-floating in the mixture. Applying a cutoff to remove cells with low UMI counts, as described in the previous section, should remove the majority of empty droplets. However, empty droplets with a larger UMI count than the empirical cutoff will be inaccurately called as cell-containing droplets. Similarly, cell-containing droplets with a lower UMI count than the empirical cutoff will be inaccurately called as empty droplets. One method that attempts to rescue cell-containing droplets with a lower UMI count is emptyDrops (Lun et al., 2019).

Furthermore, ambient mRNA can contaminate cell-containing droplets. This contamination can vary from cell to cell within the same experiment, and across different experiments, depending on how much ambient mRNA is present in the sample. The presence of ambient mRNA can confound analyses, for example resulting in the detection of cell-type specific marker genes in cells which do not express these genes. Methods to remove ambient mRNA signal from scRNA-Seq data exist, such as DecontX (Yang et al., 2020) and SoupX (Young et al., 2020). These methods correct the gene expression profiles of individual cells after removal of ambient mRNA signal. The following section discusses an orthogonal approach to detect empty droplets containing predominantly ambient mRNA within scRNA-Seq data.

## 5.3 DETECTING EMPTY DROPLETS

Computational approaches exist for removing ambient mRNA signal from individual cells, including DecontX (Yang et al., 2020) and SoupX (Young et al., 2020), and rescuing cell-containing droplets with low UMI counts, including emptyDrops (Lun et al., 2019), but there is a lack of tools designed specifically to remove empty droplets which were incorrectly labeled as cell-containing droplets. In this section, we describe a novel approach to detect and remove these cells. We apply the method to scRNA-Seq of day 50 brain organoids generated with inDrops (Klein et al., 2015) and 10X (Zheng et al., 2017) protocols, and highlight the utility of this critical pre-processing step in the analysis of scRNA-Seq data.

### 5.3.1 Simulating empty droplets

Under the assumption that an empty droplet contains purely ambient mRNA from the sample, the transcriptomic profile of an empty droplet should match the background profile of all mRNA molecules in the sample. However, building a background mRNA profile is not entirely straightforward. One approach is to pool the mRNA signal among droplets which contain low UMI counts, under the assumption that these droplets did not contain cells, but rather the ambient mRNA in the sample. This approach is used by SoupX. However, there may be a disproportionate number of stressed cells or cells undergoing apoptosis in the droplets containing low UMI counts, which will result in a high mitochondrial mRNA signal, thereby skewing the true ambient mRNA profile. In order to mitigate this effect, a background count profile is defined as follows. Let $C$ be the count matrix of a scRNA-Seq experiment, with $C_{ij}$ = number of UMI counts for gene $i$ in cell barcode $j$. Then, define $B_i$ as the sum of UMI counts for $i$ across all cell barcodes,

$$B_i = \sum_j C_{ij}. \tag{5.1}$$

Background droplets can then be simulated by sampling from a multinomial distribution built from this background count profile. First, normalize the background count profile to sum to 1 as follows:

$$P_i = \frac{B_i}{\sum_i B_i}. \tag{5.2}$$

$P_i$ represents the probability of drawing a count for gene $i$ from the summed background count profile. To simulate a cell with $N$ UMI counts containing an ambient mRNA signal, draw $N$ random samples from the multinomial probability distribution $P = \{P_i\}_{i=1,...,M}$ where $M$ is the total number of genes. To ensure $N$ reflects the UMI count distribution of cell-containing droplets, $N$ is randomly chosen from the distribution of UMI counts for cell barcodes with a UMI count of at least $K$, where $K$ is the UMI count cutoff for determining cell-containing droplets. Figure 5.5 shows a joint UMAP embedding of real cells and 1,000 simulated cells containing ambient mRNA from the inDrops scRNA-Seq dataset of day 50 Triple-i derived brain organoids.

The next step is to measure the similarity of real cells in the dataset to the set of simulated cells containing ambient mRNA signal, and classify cells in the dataset with high similarilty to the simulated cells as empty droplets. One approach to do this is to co-cluster the real cells from the dataset with the simulated cells. All real cells in the dataset which co-cluster with simulated cells are labeled as "background" droplets. One of the downsides of using such an approach is that the optimal granularity of clustering is not known a priori. Another metric which estimates the similarity of real cells in the dataset to the set of simulated cells is the proportion of simulated cells in the $k$-nearest neighbors of all reall cells, a quantity which the doublet detection approaches Scrublet and DoubletFinder utilize when simulating artificial doublets. The proportion of simulated cells in the $k$-nearest neighbors of all real cells can be used to inform the granularity

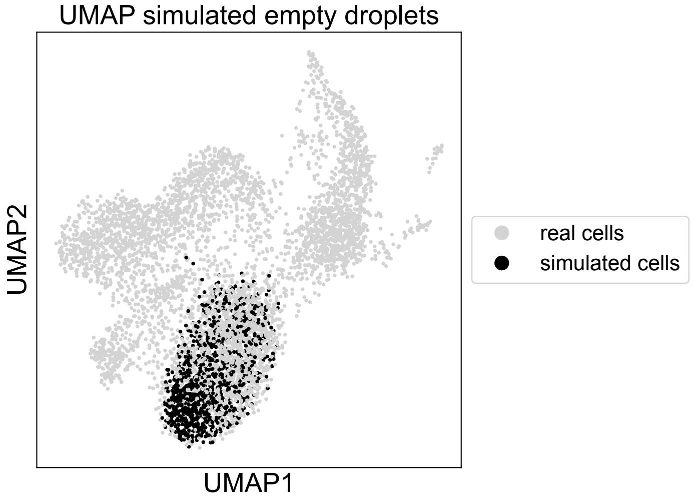

**Figure 5.5: UMAP embedding of real cells and simulated cells.** Plotting simulated cells and real cells from the dataset in a joint UMAP highlights a strong overlap between a subset of real cells and simulated cells.

of clustering, thereby reducing the dependency of results on pre-specified hyperparameters. In the following section, this approach will be described in detail.

### 5.3.2 Classification of cell barcodes as background using co-clustering

One approach to determine if a cell barcode contains mainly ambient mRNA is to co-cluster the simulated background cells with the real cells in the dataset. Ideally, all simulated cells will belong to one cluster, along with real cells from the dataset, which can then be labeled as empty droplets. However, choosing an optimal resolution parameter a priori for Louvain clustering is not straightforward. A resolution parameter which is too low will lead to underclustering, potentially resulting in the inaccurate assignment of cell-containing droplets as empty droplets, while a resolution parameter which is too high may lead to the simulated cells splitting into multiple subclusters, as seen in Figure 5.6. A joint approach with co-clustering and incorporating a metric of similarity with simulated cells derived from the $k$-nearest neighbors graph can help to mitigate these issues.

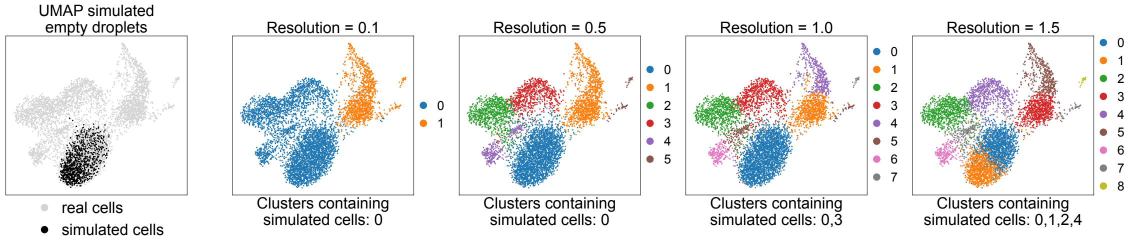

**Figure 5.6: Co-clustering real and simulated cells with varying Louvain resolution parameter.** The UMAPs show clustering results after merging real and simulated background cells from the dataset with varying resolution parameters in Louvain clustering using $k = 15$ nearest neighbors. A resolution parameter of 0.1 leads to under-clustering of the cells, resulting in many cells co-clustering with simulated background cells, while a resolution parameter of 1 or 1.5 leads to multiple clusters containing simulated background cells.

The proportion of simulated cells in the $k$-nearest neighbors of all real cells can be a useful metric to quantify the similarity of a real cell to the simulated cells. To begin, build a $k$-nearest neighbors graph including both simulated empty droplets and real cells from the dataset. Define $n_{si}$ as the number of simulated cells which are connected by an edge to cell $i$ in the $k$-nearest neighbors graph, and $n_{ri}$ as the number of real cells in the dataset which are connected by an edge to cell $i$ in the $k$-nearest neighbors graph. The proportion of simulated cells connected to cell $i$ in the $k$-nearest neighbors graph, defined as $P_{si}$, is then:

$$P_{si} = \frac{n_{si}}{n_{si} + n_{ri}}. \tag{5.3}$$

Since all simulated cells are drawn from the same background distribution of ambient mRNA, they should be strongly interconnected, and have a large $P_{si}$ value. Any real cell in the dataset which has a $P_{si}$ value within the range of $P_{si}$ values of simulated cells is likely to resemble a droplet containing purely ambient mRNA. Figure 5.7 shows the distribution of $P_{si}$ values for both simulated and real cells in the dataset, as well as the minimal value of $P_{si}$ in all simulated cells.

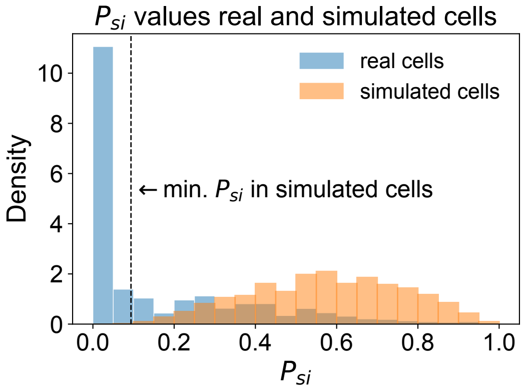

**Figure 5.7: $P_{si}$ values in both real and simulated cells.** The simulated cells have much higher $P_{si}$ values than the cells from the dataset. However, some of the cells from the dataset have $P_{si}$ value within the range of those derived from simulated cells. A cell from the dataset within this range has a strong likelihood of containing a large proportion of ambient mRNA.

The estimates for $P_{si}$ in real cells can be used to find an optimal resolution parameter for the co-clustering of real and simulated cells in the following way. For a given clustering of real and simulated cells using Louvain clustering and resolution parameter, $r$, calculate the average $P_{si}$ of all real cells in clusters containing at least 5% simulated cells, and the average $P_{si}$ for all real cells in clusters containing less than 5% simulated cells. Define the difference of these two terms as $D_r$. Then, increase the resolution parameter $r$ from 0.01 to 2 with a step size of 0.01, and record the difference $D_r$. The quantity $D_r$ represents the enrichment in the connectivity of real cells with simulated cells among those which co-cluster with simulated cells compared to those which do not. Ideally, this parameter will be high if the Louvain clustering selectively co-clusters simulated cells and real cells with a high connectivity. Therefore, the resolution parameter should be chosen to maximize this quantity. For each cluster, define $P_{sc}$ as the average of the $P_{si}$ values across all reals cells, $i$, in cluster $c$. Then, for each cluster $c$ that has a $P_{sc} \geq 0.05$, all cells in cluster $c$ are annotated as "empty droplets". In pseudocode, the above procedure is formulated in Algorithm 3.

Figure 5.8 highlights the estimates of $D_r$ for the inDrops scRNA-Seq dataset of day 50 Triple-i derived brain organoids shown in the previous sections using the above procedure. The resolution parameter $r = 0.76$ maximizes the quantity $D_r$.

In this example, only one cluster contains simulated cells using resolution parameter $\text{argmax}_r\, D_r = 0.76$, cluster 0, and the $P_{sc}$ estimate for this cluster is 0.37, with all others clusters having $P_{sc}$ estimates less than 0.02. Therefore, the real cells in the dataset assigned to cluster 0 are annotated as empty droplets. These cells have elevated $P_{si}$ values and are also in close proximity to the simulated cells in the UMAP, shown in Figure 5.9, a further indication of their similarity with the ambient mRNA expression profile.

**Algorithm 3:** Classify cell barcodes as cell-containing or “empty” droplets.

```
Simulate $1,000$ cells as described in Section 5.3.1.
Estimate $k$-nearest neighbors graph with $k = 15$.
Set Louvain clustering resolution parameter $r = 0.01$.
while $r \leq 2$ do
    Cluster cells using Louvain clustering and resolution parameter $r$.
    Estimate $A_r$ = average $P_{si}$ for all real cells in clusters containing $\geq 5\%$ simulated cells.
    Estimate $B_r$ = average $P_{si}$ for all real cells in clusters containing $< 5\%$ simulated cells.
    Record $D_r = \min(0, A_r - B_r)$.
    Set $r = r + 0.01$.
end while
Calculate $\operatorname{argmax}_r D_r$.
Estimate $P_{sc}$ = average $P_{si}$ estimates for all real cells, $i$, in cluster $c$ with resolution
$r = \operatorname{argmax}_r D_r$.
For each cluster $c$ with $P_{sc} \geq 0.05$, annotate cell barcodes within cluster $c$ as “empty”.
```

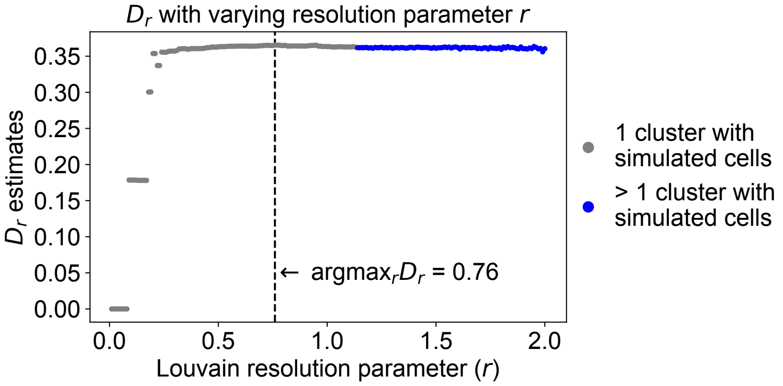

**Figure 5.8:** $D_r$ **estimates with varying Louvain resolution parameter** $r$**.** Different values for the Louvain resolution parameter, $r$, are plotted against $D_r$. Values for $r$ which produce only 1 cluster containing simulated cells are highlighted in gray, and those which produce more than 1 cluster containing simulated cells are in blue. For values of $r \leq 0.08$, only 1 cluster is found in the data, hence the $D_r$ estimates for these values of $r$ is 0.

Another piece of evidence which further supports the classification of these barcodes as empty droplets is their relative number of UMI counts to cell-containing droplets. Figure 5.10 highlights the distribution of UMI counts in annotated cell-containing and empty droplets. There is a significant decrease in the number of UMIs in the empty droplets compared to cell-containing droplets ($P = 7.63e - 220$, Mann-Whitney U Test). This is expected, given that cell-containing droplets should contain cells with a higher concentration of mRNA molecules than the ambient mRNA in the sample.

Finally, when running a differential gene expression analysis, comparing the expression levels across all cells in each cluster against the set of simulated cells based on the clustering from Figure 5.9, using a Wilcoxon rank sum-test with Benajmini-Hochberg multiple hypothesis correction,

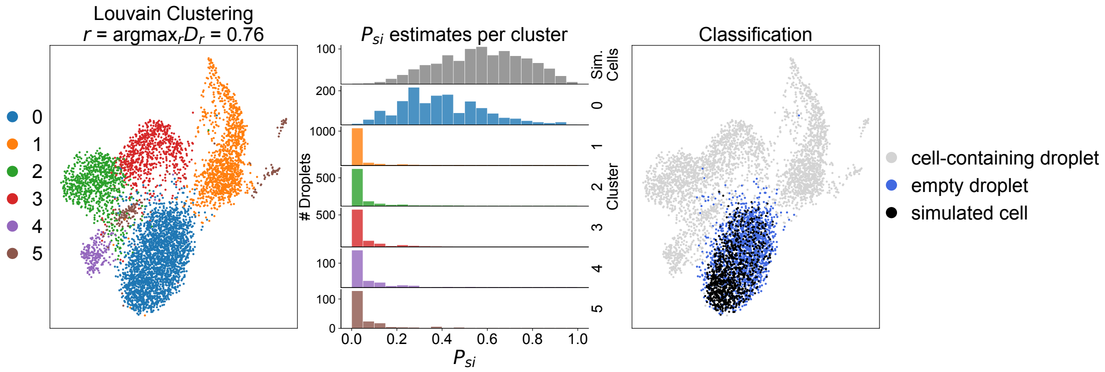

**Figure 5.9: Classification of cells as empty droplets.** The UMAP on the left shows the clustering of cells with Louvain resolution parameter $\mathrm{argmax}_r\, D_r = 0.76$. All cells from the dataset which co-cluster with simulated background cells (cluster 0) are classified as empty droplets. These cells have elevated $P_{si}$ values, as shown in the middle histograms. The classifications of droplets as empty or cell-containing are highlighted in the UMAP on the right.

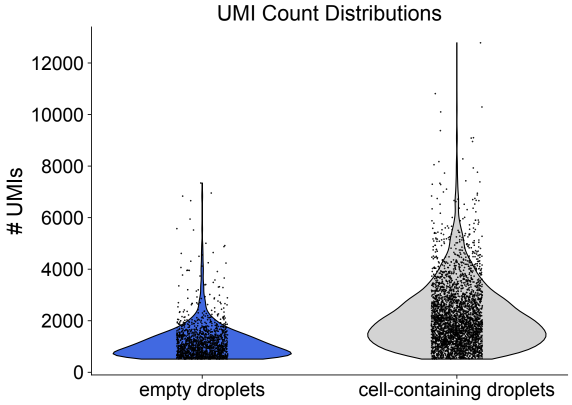

**Figure 5.10: UMI counts in empty droplets.** The violin plot highlights the distribution of UMI counts in annotated empty droplets and cell-containing droplets. Points highlight the UMI counts corresponding to individual cell barcodes.

0 genes were found to be up-regulated in the cells belonging to cluster 0, and only 3 were found to be up-regulated in the simulated background cells, as shown in Figure A1. All other clusters contained many significantly differentially expressed genes. This further supports the claim that the cells belonging to cluster 0 are empty droplets.

### 5.3.3 Genotyping individual cells to detect empty droplets

When samples from multiple cell lines or individuals are included in a single scRNA-Seq library, it is possible to utilize cell-line specific mutations to genotype the reads or UMIs from a given cell barcode and assign it a cell line of origin (Kang et al., 2018). Furthermore, since the ambient mRNA pool will contain a mixture of mRNA molecules from all input cells, any cell barcode which contains mainly ambient mRNA should contain a mixture of genomic material reflecting the relative proportions of cells from the input populations. By genotyping individual cells to a cell line or individual of origin, a process also known as demultiplexing, it is possible to detect cell barcodes containing ambient mRNA in a completely orthogonal manner to the approaches described in the previous section, enabling a further validation of cell barcode annotation as empty or cell-containing. In this section, we present a method to demultiplex cells in a scRNA-Seq sample derived from two individuals, or cell lines, with known genotypes. Here, we only consider single nucleotide polymorphisms (SNP), which consist of a single base-pair substitution unique to one of the individuals.

Define the following variables:

$$\begin{aligned}
N_{i1} &= \text{number of UMIs which support cell line 1 genotype in cell } i,\\
N_{i2} &= \text{number of UMIs which support cell line 2 genotype in cell } i,\\
c_i &= \text{\% ambient mRNA UMI counts in cell } i,\\
p_1 &= \text{fraction of UMIs in background mRNA profile belonging to cell line 1},\\
p_2 &= \text{fraction of UMIs in background mRNA profile belonging to cell line 2}.
\end{aligned}$$

Here, $p_1$ and $p_2$ can be estimated from the sample as follows. If $N_1$ = number of UMIs which support cell line 1, and $N_2$ = number of UMIs which support cell line 2, measured across all UMIs in all cell barcodes in the sample, then these values can be estimated as $p_1 = \frac{N_1}{N_1+N_2}$ and $p_2 = \frac{N_2}{N_1+N_2}$. For each cell-containing droplet, both the captured cell and ambient mRNA in the droplet contribute to $N_{i1}$ and $N_{i2}$. Let $N_i = N_{i1} + N_{i2}$. Then, if cell $i$ was derived from cell line 1,

$$\begin{aligned}
N_{i1} &= (1 - c_i)N_i + c_i p_1 N_i,\\
N_{i2} &= c_i p_2 N_i.
\end{aligned}$$

Similarly, if cell $i$ was derived from cell line 2,

$$\begin{aligned}
N_{i2} &= (1 - c_i)N_i + c_i p_2 N_i,\\
N_{i1} &= c_i p_1 N_i.
\end{aligned}$$

The only unknown variable in the above model for a given cell $i$ is $c_i$. Under the assumption that $N_{i1}$ follows a binomial distribution, the likelihood of observing $N_{i1}$ UMIs which support cell line 1 genotype in cell $i$ is:

$$\begin{aligned}
\mathcal{L}(N_{i1}|c_i; \text{cell } i \text{ from cell line 1}) &= \text{Binom}\left(N_i, \frac{N_{i1}}{N_i}\right)\\
&= \text{Binom}(N_i, 1 - c_i p_2) \qquad (5.4)\\
&= \binom{N_i}{N_{i1}}(1 - c_i p_2)^{N_{i1}}(c_i p_2)^{N_{i2}}.
\end{aligned}$$

Solving for $c_i$ which maximizes this likelihood will provide the maximum likelihood estimate for the level of ambient mRNA in cell $i$, under the assumption that cell $i$ was derived from cell line

1. This is equivalent to solving for $c_i$ which maximizes the log-likelihood, which can be solved as follows,

$$\frac{d}{dc_i}\left(\ln\left(\mathcal{L}(N_{i1}\right)\right) = \frac{d}{dc_i}\left(\ln\binom{N_i}{N_{i1}} + N_{i1}\ln(1-c_ip_2) + N_{i2}\ln(c_ip_2)\right) = 0. \tag{5.5}$$

It is possible to solve for this solution analytically, and when doing so, provides the following solution,

$$\hat{c}_i = \frac{N_{i2}}{N_ip_2}. \tag{5.6}$$

Similarly, under the assumption that cell $i$ was derived from cell line 2, the maximum likelihood estimate for $c_i$ is,

$$\hat{c}_i = \frac{N_{i1}}{N_ip_1}. \tag{5.7}$$

Finally, the maximum likelihood estimate is chosen to be $\min\left(\frac{N_{i2}}{N_ip_2}, \frac{N_{i1}}{N_ip_1}\right)$. If $\hat{c}_i$ is 1, then cell barcode $i$ most likely contains purely ambient mRNA, while if $\hat{c}_i$ is 0, then cell barcode $i$ most likely does not contain any ambient mRNA.

The inDrops scRNA-Seq dataset of day 50 Triple-i brain organoids described in the previous section was derived from pooling together organoids derived from the iPSC lines ZIP8K8 and ZIP13K5. The above method was run on this sample, and Figure 5.11 shows the estimates of $\hat{c}_i$ for corresponding estimates of $N_{i1}$ and $N_{i2}$ for each cell barcode $i$ in the sample.

This provides an orthogonal piece of evidence, when compared to the method described in Section 5.3.2, to label cells as empty droplets. If a cell barcode was annotated as an empty droplet, then the cell should also have a high $\hat{c}_i$ estimate. Figure 5.12 shows highly elevated $\hat{c}_i$ estimates for cell barcodes which were annotated as empty droplets (cluster 0; median $\hat{c}_i = 0.80$), indicating that these barcodes predominantly contain a mixture of genetic material from both cell lines. This result validates the finding that these are in fact empty droplets. Furthermore, the median $\hat{c}_i$ estimates for cells in the remaining clusters is 0.13, indicating an estimated global background ambient mRNA contribution of 13% in the sample.

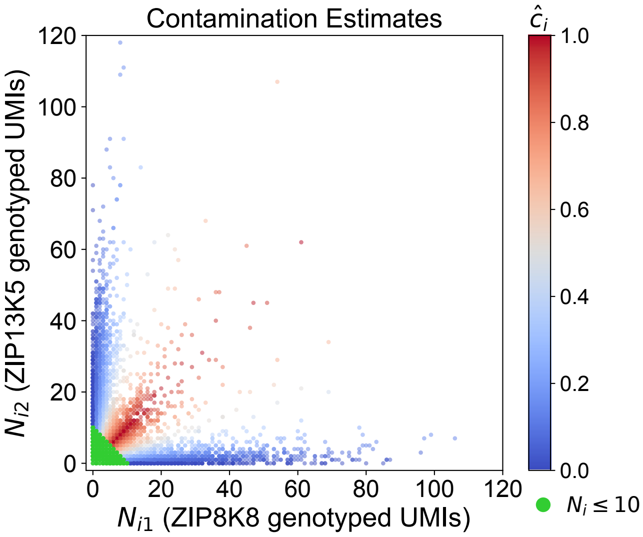

**Figure 5.11: $\hat{c}_i$ estimates for corresponding $N_{i1}$ and $N_{i2}$ estimates in each cell barcode.** $N_{i1}$ is plotted on the x-axis and $N_{i2}$ is plotted on the y-axis for each cell barcode $i$. Cell barcodes are colored according to their $\hat{c}_i$ MLE estimates. Cell barcodes with 10 or fewer genotyped UMIs ($N_i \leq 10$) are shown in green. $\hat{c}_i$ is not estimated for these cells due to small sample size and thus reduced confidence in $\hat{c}_i$ estimates. For this sample, $p_1$ was estimated to be 0.499 and $p_2$ to be 0.501, indicating a similar contribution of both cell lines to the sample.

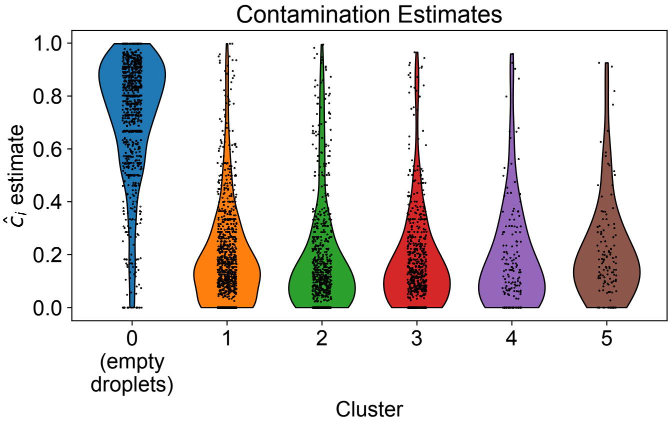

**Figure 5.12: $\hat{c}_i$ estimates per cluster.** The violin plot shows the $\hat{c}_i$ estimates for each cluster, which were derived using the procedure from Section 5.3.2, and are also shown in Figure 5.9. Cells in cluster 0 were annotated as empty droplets. The high $\hat{c}_i$ estimates for cells in this cluster validate this annotation.

### 5.3.4 Comparison with other methods

Other computational methods exist which address the issue of empty droplets and ambient RNA detection in scRNA-Seq datasets. These methods include emptyDrops (Lun et al., 2019), DecontX (Yang et al., 2020), and SoupX (Young et al., 2020). The emptyDrops method is designed to identify cell barcodes which correspond to non-empty droplets by measuring the likelihood that a droplet matches the background transcriptomic profile. The background, or ambient, transcriptomic profile used in emptyDrops is built from cell barcodes with low UMI counts. In the emptyDrops method, the likelihood that a droplet contains purely ambient mRNA is estimated using a Dirichlet-multinomial model. Cells which significantly deviate in their transcriptomic profile from the aggregated background transcriptomic profile are labeled as cell-containing droplets, while cells which do not significantly differ from the background profile are labeled as empty droplets. This method was designed to rescue cell-containing droplets with low UMI counts, and is not as sensitive at removing droplets with higher UMI counts that contain mainly ambient mRNA. When running emptyDrops on the day 50 brain organoids dataset presented in the previous section, the method detects 13% of cells in cluster 0 (labeled as empty droplets by the approach in Section 5.3.2) as empty droplets, while all remaining clusters contained less than 5% empty droplets. Thus, while emptyDrops is able to detect an elevated presence of empty droplets in this cluster, it incorrectly labels the majority of cells in the cluster as cell-containing droplets.

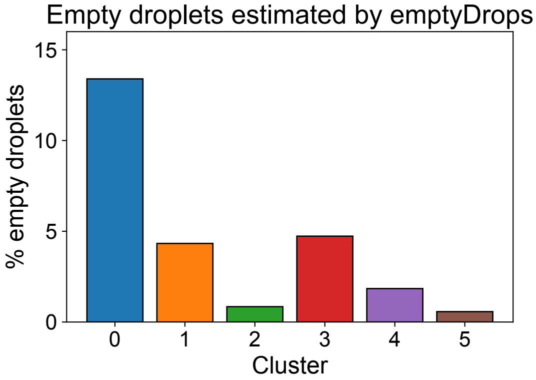

**Figure 5.13: Percentage of empty droplets per cluster estimated by emptyDrops.** The bar plot shows the percentage of empty droplets estimated by emptyDrops for each cluster from Figure 5.9. While cluster 0 has an elevated rate of empty droplets, the majority of cells in this cluster are labeled as “cell-containing droplets”.

SoupX also builds a background gene expression profile from cell barcodes with low UMI counts, under the assumption that these barcodes truly contain a background signal. The method then estimates contamination values for cluster-specific genes independently, where cluster-specific genes are differentially expressed in cells in the cluster compared to the background contamination distribution, and clusters are defined a priori. The method then sets a global contamination estimate to the mode of the contamination values estimated across each cluster-specific gene independently. These cluster-specific genes may also be user-specified based on known cell-type specific marker genes which should be present in the dataset. After this, SoupX removes the contaminating counts from cells one cluster at a time by distributing contaminating counts across all cells in a given cluster. This is estimated by multiplying the global contamination rate, total UMI counts across all cells in a given cluster, and the proportion of counts for a given gene in the background gene expression profile together, and subtracting these counts from each cell in the cluster individually. While SoupX can be useful to remove ambient mRNA contamination from individual cells to clean up the signal, it is not able to identify the droplets which contain purely ambient mRNA signal, in part due to the assumption that every droplet in the sample is affected by ambient mRNA at similar levels. When measuring the effective contamination per cell in the day 50 brain organoids dataset as the fraction of removed counts by application of SoupX, shown in Figure 5.14, cells from cluster 0 did not have an elevated effective contamination level.

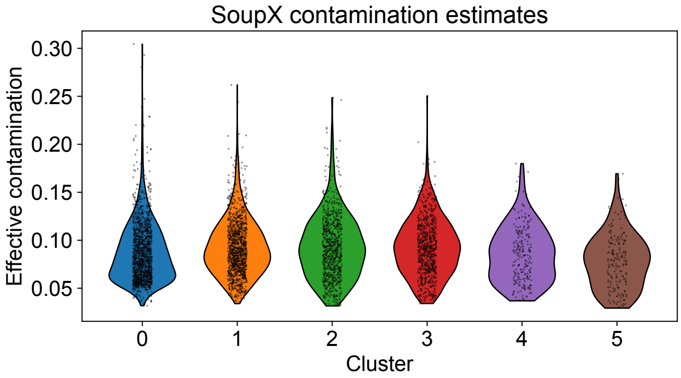

**Figure 5.14: SoupX contamination estimates.** The violin plot shows the effective contamination estimates (the fraction of removed counts) per cell using SoupX for each cluster from Figure 5.9.

Instead of using cell barodes with low UMI counts to build a background count distribution, DecontX builds the contamination distribution as a weighted combination of the count distribution across all other cell populations (clusters) in the sample. This can be especially beneficial if the full count matrix including cell barcodes with low UMI counts is not available. Essentially, DecontX treats the counts in each cell as a mixture of multinomial distributions over genes, one from the native cell population (cluster) and another from the contamination distribution, measured from all other clusters in the sample, and estimates the contamination level in each cell using variational inference. DecontX relies heavily on accurate clustering of the data a priori to estimate and remove ambient mRNA signal from individual cells, since the cluster based expression profiles will be used to generate the contamination count distribution rather than the background expression profile gathered from cell barcodes with low UMI counts. Minor changes in the input clustering to DecontX give rise to contradicting contamination estimates per cell, as shown in Figure 5.15. While DecontX confirms cluster 0 contains cells with elevated contamination levels when input clusters were generated using Louvain clustering and resolution parameter 0.3, this trend is not apparent with a resolution parameter of 0.4. While these results highlight the selective ability of DecontX to detect cells with high levels of ambient mRNA contamination, they also highlight the strong dependency DecontX has on the input cluster annotation.

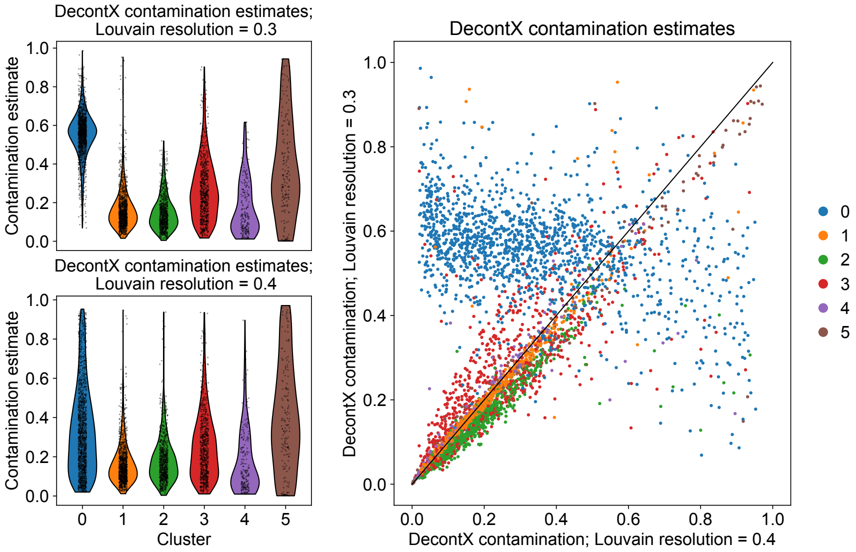

**Figure 5.15: DecontX contamination estimates.** The violin plots on the left show the contamination estimates using DecontX for each cluster from Figure 5.9, when using different input cluster annotations (Louvain resolution 0.3 for top violin plot, Louvain resolution 0.4 for bottom violin plot). The plot on the right compares these contamination estimates in a scatter plot, where cells are colored by cluster.

### 5.3.5 Application to inDrops scRNA-Seq of day 50 brain organoids

In order to test the utility of the procedure described in Section 5.3.2, the approach was applied to inDrops scRNA-Seq data of day 50 organoids derived from three different protocols - Triple-i, Dual SMAD-i and Inhibitor-free. In this analysis, doublets were detected using Scrublet and removed initially during pre-processing. Cells were then clustered and clusters subsequently annotated with a corresponding cell type, both with and without cells which were classified as empty droplets, in order to measure if the removal of empty droplets using this procedure enabled a more accurate interpretation of the data.

Clusters were annotated to corresponding cell types using the relative expression of curated cell-type and region-specific marker genes. The relative expression levels and annotated cell types are highlighted in Figure 5.16. When including empty droplets in the analysis, one cluster (cluster 0) contains predominantly empty droplets (94%) exclusively from Dual SMAD-i and Triple-i organoids. The cells in this cluster exhibit a strong enrichment of cortical genes, and express both neural stem cell and neuronal genes, and were therefore labeled as “cortical transitory” cells. However, there is no corresponding cluster which matches the expression profile of this cluster after removing all empty droplets, indicating that this cluster is most likely artifactual. The strong enrichment of cortical genes results from the ambient mRNA expression profiles of the Dual SMAD-i and Triple-i organoids, which contributed to this cluster. These organoids contain predominantly cortical cell types, and therefore the background ambient mRNA pool contains many cortical genes.

The retinal progenitor cluster (cluster 2) also has a high proportion of empty droplets (28%) contributed mainly by Inhibitor-free organoids. This cluster exhibits a strong expression of optic vesicle genes, as well as elevated expression levels of mesenchymal genes. After removal of empty droplets, the expression levels of the mesenchymal genes in the retinal progenitor cluster substantially decreased, indicating that filtering empty droplets removed this source of contamination. This example also highlights the importance of removing empty droplets within each sample separately. While the empty droplets from the Triple-i and Dual SMAD-i experiments have similar expression profiles and thus co-cluster, the empty droplets from Inhibitor-free organoids exhibit a distinct expression profile. This is due to the underlying heterogeneity of the cell type compositions in the different samples, resulting in distinct ambient mRNA expression profiles. Thus, it is essential to build a sample-specific background expression profile when performing the procedure described in Section 5.3.2. Based on these results, Triple-i organoids were comprised of 87% cortical cells, Dual SMAD-i organoids were comprised of 71% cortical cells, and Inhibitor-free organoids contained only 5% cortical cells. These results also agree with the findings from Chapter 4, namely that the Triple-i derivation protocol enriches for cortical populations.

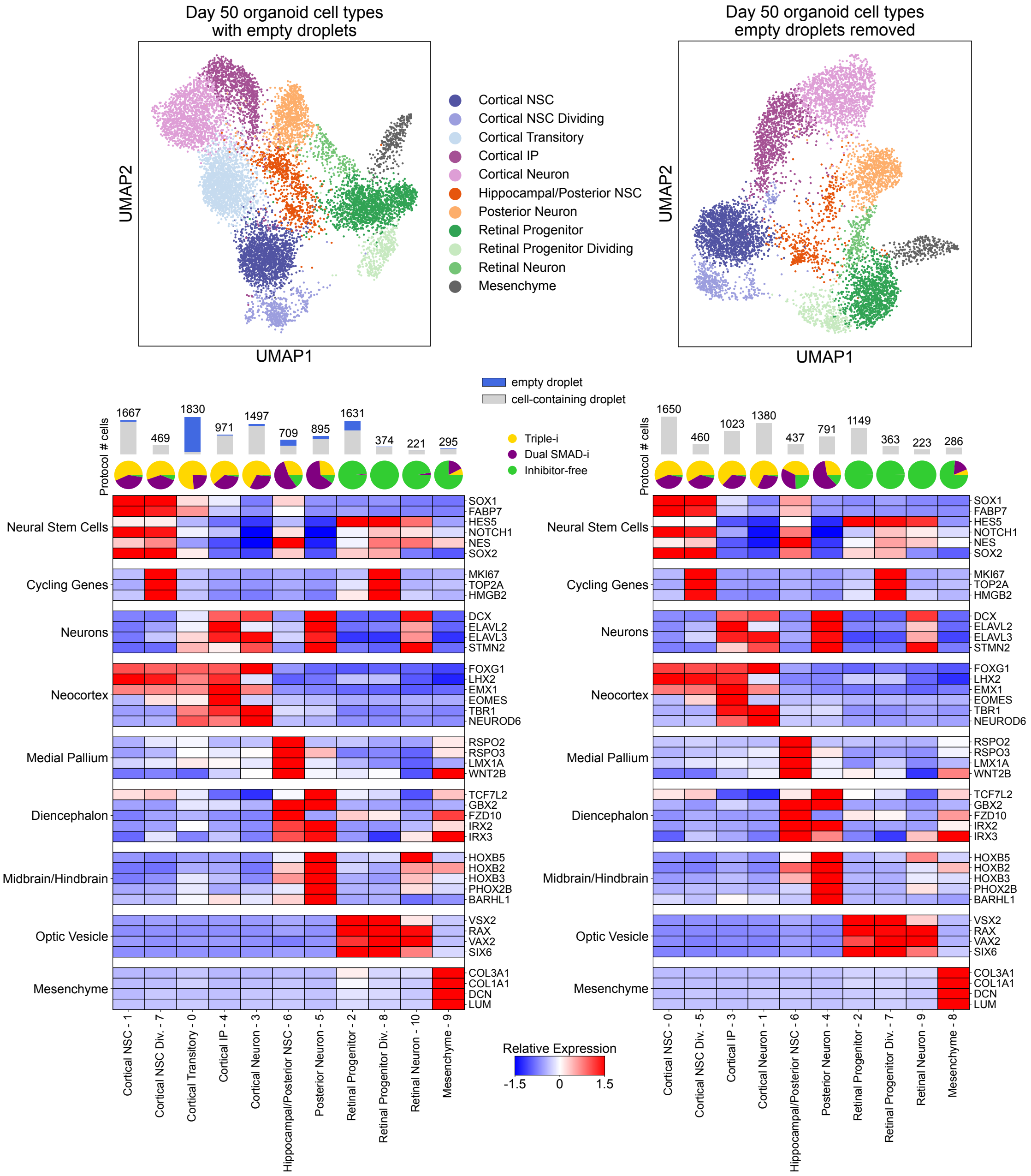

**Figure 5.16: Clustering inDrops scRNA-Seq of day 50 organoid samples with and without empty droplets.** The UMAP on the left highlights cells from inDrops scRNA-Seq datasets of day 50 brain organoids derived from Triple-i, Dual SMAD-i and Inhibitor-free protocols, annotated by their corresponding cell type, and including all empty droplets. The UMAP on the right highlights cell types after removal of empty droplets. The heatmaps below (left - with empty droplets; right - without empty droplets) highlight the relative expression levels of curated cell-type and region-specific marker genes using a z-score normalization of the mean expression levels across all cells in each cluster.

### 5.3.6 Application to 10X scRNA-Seq of day 50 brain organoids

The previous section highlighted the importance of detecting and removing empty droplets in scRNA-Seq data, and how this artifact can confound downstream analyses. The most widespread technology used to generate libraries for scRNA-Seq is developed by 10X Genomics, and initial pre-processing including alignment, filtering, barcode counting, UMI counting and cell barcode filtering are typically performed by the CellRanger software (Zheng et al., 2017). This pipeline incorporates a cell-containing droplet calling algorithm which works as follows. First, droplets with a high RNA content are called as cell-containing droplets. Then, the algorithm uses an approach based on the emptyDrops method (Lun et al., 2019) to determine if droplets with a lower UMI count have a different gene expression profile than the background expression profile, built from barcodes with low UMI counts. The cutoff used by CellRanger to define droplets with a high RNA content is larger than the estimated inflection point, as highlighted in Figure 5.17. This figure also highlights a stark difference in the distribution of UMI counts in each droplet compared to the inDrops protocol, as seen in Figure 5.2. For the inDrops protocol, the distinction between droplets with a high RNA content and those with a lower RNA content is less pronounced than in the 10X dataset.

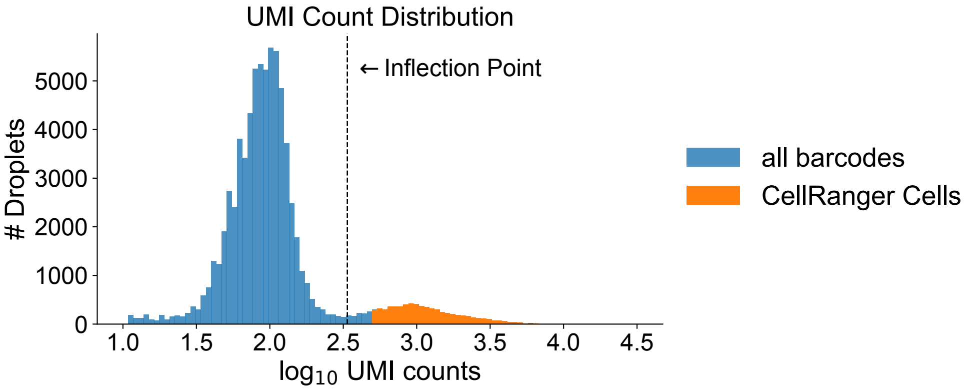

**Figure 5.17: Distribution of UMI counts per cell barcode in 10x scRNA-Seq.** The histogram displays the distribution of UMI counts across all droplets in a 10X scRNA-Seq dataset of day 50 brain organoids derived from the cell line ZIP8K8 using the Triple-i protocol. Cell-containing droplets called by CellRanger are highlighted in orange.

The approach described in Section 5.3.2 was run on a comprehensive dataset of 10 scRNA-Seq experiments of day 50 brain organoids generated using the 10X protocol (Rosebrock et al., 2022), and also described in Chapter 4. Table 5.1 highlights the percentage of annotated empty droplets across cells above the inflection point, as well as across cells called by CellRanger, across these 10X samples. For comparison, the percentage of empty droplets estimated in the inDrops experiment discussed in Section 5.3.5 are also highlighted.

Interestingly, the percentage of empty droplets across cells with a higher UMI count than the inflection point is substantially lower in the 10X datasets when compared to the inDrops datasets (final column in Table 5.1). This suggests that the two technologies differ in the relative capture rate of ambient mRNA in an individual droplet relative to the capture rate of mRNA from a cell. Also, as seen in Table 5.1, the percentage of empty droplets is substantially higher when including all cells above the inflection point compared with only including CellRanger cells. This is expected, since cell barcodes with a lower UMI count are more likely to contain purely ambient mRNA signal, and not an actual cell. Furthermore, as seen in Figure A2, one cluster (cluster 42) contained a substantial proportion of empty droplets (38%), while all other clusters contained $< 10\%$ empty droplets. Cells in this cluster also co-expressed neural stem cell, neuronal, and cell division genes, as highlighted in the heatmap, a further indication that this cluster may be comprised of empty droplets.

Including cells originally removed by CellRanger will increase the number of cells in a scRNA-Seq experiment, however, these cells have low UMI counts and hence a decreased level of signal. As an example, cells labeled as cell-containing droplets using our method that were originally removed by CellRanger, were re-introduced in the 10X scRNA-Seq dataset of ZIP8K8-derived day 50 Triple-i organoids. When merging these cells with the original dataset in a joint UMAP shown in Figure 5.18, they overlap well with the original CellRanger cells. Interestingly, a larger relative proportion of cells originally removed by CellRanger overlap with the cortical NSC population, and a smaller relative proportion overlap with the choroid plexus population. This is most likely a result of the differences in the UMI counts across cell types, with choroid plexus cells having a median UMI count of 2,286, and cortical NSCs having a median UMI count of only 891. These differences are also shown in Figure 5.18. Hence, cells originally removed by CellRanger will disproportionately contain cell types with a lower UMI count, either resulting from differences in the total amount of mRNA in cells from each population, or differences in the mRNA capture rates by the 10X technology in different cell types.

In some cases, cell populations with very low mRNA levels may be filtered out when excluding cell barcodes with low UMI counts, however, in this example of ZIP8K8-derived day 50 Triple-i organoids, no new cell populations were detected which were not present in the original dataset. Nonetheless, to get a more accurate estimate of the relative proportions of cell types present in a sample, it may be beneficial to include cell-containing droplets with lower UMI counts.

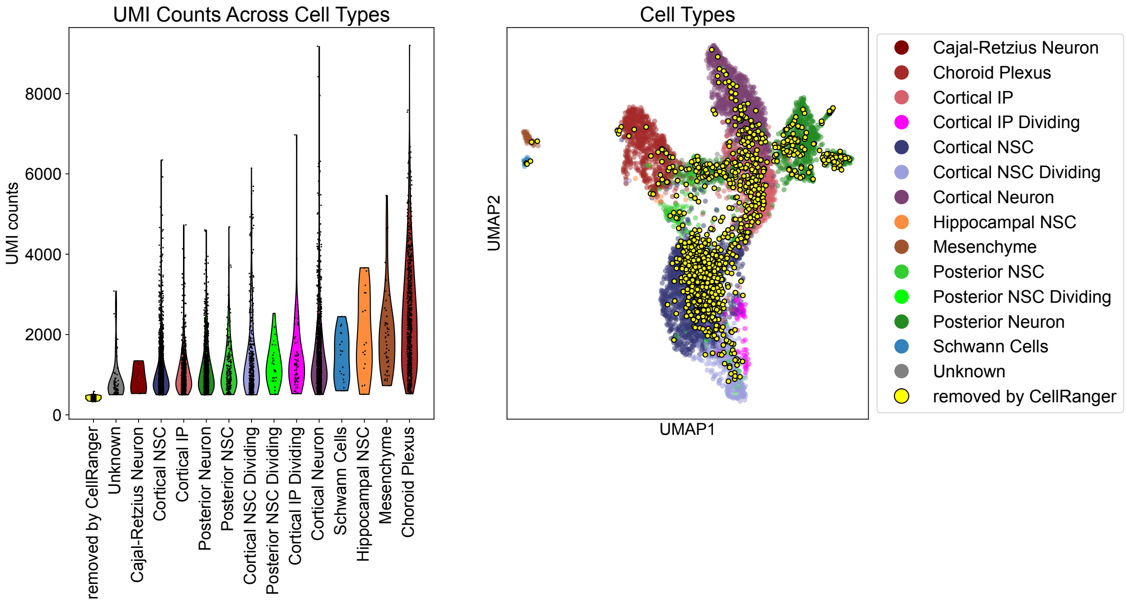

**Figure 5.18: Including cells with low UMI counts in 10X zip8k8-derived day 50 Triple-i organoids.** The violin plot on the left displays the UMI counts across individual cells grouped together by annotated cell type from the publication (Rosebrock et al., 2022), as well as cells originally removed by CellRanger. The UMAP on the right displays the joint dataset of cells from the original publication and those originally removed by CellRanger, which were detected as cell-containing by our method.

**Table 5.1:** Empty droplets in 10X and inDrops scRNA-Seq of day 50 brain organoids.

| Cell Line | Organoid Protocol | Library Generation Technology | # Cell-Ranger Cells | % Empty Droplets (Cell-Ranger Cells) | # Cells Above Inflection Point | % Empty Droplets (Cells Above Inflection Point) |
|---|---|---|---|---|---|---|
| ZIP8K8 | Triple-i | 10X Single Cell 3' v3 | 6683 | 0.70% | 7717 | 1.3% |
| ZIP8K8 | Dual SMAD-i | 10X Single Cell 3' v3 | 9479 | 2.11% | 11897 | 2.56% |
| ZIP8K8 | Inhibitor-free | 10X Single Cell 3' v3 | 13825 | 1.06% | 16300 | 2.30% |
| ZIP13K5 | Triple-i | 10X Single Cell 3' v3 | 7442 | 0.79% | 8286 | 1.01% |
| ZIP13K5 | Dual SMAD-i | 10X Single Cell 3' v3 | 6583 | 1.01% | 7672 | 1.56% |
| ZIP13K5 | Inhibitor-free | 10X Single Cell 3' v3 | 8763 | 6.47% | 10508 | 8.09% |
| KUCG2 | Triple-i | 10X Single Cell 3' v3 | 12604 | 1.98% | 16465 | 2.48% |
| KUCG2 | Dual SMAD-i | 10X Single Cell 3' v3 | 12171 | 3.86% | 18231 | 7.72% |
| FOK1 | Triple-i | 10X Single Cell 3' v3 | 11209 | 4.27% | 14619 | 9.52% |
| FOK1 | Dual SMAD-i | 10X Single Cell 3' v3 | 7695 | 1.56% | 12378 | 3.34% |
| ZIP8K8 + ZIP13K5 | Triple-i | inDrops v2 | N.A. | N.A. | 4684 | 31.55% |
| ZIP8K8 + ZIP13K5 | Dual SMAD-i | inDrops v2 | N.A. | N.A. | 3140 | 27.42% |
| ZIP8K8 + ZIP13K5 | Inhibitor-free | inDrops v2 | N.A. | N.A. | 2735 | 16.75% |

## 5.4 SUMMARY

In this chapter, we described two potential sources of technical artifacts that can arise in scRNA-Seq data, multiplets and ambient mRNA, and presents a method which is able identify droplets containing a purely ambient mRNA signal. This is accomplished by simulating empty droplets from the background transcriptomic profile, and co-clustering these simulated cells with the cells from the sample. This method is described in Section 5.3.2. When two or more cell lines, or cells with different genotypes, are present in a given sample, it is possible to genotype individual cells to a cell line of origin, as shown in Section 5.3.3, and thereby also identify empty droplets which contain a mixture of genetic material matching the background level of contamination. However, not all scRNA-Seq experiments consist of a mixture of cells with different genotypes, and therefore an approach to identify empty droplets using solely the expression data in the form of a count matrix is needed. By genotyping individual cells and using this as an orthogonal approach to identify empty droplets, we validate that co-clustering cells in the dataset with simulated empty droplets enables the identification of droplets which contain mainly ambient mRNA signal. By optimizing the average background connectivity enrichment in cells which co-cluster with simulated empty droplets ($D_r$ score) across various resolution parameters in Louvain clustering, we establish an automated pipeline to identify empty droplets in scRNA-Seq datasets.

As opposed to this binary classification type, the methods DecontX and SoupX measure contamination estimates per cell and remove the contaminated signal from the UMI count matrix. A third method, emptyDrops, is useful in rescuing cell barcodes with low UMI counts. When using these methods on the brain organoid dataset in Section 5.3.4, we find that SoupX and emptyDrops are unable to distinguish the majority of droplets which contain mainly ambient mRNA signal from those which contain a cell, and DecontX is highly dependent on input cluster annotation, producing conflicting results when the input clusters are slightly different. This underscores the utility of the orthogonal approach presented in Section 5.3.2 to detect and remove empty droplets.

In Section 5.3.5, we highlight the importance of addressing this artifact in a dataset of inDrops scRNA-Seq experiments of day 50 brain organoids. In these datasets, including empty droplets in the analysis led to the detection of a cluster with an enrichment of cortical genes, and mixed expression of neural stem cell and neuronal genes. We were able to classify the majority of cells in this cluster as empty droplets, and the filtering of empty droplets resulted in the removal of this cluster. Thus, if not addressed properly, empty droplets may result in the detection of artifactual cell populations which can have a profoundly negative effect on downstream analyses.

Finally, we compare the rates of empty droplets detected across different technologies, in particular inDrops and 10X, in Section 5.3.6. While the rates of empty droplets vary across samples from the same technology, there was a substantially higher rate of empty droplets in the inDrops datasets compare to the 10X datasets. We also highlight the utility of the approach to rescue cell barcodes with low UMI counts that were originally filtered from the dataset using the CellRanger pipeline.

In summary, careful pre-processing of a scRNA-Seq dataset is essential for accurate downstream analyses. This requires an intricate understanding of the possible error modes that can arise during library generation and sequencing. In the following section, we will discuss one of the downstream applications of scRNA-Seq data analysis, namely, the ordering of cells along a differentiation trajectory, and how to measure the dynamics of gene expression and gene interactions as cells undergo differentiation, with a particular emphasis on the differentiation of cortical neural stem cells into neurons within brain organoids.

# 6 MEASURING TRANSCRIPTIONAL CASCADES IN SCRNA-SEQ

One of the fundamental questions underlying developmental biology is how stem cells differentiate into specialized cell types. Changes in gene expression underlie the intrinsic molecular processes governing differentiation, enabling cells to change their morphology and function. These changes in part occur due to environmental cues from signaling molecules, which can activate or repress different transcription factors that are essential for the expression of certain lineage specific genes. These signaling molecules work by initiating a signal transduction pathway, which is typically induced by a ligand released by one cell binding to a receptor on the cell surface membrane of another cell, eventually resulting in the regulation of downstream target genes (Gilbert, 2009). Other environmental cues that can affect cellular differentiation include temperature (Wang et al., 2020) and oxygen levels in the organism's environment (Holzwarth et al., 2010). Changes in chromatin modification are also known to play an important role in regulating gene expression during cellular differentiation (Chen et al., 2014). Finally, asymmetric cell division, a cell division process which produces two daughter cells with different cellular fates, is another mechanism leading to cellular differentiation. Asymmetric cell division can broadly be categorized into two mechanisms, the first relying on the asymmetric distribution of cellular components, for example proteins and mRNA molecules, across the two daughter cells, and the second relying on the differential placement of the two daughter cells relative to external signaling cues (Morrison et al., 2006).

These mechanisms of gene regulation during cellular differentiation ultimately result in modifying the expression levels of genes which are critical for cell-fate specification. The most important genes for cell-fate specification are transcription factors, which can initiate or block the expression of downstream genes by binding to the DNA in the promoter region, or enhancer regions, of their target genes. Transcription factors form the key players in gene regulatory networks, which define the relationships between regulators of gene expression and their target genes. For example, the transcription factors PAX6, EOMES and TBR1 form an essential gene regulatory network underlying the differentiation of cortical neural stem cells into neurons via intermediate progenitor cells, as described in Section 2.2.3.

Single-cell RNA sequencing enables sampling the gene expression profile of thousands of cells in an individual sample. However, it is necessary to destroy the cell in order to measure its transcriptome, thereby making it impossible to observe how the cell and its gene expression profile would have altered in the future. Nonetheless, it is possible to order cells along a trajectory which accurately recapitulates the progression of cells as they differentiate. This ordering of the cells along a differentiation trajectory is known as pseudotemporal ordering, or pseudotime. Pseudotime is essentially a mapping of single cell transcriptomes to a developmental timeline. Pseudotime methods work under the assumption that cell state changes occur through transitional states, and that these can be measured as gradual shifts in gene expression in individual cells from scRNA-Seq datasets. There is a large variety of methods which estimate pseudotime from scRNA-Seq data (Campbell et al., 2019; Cao et al., 2019; Haghverdi et al., 2016; Lange et al., 2022; Setty et al., 2019; Street et al., 2018). For a comprehensive overview of different pseudotime algorithms, we refer the reader to (Saelens et al., 2019).

Based on pseudotemporal orderings of cells along a differentiation trajectory, it is possible to measure the dynamics of gene expression as cells differentiate. Current algorithms typically measure the dynamics of genes along pseudotemporal trajectories by fitting their expression profiles using generalized linear models (Berge et al., 2020; Cao et al., 2019; Ji et al., 2016), with the ultimate goal of determining if gene expression significantly varies as a function of pseudotime. Other methods attempt to measure pseudotime-dependent gene interactions by calculating a similarity measure between the expression levels of the "present" of one gene, and the "past" of another gene using correlation (Specht et al., 2016) or mutual information (Qiu et al., 2020). However, these methods do not return an explicit ordering of gene dynamics along a pseudotime trajectory, and require user-defined cutoffs for determining meaningful interactions.

In this chapter, we explore how to explicitly model gene expression over a pseudotime trajectory using a variety of functions that reflect biological state switches, and that model the dynamic behaviors of gene expression within cells as they differentiate. We formulate the problem in terms of a Bayesian inference problem and use MCMC to sample from the posterior distributions over the parameter space of the various functions. This provides an explicit ordering of genes along a pseudotemporal trajectory, enabling the description of gene dynamics in terms of transcriptional cascades, statistical testing for differences in switch times of gene expression, and annotation of potentially causal gene interactions in gene regulatory networks.

## 6.1 MODELING DYNAMICS OF GENE EXPRESSION ALONG PSEUDOTEMPORAL TRAJECTORIES

The ultimate goal of the method described in this section is to measure state changes in the expression levels of a given gene along a pseudotime trajectory, and pinpoint at what pseudotime these changes occur. From the pseudotemporal ordering of state changes, it is possible to deduce a wide variety of biologically meaningful interpretations, such as measuring the potential upstream regulators of a given gene of interest, measuring transcriptional cascades, and reconstructing causal gene regulatory networks. The initial input to the method consists of a set of cells ordered by their pseudotemporal ordering, $t = 1, ..., N$, and the expression levels of genes within those cells. Figure 6.1 highlights the expression levels of the key cortical neurogenesis transcription factors PAX6, EOMES and TBR1, in non-dividing cortical cell types in day 30 brain organoids after ordering the cells using diffusion pseudotime (Haghverdi et al., 2016). All dividing cells were excluded for the pseudotime estimation because they express a transcriptional program that is independent of the underlying cell type, which can potentially confound the pseudotime estimates.

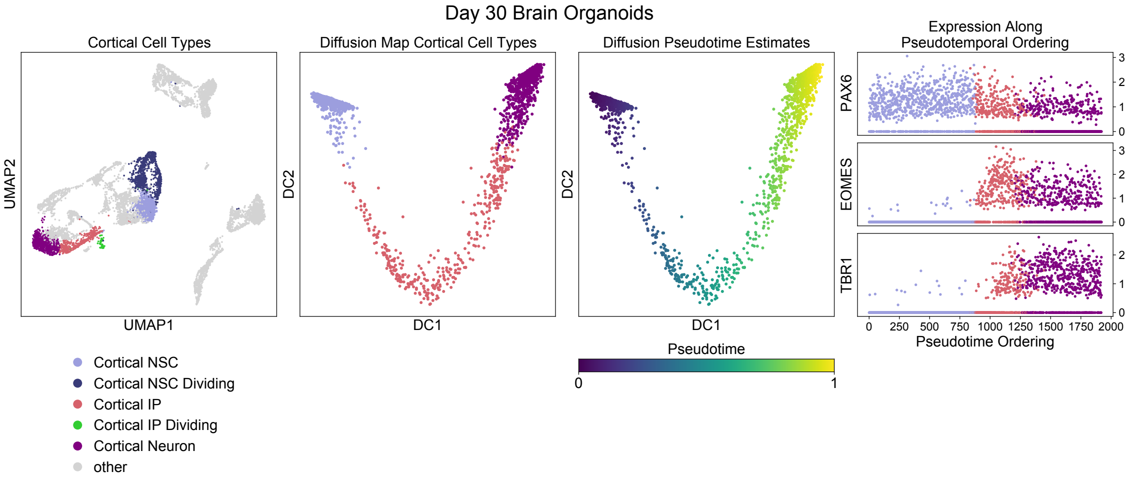

**Figure 6.1: Ordering cells along cortical neuronal differentiation trajectory.** The left-most panel displays a UMAP of cells in a scRNA-Seq experiment of day 30 brain organoids derived using the Triple-i protocol. Cortical cell types are highlighted in the plot, including dividing and non-dividing cortical neural stem cells (NSCs), dividing and non-dividing intermediate progenitors (IP), and neurons. All other cells are highlighted as "other". The second panel highlights a diffusion map embedding of the cortical NSC, IP and neuronal populations, after removal of all dividing cell types, with the third panel highlighting the same cells in the diffusion map colored by diffusion pseudotime estimates, scaled from 0 to 1. The rightmost panel displays the expression of key developmental transcription factors for cortical neurogenesis, PAX6, EOMES and TBR1, within cells after ordering them according to their relative pseudotemporal ordering using a fixed time step of 1, with cells colored by their respective cell type annotations.

From the expression levels of genes along a pseudotemporal ordering such as that in Figure 6.1, the goal of the method in this section is to decide if a state switch (up-to-down regulation or down-to-up regulation) occurs along the trajectory, and at what pseudotime these switches occur. In order to do this, first define a set of functions which can model a wide variety of expression dynamics, and for which state changes are well-defined and interpretable, namely at the inflection points of each function. The functions used are defined as follows,

$$\begin{aligned} f(x;a,b,x_0,\sigma) &= ae^{-\frac{(x-x_0)^2}{\sigma^2}}+b, \\ g(x;k,L,x_0,b_{min}) &= \frac{L}{1+e^{-k(x-x_0)}}+b_{min}, \\ h(x;k1,k_2,x_1,x_2,b_{min},b_{mid},b_{max}) &= b_{min}+\frac{b_{mid}-b_{min}}{1+e^{-k_1(x-x_1)}}+\frac{b_{max}-b_{mid}}{1+e^{-k_2(x-x_2)}}, \\ u(x) &= b. \end{aligned} \tag{6.1}$$

Here, $f(x)$ is a Gaussian function with parameter constraints $a > 0$, $b > 0$, $\sigma > 0$, and $1 \leq x_0 \leq N$, with $N =$ number of cells in the pseudotime trajectory. $g(x)$ is a sigmoidal function with parameter constraints $L > 0$, $b > 0$, and $1 \leq x_0 \leq N$. $h(x)$ is a double sigmoidal function with the formulation described in (Baione et al., 2021) and parameter constraints $b_{min} > 0$, $b_{mid} > 0$, $b_{max} > 0$, $k_1 > 0$, $k_2 > 0$, and $1 \leq x_1 < x_2 \leq N$. Finally, $u(x)$ is a uniform function with $b > 0$, which models the absence of dynamics in gene expression along a pseudotime trajectory. The motivation for using these functions is in part based on observations from biological scenarios during development (Bar-Joseph et al., 2012). For instance, genes in developmental studies can display a shift from one steady state to another, which can be modelled using a sigmoidal function. They can also exhibit impulse patterns of up-regulation followed by a return to basal levels, which can be modelled using a Gaussian function. Finally, double sigmoidal functions can model impulse patterns with asymmetric increase and decrease rates and different initial and terminal basal levels, as well as step-wise up-patterns and step-wise down-patterns. For example, Figure 6.2 highlights the types of expression dynamics which these functions can model, as well as what the individual parameters specify.

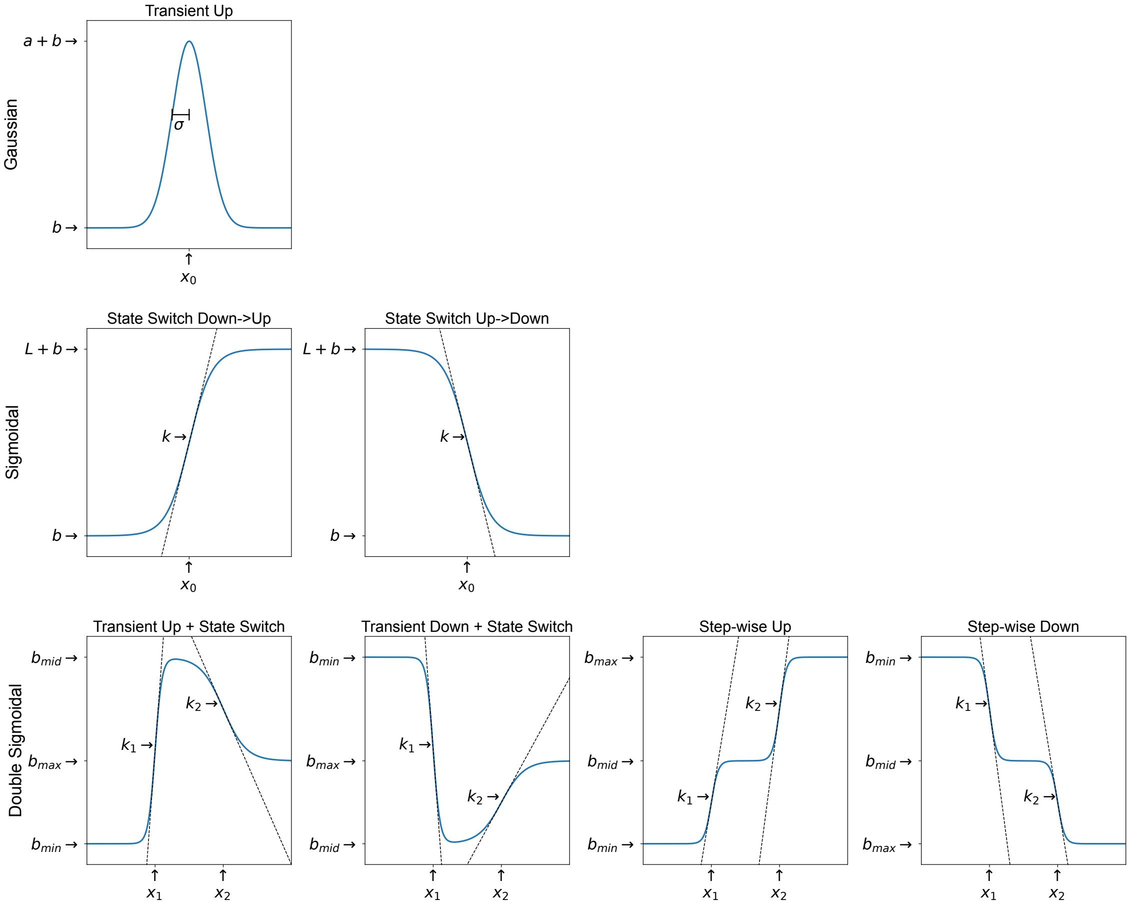

**Figure 6.2: Modeling gene expression dynamics.** Gaussian, sigmoidal, and double sigmoidal functions can be used to model a variety of expression dynamics along differentiation trajectories. The plots highlight the different expression dynamics which can be modeled with each function. For the Gaussian function, $x_0$ specifies the location of the peak, $b$ the basal expression level, $a + b$ the peak expression level, and $\sigma$ the standard deviation. For the sigmoidal function, $x_0$ specifies the location of the inflection point, $b$ the basal expression level, $L + b$ the peak expression level, and $k$ the rate of increase or decrease from basal to peak expression level. For the double sigmoidal function, $x_1$ and $x_2$ specify the location of the first and second inflection points, $b_{min}$, $b_{mid}$ and $b_{max}$ specify the initial, transitional, and final steady state expression levels, and $k_1$ and $k_2$ specify the rates of increase or decrease at each inflection point, or state change.

### 6.1.1 Establishing a likelihood model

The negative binomial distribution has been shown to accurately describe the count data generated in scRNA-Seq experiments without the need to account for zero-inflation resulting from "dropout" events (Svensson, 2020). The probability mass function for the negative binomial distribution can be parameterized using the mean, $\mu \in \mathbb{R}^+$, and dispersion parameter, $\phi \in \mathbb{R}^+$, with $y \in \mathbb{N}$, as follows:

$$p(y|\mu,\phi) = \binom{y+\phi-1}{y}\left(\frac{\mu}{\mu+\phi}\right)^y\left(\frac{\phi}{\mu+\phi}\right)^\phi. \tag{6.2}$$

The mean and variance of the random variable $Y \sim \text{NB}(\mu,\phi)$ which follows a negative binomial distribution is then $\mathbb{E}[Y] = \mu$ and $\text{Var}[Y] = \mu + \frac{\mu^2}{\phi}$. For a gene with measured counts of $\vec{Y} = \{y_t\}_{t=1,\dots,N}$ along a pseudotime trajectory with fixed pseudotime-step interval, $\vec{\mu} = \{\mu_t\}_{t=1,\dots N}$ and $\vec{\phi} = \{\phi_t\}_{t=1,\dots,N}$ the mean and dispersion at corresponding pseudotimes, the full likelihood of observing $\vec{Y}$ is:

$$\mathcal{L}(\vec{Y}|\vec{\mu},\vec{\phi}) = \prod_{t=1}^{N} p(y_t|\mu_t,\phi_t), \tag{6.3}$$

where $p(y_t|\mu_t,\phi_t)$ is the negative binomial probability mass function. The full log-likelihood is then:

$$\ln\left(\mathcal{L}(\vec{Y}|\vec{\mu},\vec{\phi})\right) = \sum_{t=1}^{N} \ln(p(y_t|\mu_t,\phi_t)). \tag{6.4}$$

Thus, the problem of fitting a curve to the pseudotime-ordered expression profile of a gene can be formulated as solving for $\mu(t)$.

### 6.1.2 Estimation of global dispersion parameter

It was shown that when fitting scRNA-Seq UMI count data to a negative binomial model, data are consistent with a global dispersion parameter independent of the expression level of a given gene, and that fitting a dispersion parameter to each gene individually leads to overfitting (Lause et al., 2021). Therefore, a global estimate of $\phi$ can be used for every gene independent of pseudotime, and $\vec{\phi} = \{\phi_t\}_{t=1,\dots,N}$ is replaced with a constant $\phi$ in Equation 6.4. A dataset specific $\phi$ using genes which exhibit lower levels of overdispersion is estimated, since the expression levels in these genes reflect the technical rather than the biological variability. To do this, the log10 mean counts for each gene are binned into five equally spaced bins, and a linear fit between log10 mean and log10 variance of counts in each bin is estimated. Genes within the top 20th percentile of the difference between the estimated variance and the expected variance using the linear fit in each bin are then filtered. The remaining genes are used to fit the non-linear relationship between the mean ($\mu$) and variance ($\sigma^2 = \mu + \frac{\mu^2}{\phi}$) using unconstrained non-linear least squares. Figure 6.3 displays the least-squares fit for $\phi$ among the non-dividing cortical cells highlighted in Figure 6.1.

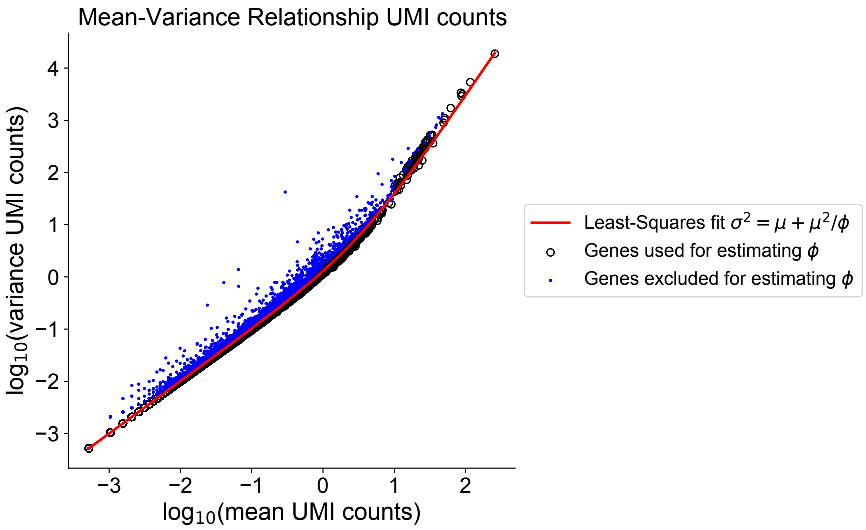

**Figure 6.3: Estimating $\phi$ in scRNA-Seq UMI count data.** A least-squares fit of the form $\sigma^2 = \mu + \frac{\mu^2}{\phi}$ in non-dividing cortical cells from 10X scRNA-Seq of day 30 brain organoids. Fits are performed in the raw space, and are plotted after log10 transformation of both mean and variance of UMI counts.

Here, $\phi$ estimates the dispersion based on genes which do not exhibit high variability in the dataset, and therefore captures the technical variability in the dataset. This technical variability is in large part driven by the varying number of UMI counts captured in each cell, as well as other factors including library quality and amplification bias. Thus, the full log-likelihood of observing counts $\vec{Y} = \{y_t\}_{t=1,..,N}$ along a pseudotime trajectory given the mean at corresponding pseudotime points $\vec{\mu} = \{\mu_t\}_{t=1,...N}$, becomes

$$\ln\left(\mathcal{L}(\vec{Y}|\vec{\mu}, \phi)\right) = \sum_{t=1}^{N} \ln(p(y_t|\mu_t, \phi)), \tag{6.5}$$

where $\phi$ is a global parameter estimated using the procedure described above.

### 6.1.3 Mapping from log-normalized to count space

For scRNA-Seq methods which sequence only from one end of the transcript and not full-length protocols, normalization does not need to account for the total transcript length. In this case, for a given cell $i$, let $K_i$ be a size factor for cell $i$, $T_i$ be the number of UMIs in cell $i$, and $y_{gi}$ be the number of UMIs for gene $g$ in cell $i$. The log-normalized expression levels for gene $g$ in cell $i$ is then defined as,

$$x_{gi} = \ln\left(\frac{y_{gi}}{T_i}K_i + 1\right), \tag{6.6}$$

where a pseudocount of 1 has been added to the normalized expression. The inverse relationship from log-normalized expression space to UMI count space is then,

$$c_{gi} = \frac{T_i}{K_i}\left(e^{x_{gi}} - 1\right). \tag{6.7}$$

Note that for an arbitrary $x_{gi}$, the value $c_{gi}$ is not guaranteed to be an integer. However, since this mapping is used to convert the mean expression level in log-normalized space to the mean expression level in count space, which is then used to parameterize $\mu \in \mathbb{R}^+$ in the negative binomial model, $c_{gi}$ is not constrained to $\mathbb{N}$ for this purpose. In widely used normalization techniques, a fixed $K_i$ is used for all cells, such as 10e6 for counts per million (CPM) normalization, the median number of UMIs across all cells in the dataset (Wolf et al., 2018) or 10,000 (Butler et al., 2018). Other methods for estimating cell-specific size factors exist, for example scran (L. Lun et al., 2016), which estimates size factors by pooling together cells with similar UMI counts in order to overcome issues arising from the dominance of low and zero counts. Pool-based size factors are then deconvolved into cell-specific size factors by solving a linear regression model over all genes. Normalizing for the total number of UMI counts in cell $i$, $T_i$, is not needed when using scran, as the size factor takes this into consideration.

Given a function, $q(t), t = 1, ..., N$, which models the log-normalized expression values $\{X_t\}_{t=1,...,N}$, the mean function $\mu_t$ is defined by mapping the expression values in log-normalized expression space evaluated at $t = 1, ..., N$ back to count space using the mapping in Equation 6.6. The full log-likelihood estimate is then evaluated by plugging in the estimates $\mu_t$ and global estimate for $\phi$ into Equation 6.5. Fitting the functions in log-normalized expression space is necessary, as the log-transformation reduces the impact of large stochastic fluctuations in the count data, which can have a large influence on parameter inference.

### 6.1.4 Model inference using MCMC

Under the framework presented above, solving for $\mu(t)$ can be formulated as a Bayesian inference problem, which will be estimated using an MCMC approach. MCMC provides an estimate of the posterior distribution over the parameter space for each of the parameters in the different functions defined in Equation 6.1. For each of the models, the priors used for the different parameters are summarized in Table 6.1.

| Uniform | Gaussian | Sigmoidal | Double Sigmoidal |
|---|---|---|---|
| $b \sim \text{Unif}(0, M)$ | $b \sim \text{Unif}(0, M)$<br>$a \sim \text{Unif}(0, M)$<br>$x_0 \sim \text{Unif}(1, N)$<br>$\sigma \sim \mathcal{FN}(0, N/10)$ | $b \sim \text{Unif}(0, M)$<br>$L \sim \text{Unif}(0, M)$<br>$x_0 \sim \text{Unif}(1, N)$<br>$k \sim \mathcal{FN}(0, 0.1)$ | $b_{min} \sim \text{Unif}(0, M)$<br>$b_{mid} \sim \text{Unif}(0, M)$<br>$b_{max} \sim \text{Unif}(0, M)$<br>$x_1 \sim \text{Unif}(1, N)$<br>$x_2 \sim \text{Unif}(1, N)$<br>$k_1 \sim \mathcal{FN}(0, 0.1)$<br>$k_2 \sim \mathcal{FN}(0, 0.1)$ |

**Table 6.1: Priors on Function Parameters.** $M = \max\left(\{X_t\}_{t=1,\dots,N}\right)$ = maximum expression level in log-normalized space across all cells. Unif$(a, b)$ refers to the uniform distribution on the open interval $(a, b)$, and $\mathcal{FN}(\mu, \sigma^2)$ refers to a folded normal distribution with parameters $\mu$ and $\sigma$, as defined in 6.8.

Note, in Table 6.1, the folded normal distribution is parameterized by $\mu > 0$ and $\sigma > 0$ with probability density function,

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} + \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x+\mu)^2}{2\sigma^2}}. \tag{6.8}$$

The uniform priors in Table 6.1 are uninformative, however, they provide bounds on the parameters to keep them in interpretable and meaningful ranges. The slope parameters $k$ in the sigmoidal function, and $k_1$ and $k_2$ in the double sigmoidal function, have a folded normal prior with 0-mean and 0.1 variance, which is used to ensure that the slope has a low magnitude. This prior is used because differences in the function once the slope becomes relatively large are minimal. Finally, the folded normal prior on $\sigma$ in the Gaussian with 0-mean and $N/10$ variance is used to ensure that the curve does not become very flat.

A variety of MCMC algorithms exist, some of which were discussed in detail in Section 3.1.3. In this section, the affine-invariant ensemble sampler proposed by Goodman & Weare in 2010 (2010) was used with implementation by Foreman-Mackey et al. (2013). This method was described in detail in Section 3.1.3. An initial guess is needed as a starting point from which a walker begins in the ensemble sampler. For the Gaussian and sigmoidal functions, initial guesses are derived from a non-linear least squares fit for each function on the log-normalized pseudotime expression levels, with added Gaussian noise. For the double sigmoidal function, initial guesses are randomly chosen to cover the varieties of different forms the functions can have. For the uniform function, initial guesses are randomly chosen from a uniform distribution over the interval 0.01 and maximum expression level for the gene of interest. A separate MCMC is run for each of the functions. The number of walkers used is four times the number of parameters for each function - 28 for the double sigmoidal fit, 16 for the Gaussian fit, 16 for the sigmoidal fit, and 4 for the uniform fit. This enables a wide sampling across the search space of parameters.

The MCMC is then run for a total of 10,000 iterations. There is generally no consensus on how many iterations to run an MCMC algorithm (Foreman-Mackey et al., 2013). Thousands of iterations are typically desirable to allow the process to reach a steady-state. After reaching the steady-state, the MCMC will sample from the posterior distribution over the parameter space, enabling an estimate of the posterior distribution for each parameter. Iterations before reaching the steady-state are discarded, as these are not sampled from the target distribution. This is called the "burn-in" phase. For this implementation, a burn-in of $5,000$ iterations was used. An example MCMC trace for the double sigmoidal fit for EOMES is shown in Figure 6.4.

Some MCMC walkers can get stuck near a local maximum. These walkers typically have a low acceptance rate, that is the proportion of moves for which the MCMC sampler generated parameter values that differed from the previous sample. One common practice is to prune these walkers from the final MCMC output. For example, walkers can be pruned which get stuck in irrelevant local optima by clustering the likelihood of the walkers and removing the clusters with lower likelihoods (Hou et al., 2011). For this implementation, half of the MCMC walkers are pruned with the lowest acceptance rate in order to remove potentially stuck walkers. Figure 6.5 highlights the acceptance rates across individual walkers for double sigmoidal, Gaussian, sigmoidal and uniform fits for EOMES.

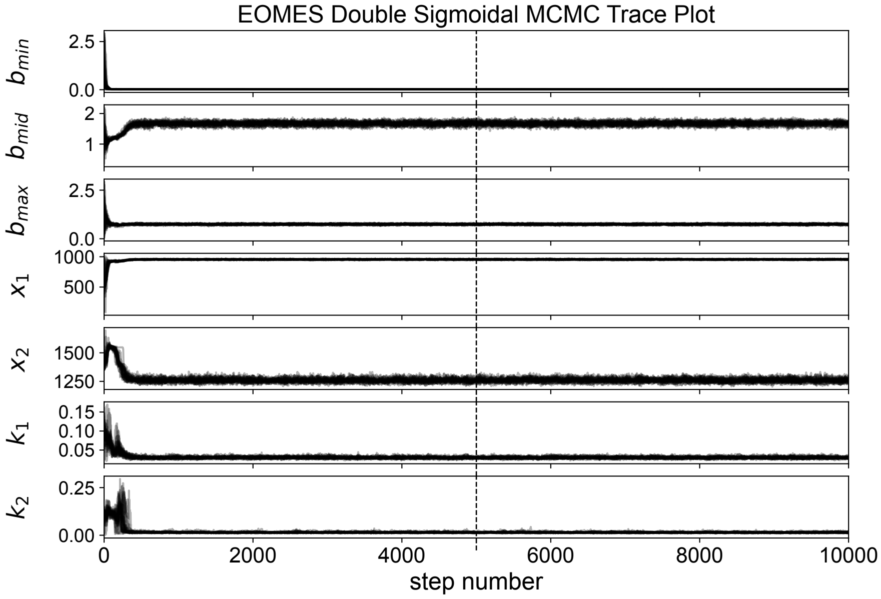

**Figure 6.4: MCMC trace plot for EOMES in cortical cells derived from day 30 brain organoids.** MCMC trace plot for parameters $b_{min}$, $b_{mid}$, $b_{max}$, $x_1$, $x_2$, $k_1$, and $k_2$ in the double sigmoidal MCMC ensemble sampler for EOMES. The dashed line represents the burn-in time at 5,000 iterations. In this example, only 500 iterations are needed for the sampler to reach a steady-state.

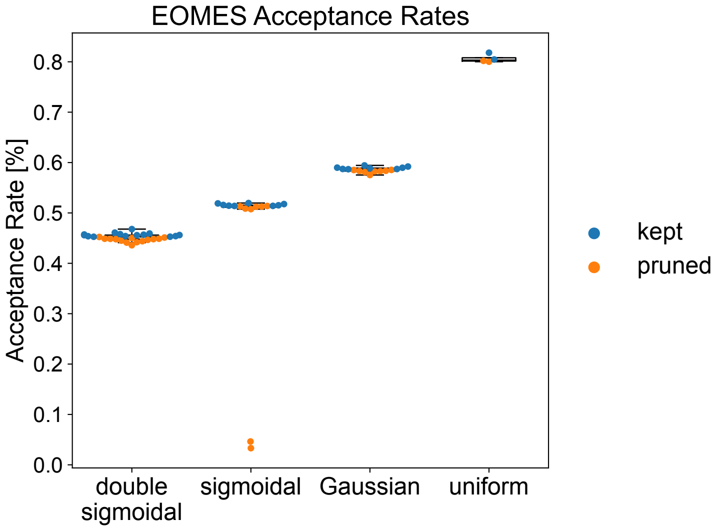

**Figure 6.5: Acceptance rates for individual MCMC walkers for EOMES in cortical cells derived from day 30 brain organoids.** The boxplots highlight the distribution of acceptance rates for each walker in each MCMC run. Walkers with the lowest acceptance rates (plotted in orange) are pruned and discarded from downstream analyses.

Each of the MCMC runs produces a sampling from the posterior distribution over the parameters for each model. Using the likelihoods of observing the data given the sampling over the parameter space, it is possible to compare the different models to choose an optimal model to fit the data. The next section discusses how to choose the optimal model from the MCMC outputs.

### 6.1.5 Model Selection

The distribution over likelihoods of observing the data given the parameter selections across the MCMC runs can be used to measure a goodness of fit for each model. Figure 6.6 highlights the distribution of log-likelihoods for each of the MCMC runs.

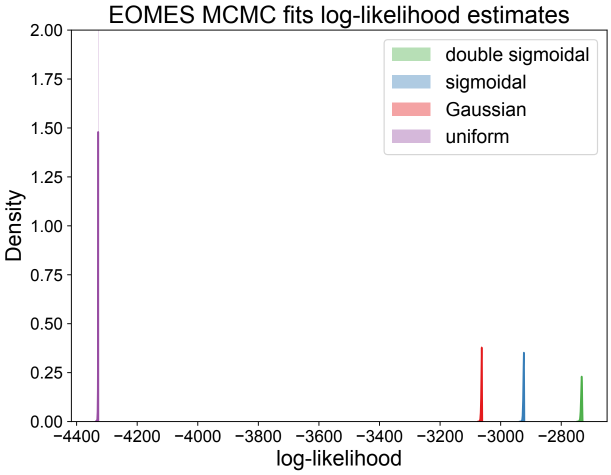

**Figure 6.6: Distributions of log-likelihoods for MCMC fits for EOMES in cortical cells derived from day 30 brain organoids.** The plot displays the distributions of log-likelihood estimates in individual iterations across all MCMC samplers after removing the first 5,000 iterations (burn-in), and removing the samplers in the bottom half of acceptance rates. The double sigmoidal fit has the highest log-likelihood estimates.

Simply comparing the log-likelihoods between the different models and choosing the model with the highest modal log-likelihood across MCMC iterations is one solution to select for the best model. However, it is advantageous to penalize models which are more complex than others in order to prevent overfitting. The different models vary in their complexity, which can be measured by the number of parameters to be specified in each model. For the double sigmoidal model, seven parameters are used. For the Gaussian and sigmoidal models, four parameters are used. And for the uniform model, only one parameter needs to be specified. Using probabilistic model selection techniques such as the Akaike information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) (Schwarz, 1978), or deviance information criterion (DIC) (Gelman, 2014), it is possible to score the different models using the log-likelihood estimates with the addition of a penalty term which compensates for overfitting with more complex models, and are defined as follows,

$$\begin{aligned} \text{AIC} &= 2k - 2\ln\left(\hat{L}\right), \\ \text{BIC} &= k\ln(n) - 2\ln\left(\hat{L}\right), \\ \text{DIC} &= 2\text{var}\left[\ln(L)\right] - 2\left\langle \ln(L) \right\rangle, \end{aligned} \tag{6.9}$$

where $n$ = number of data points, $k$ = number of parameters in the model, $\hat{L}$ = maximized value of the likelihood function, var $[\ln(L)]$ = variance of log-likelihood estimates across MCMC iterations, and $\langle \ln(L) \rangle$ = mean of log-likelihood estimates across MCMC iterations. In the original formulation of the AIC and BIC, the value $\hat{L}$ was derived from maximum likelihood estimation. When using an MCMC for model inference, the output consists of a sampling or distribution over the parameter space. It is advantageous to use a likelihood estimate which more closely reflects the optimal parameter regime estimated from the MCMC instead of the parameter regime which maximizes the likelihood. To this end, $\hat{L}$ in the AIC and BIC is replaced with $P(y| \langle \theta \rangle)$, the likelihood of observing the data given $\langle \theta \rangle$, where $\langle \theta \rangle$ = mean over the parameter estimates across all MCMC iterations. However, the AIC and BIC are still constructed from point estimates of the likelihood function. The DIC is particularly well-suited for model comparison directly from the MCMC results, since it is directly estimated from the log-likelihoods across all MCMC iterations. However, the DIC is prone to choosing more complex models, resulting in overfitting (Spiegelhalter et al., 2014). The BIC is more selective than the AIC since the penalty term is higher in the BIC for more complex models when $n$ is large.

When using these different criteria to measure the best-fitting model to the data, the BIC assigns 3,625 out of 13,736 genes (26%) expressed in at least 1% of cells to a non-uniform fit, the AIC assigns 5,368 out of 13,736 genes (39%) to a non-uniform fit, and the DIC assigns 6,004 out of 13,736 genes (44%) to a non-uniform fit.

To improve the generalizability of a model fit to a dataset, and remove the bias of outliers, another approach called cross-validation can be used. This method consists of splitting the dataset into a training set and test set, where model inference is run on the training set only, and a goodness of fit is measured based on the predictions made using the model on the test set. This provides an idea of how well a model will perform on unobserved data points. Variations of cross-validation include leave-one-out cross validation (LOO-CV) in which one data point is selected for the test set, and leave-p-out cross validation (LpO-CV) in which $p$ data points are selected for the test set. The dataset is then split into many partitions and model inference run on these partitions to get a more reliable estimate using cross-validation. For this application, model inference using MCMC is time-consuming, and therefore running an LpO-CV or LOO-CV on many subsets of the data is impractical. Thus, a variation of cross-validation is used here for model selection, highlighted in Algorithm 4.

Note, in Algorithm 4 the BIC is used for model selection due to its selective ability to favor less complex models, thereby reducing the problem of overfitting. Instead of cross-validating a model estimated from a training set on a test set, the full dataset is used for model inference and tested on random subsets of the dataset. When using this approach to select the model with the best fit, 1,475 out of 13,736 genes (11%) were assigned to a non-uniform fit. Figure 6.7 highlights a random sampling of the parameters over the MCMC runs using a double sigmoidal, Gaussian, sigmoidal and uniform model, as well the BIC estimates on random 98% subsets of the data.

It is worth noting that the double sigmoidal function can also closely take the form of the Gaussian and sigmoidal functions. It would be possible to use the double sigmoidal function alone, instead of including the Gaussian and sigmoidal functions, to model the dynamics of gene expression. However, the double sigmoidal function will force the presence of two inflection

```
Algorithm 4: Perform Model Selection based on MCMC runs.
 1: Measure average parameter estimates, $\langle\theta\rangle$, across MCMC runs for each model.
 2: Remove 2% of the data chosen randomly ($y_{sub}$), and estimate BIC for each model using
    $P(y_{sub}|\langle\theta\rangle)$.
 3: Repeat Step 2. for 10,000 subsets. Define $\mathrm{BIC_x}$ as the set of BIC estimates across all 10,000
    subsets for a given fit, and $\langle\mathrm{BIC_x}\rangle$ as the mean BIC estimate across all 10,000 subsets.
 4: if $\max(\mathrm{BIC_{double\ sigmoidal}}) < \min(\mathrm{BIC_{uniform}})$ & $\left\langle\mathrm{BIC_{double\ sigmoidal}}\right\rangle < \left\langle\mathrm{BIC_{gauss}}\right\rangle$ &
    $\left\langle\mathrm{BIC_{double\ sigmoidal}}\right\rangle < \left\langle\mathrm{BIC_{sigmoidal}}\right\rangle$ then
 5:     Set best fit to double sigmoidal.
 6: else if $\max(\mathrm{BIC_{sigmoidal}}) < \min(\mathrm{BIC_{uniform}})$ & $\left\langle\mathrm{BIC_{sigmoidal}}\right\rangle < \left\langle\mathrm{BIC_{gauss}}\right\rangle$ then
 7:     Set best fit to sigmoidal.
 8: else if $\max(\mathrm{BIC_{gauss}}) < \min(\mathrm{BIC_{uniform}})$ & $\left\langle\mathrm{BIC_{gauss}}\right\rangle < \left\langle\mathrm{BIC_{sigmoidal}}\right\rangle$ then
 9:     Set best fit to Gaussian.
10: else
11:     Set best fit to uniform.
12: end if
```

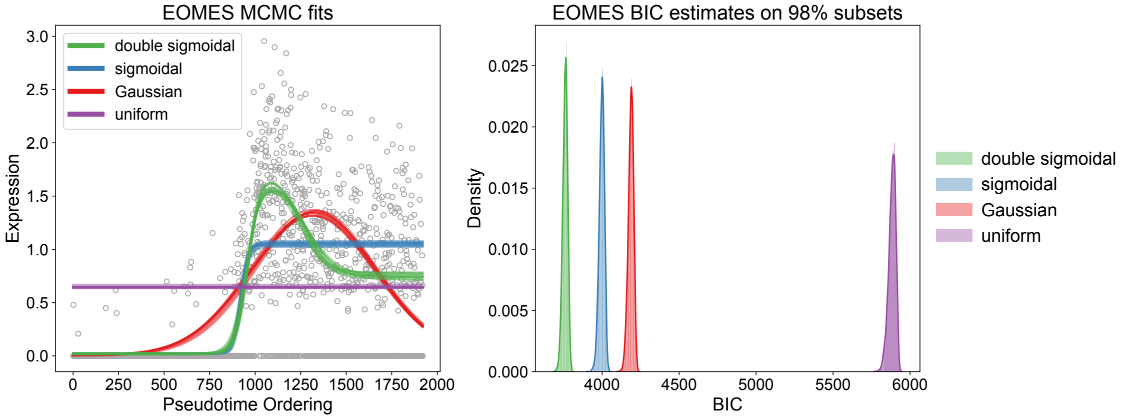

**Figure 6.7: MCMC fits and BIC estimates for EOMES in cortical cells derived from day 30 brain organoids.** The left panel in the plot displays a random sampling of the parameters from 100 iterations from the MCMC traces across all samplers for the double sigmoidal, Gaussian, sigmoidal and uniform MCMC runs, as well as the expression levels of EOMES in cells ordered according to their relative pseudotemporal ordering. The right panel highlights the BIC estimates over 10,000 subsets of the data removing a random 2% of the data. Using this approach, the doube sigmoidal model is selected as the best-fitting model.

points, whereas with the sigmoidal function will only have one inflection point, which in many cases more accurately models the gene expression dynamics of single a state-switch. Finally, a simpler model is often more favorable to use than a more complex model to prevent overfitting, and in the cases where a Gaussian function provides an equally good fit as the double sigmoidal function, then the selection of the simpler Gaussian model is preferred.

### 6.1.6 MCMC Diagnostics

In order to ensure that the MCMC adequately approximates the posterior distribution over the parameter space, a variety of heuristics exist. The MCMC trace plot shown in Figure 6.4 provides a quick visual inspection of whether the MCMC appears to have reached a steady-state. Also, the acceptance fraction across MCMC chains, with an example shown in Figure 6.5, is used to filter potentially stuck MCMC walkers. In general, there is no way to prove convergence of an MCMC sampler (Hogg et al., 2018), and therefore diagnostics are used to measure how well an MCMC run has converged to an equilibrium or steady-state. A few diagnostics are highlighted in this section to show the ability of the ensemble sampler described above to adequately converge to the posterior distribution over the parameter space.

One diagnostic metric relies on the estimate of the integrated autocorrelation time, which estimates the number of iterations needed for the MCMC to draw an independent sample. In the case of samples generated by an MCMC, the samples are not independent. This is due to the nature of the Markov process used to sample from the posterior distribution, which is dependent on the previous sampling of parameters, by definition. The integrated autocorrelation time is defined as,

$$\tau_f = \sum_{\tau=-\infty}^{\infty} \rho_f(\tau) = 1 + 2\sum_{\tau=1}^{\infty} \rho_f, \tag{6.10}$$

where $\rho_f(\tau)$ is the autocorrelation function at time delay $\tau$. Then, the effective sample size (ESS), i.e. the number of i.i.d. draws from the posterior distribution, for an ensemble sampler can be calculated as,

$$\text{ESS} = \frac{MN}{\tau_f}, \tag{6.11}$$

where $M$ = number of walkers, and $N$ = number of MCMC iterations used after discarding the burn-in. In order to estimate $\tau_f$, the marginal autocorrelation function for each parameter in the model can be estimated separately out to a certain time delay, $T$, using the average estimate across all walkers, and taking the maximum estimate of $\tau_f$ over all $T$, defined as

$$\hat{\tau}_f = \max_T \left( 1 + 2\sum_{\tau=1}^{T} < \rho_f(\tau) > \right). \tag{6.12}$$

Here, $T \in [0, 1000]$ enables an accurate estimate of $\hat{\tau}_f$ under the assumption that $\rho_f(\tau)$ approaches 0 by $\tau = T$ for each parameter. The autocorrelation function is estimated for each parameter separately. As an example, Figure 6.8 shows the estimates of $\rho_f(\tau)$ for $\tau \in [0, 1000]$ across each parameter in the model, and Figure 6.9 shows the estimates of $\tau_f$ for $T \in [0, 1000]$, with the final estimate $\hat{\tau}_f$ highlighted as a horizontal line. For this example, the average autocorrelation time across all parameters was estimated to be 96.6, equivalent to an effective sample size of approximately 725.

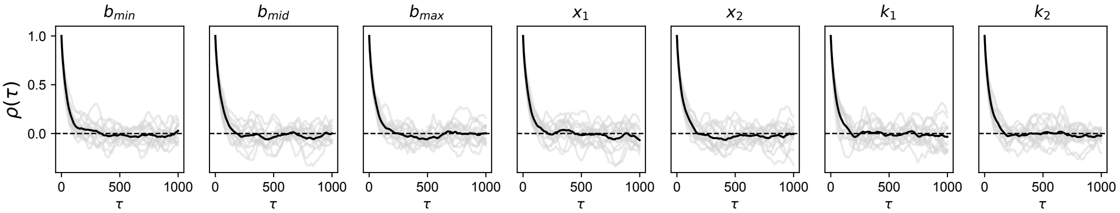

**Figure 6.8: $\rho_f(\tau)$ estimates in MCMC double sigmoidal run for EOMES in cortical cells derived from day 30 brain organoids.** The autocorrelation function, $\rho_f(\tau)$, is displayed in the y-axis at time lags $\tau \in [0, 1000]$ for each parameter across all MCMC walkers after removing the first 5,000 iterations (burn-in), and removing the samplers in the bottom half of acceptance rates. Note that the autocorrelation function decays to 0 after a certain time delay, $\tau$. Individual walkers are displayed in gray, with the average across all walkers in black.

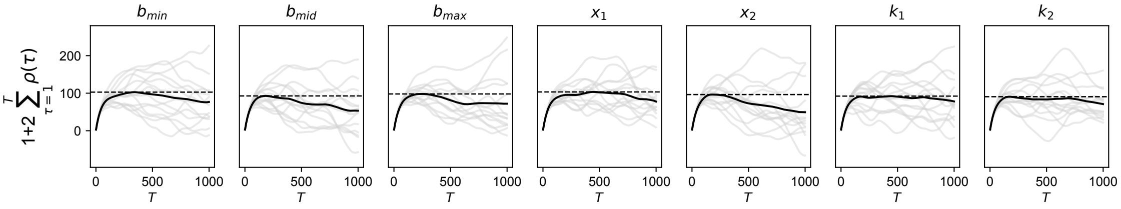

**Figure 6.9: $\hat{\tau}_f$ estimates in MCMC double sigmoidal run for EOMES in cortical cells derived from day 30 brain organoids.** The integrated autocorrelation time, $\tau_f$, is displayed in the y-axis estimated at time lags $T \in [0, 1000]$ for each parameter across all MCMC walkers after removing the first 5,000 iterations (burn-in), and removing the samplers in the bottom half of acceptance rates. Individual walkers are displayed in gray, with the average across all walkers in black. The final estimate, $\hat{\tau}_f$, as defined in Equation 6.12, is highlighted with a dashed line. Note that the integrated autocorrelation time is quite similar across all parameters, and ranges from 90 to 103.

For a general comparison, the autocorrelation times were estimated for all genes using the model with the best fit. Figure 6.10 highlights these estimates. The autocorrelation times increase with the complexity of the model (i.e. number of parameters specified in each model). This is in part expected, since a model with more parameters will generally have a lower acceptance rate due to the higher number of dimensions in which the MCMC has to make proposal moves, leading to higher autocorrelations for each parameter. Nonetheless, the autocorrelation times are fairly robust for each model.

Thinning is an approach to use every $k$-th iteration of the MCMC walkers, where $k = \tau_f$ would represent an i.i.d. sampling of the posterior distribution. However, various papers indicate that thinning is often unnecessary and results in reduced precision (Harms et al., 2018; Link et al., 2012). Therefore, no thinning of the MCMC walkers was used in this analysis.

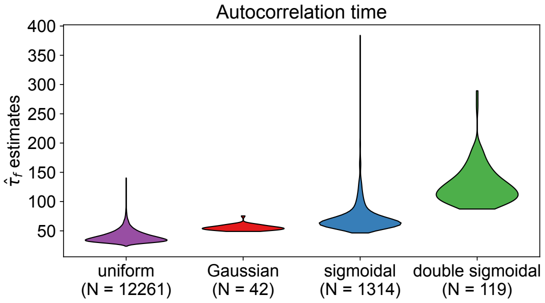

**Figure 6.10: $\hat{\tau}_f$ estimates for all genes in cortical cells derived from day 30 brain organoids.** The violin plots display the distribution of autocorrelation time estimates for all genes, grouped by the best-fitting model. The number of genes in each category is displayed in the x-axis.

Another way to visualize the posterior distribution over the parameter space derived from an MCMC is a corner plot, as shown in Figure 6.11. The corner plot highlights both the two dimensional projections over the parameter space across iterations of the MCMC, as well as the marginal posterior distribution for each individual parameter (highlighted in the upper plots). Some parameters are more correlated with each other than others, indicating underlying covariates within the model parameters. However, the marginal posterior distributions do not appear to be multimodal.

These heuristics provide some insight into the ability of the ensemble MCMC sampler to provide an accurate sampling of the posterior distribution over the parameter space.

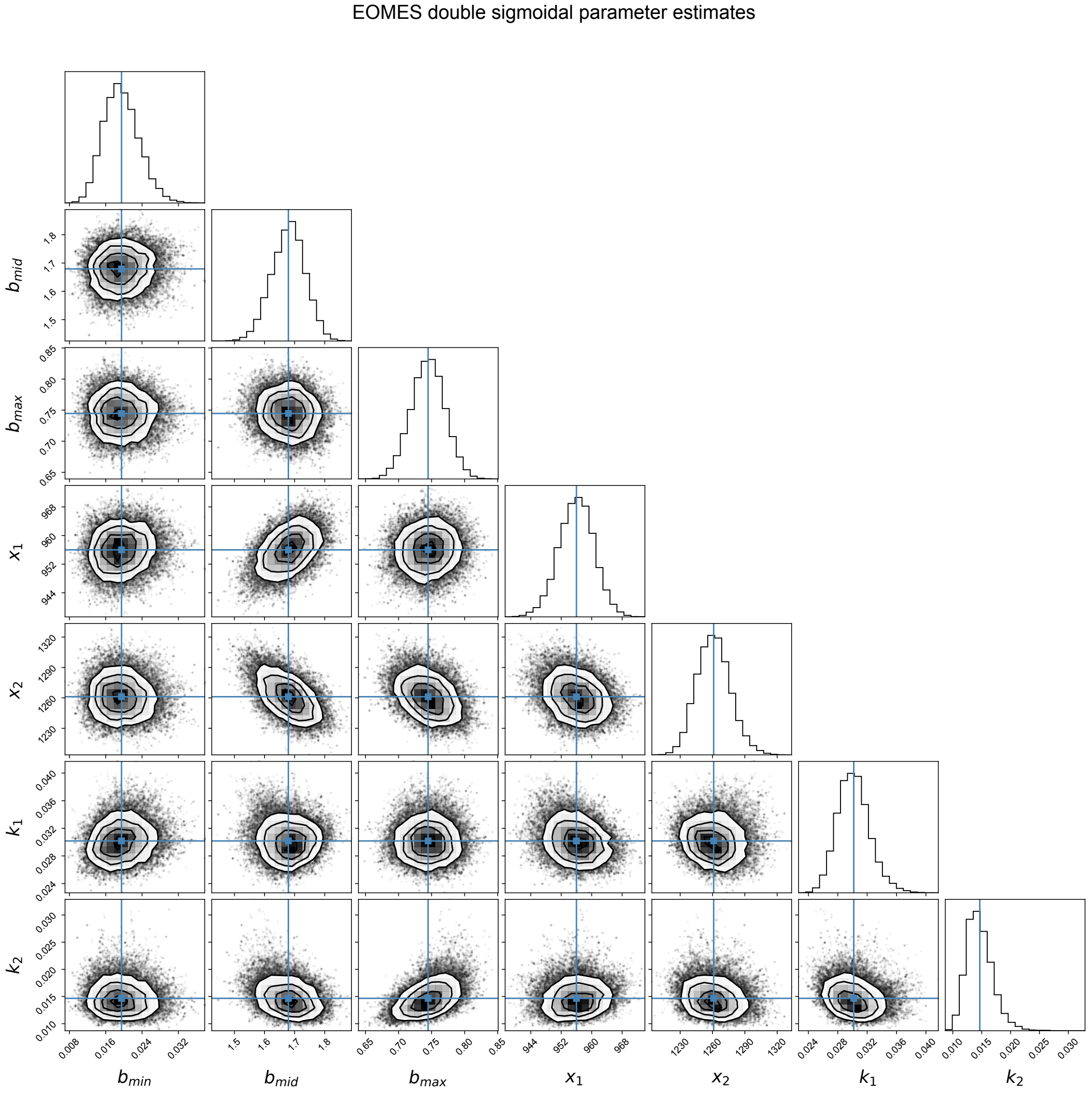

**Figure 6.11: Corner plot of MCMC double sigmoidal run for EOMES in cortical cells derived from day 30 brain organoids.** The corner plot displays the two-dimensional projections (below the diagonal) and marginal distribution for individual parameters (on the diagonal) of the posterior distribution across all MCMC walkers after removing the first 5,000 iterations (burn-in), and removing the samplers in the bottom half of acceptance rates. The mean estimate for each parameter is highlighted in the horizontal and vertical lines.

## 6.1.7 Estimating inflection points

Inflection points occur where the curvature of a function changes sign. At inflection points, the first-order derivative, or rate of change, of a function reaches a local maximum or local minimum. At an inflection point, the second-derivative of a function passes through 0 with the second derivative changing sign from positive (concave upward) to negative (concave downward) or vice versa. The inflection points of the Gaussian, sigmoidal and double sigmoidal fits can be used to compare the relative timing of when genes exhibit a state transition along a pseudotime trajectory. To estimate the inflection points of the different functions, first solve for $x$ at which the second-derivative of the function is zero. For the Gaussian function, $f(x)$, sigmoidal function $g(x)$, and double sigmoidal function $h(x)$ defined in Equation 6.1, the second derivatives are

$$
\begin{aligned}
f''(x) &= \frac{a}{\sigma^4} e^{-\frac{(x-x_0)^2}{2\sigma^2}} \left(x - (x_0 - \sigma)\right)\left(x - (x_0 + \sigma)\right), \\
g''(x) &= k^2 L \frac{e^{-k(x-x_0)} \left(e^{-k(x-x_0)} - 1\right)}{\left(1 + e^{-k(x-x_0)}\right)^3}, \\
h''(x) &= k_1^2 \left(b_{mid} - b_{min}\right) \frac{e^{-k_1(x-x_1)} \left(e^{-k_1(x-x_1)} - 1\right)}{\left(1 + e^{-k_1(x-x_1)}\right)^3} + \\
&\qquad k_2^2 \left(b_{max} - b_{mid}\right) \frac{e^{-k_2(x-x_2)} \left(e^{-k_2(x-x_2)} - 1\right)}{\left(1 + e^{-k_2(x-x_2)}\right)^3}.
\end{aligned}
$$

For the Gaussian function, two inflection points occur at $x \in (x_0 - \sigma, x_0 + \sigma)$. For the sigmoidal function, $g(x)$, one inflection point occurs at $x = x_0$. The estimates for the inflection points are then measured from the parameters $(x_0 - \sigma, x_0 + \sigma)$ for the case of the Gaussian and $x_0$ for the case of sigmoidal function at each MCMC iteration. Finally, for the double sigmoidal function, $h(x)$, the number of inflection points can vary. However, if all parameters are fixed besides $k_1$, then, $h''(x_1) \to 0$ as $k_1$ increases. Similarly, if all parameters are fixed besides $x_1$, then $h''(x_1) \to 0$ as $x_1$ decreases. That is, for $k_1 >> 0$, i.e. the transition from $b_{min}$ to $b_{mid}$ occurs rapidly, then an inflection point will occur very close to $x_1$. Similarly, for $k_2 >> 0$, i.e. the transition from $b_{mid}$ to $b_{max}$ occurs rapidly, then an inflection point will occur very close to $x_2$. Also, the further apart $x_1$ and $x_2$ are from each other, the closer the inflection points are to $x_1$ and $x_2$. To ensure the inflection points occur very close to $x_1$ and $x_2$, at each iteration of the MCMC, a move is only accepted in cases where $\text{sign}(h''(x_1 - dx)) \cdot \text{sign}(h''(x_1 + dx)) < 0$ and $\text{sign}(h''(x_2 - dx)) \cdot \text{sign}(h''(x_2 + dx)) < 0$ for $dx = 1$. The estimates for the inflection points are then measured from the parameters $x_1$ and $x_2$ at each MCMC iteration.

## 6.2 MEASURING TRANSCRIPTIONAL CASCADES AND REGULATORY INTERACTIONS

As described at the beginning of this chapter, gene regulatory interactions underlie cellular differentiation processes. Transcription factors play a key role in these processes, due to their ability to induce the expression of other genes, including other transcription factors, which are essential for the formation of specified cell types. This sequential induction or repression of transcription factors creates a cascade of gene expression, and enables a cell to turn genes on and off in a precisely timed manner as it differentiates. Thus, accurately measuring these differentiation cascades is pivotal in order to understand the endogenous factors responsible for cellular differentiation. In this section, we explore how to derive a transcriptional cascade after modeling gene expression dynamics along a developmental trajectory, as described in the previous section, with a particular emphasis on measuring the transcriptional cascades underlying neuronal differentiation in cortical cells of brain organoids.

### 6.2.1 Transcriptional cascades in cortical cells of brain organoids

Based on the best fits to either a double sigmoidal, Gaussian, sigmoidal or uniform function, genes can be sorted according to the relative pseudotemporal occurrence of inflection points. Genes for which a uniform function had the best fit are excluded due to the lack of an inflection point, and therefore a lack of underlying dynamics in gene expression. Sorting genes according to their first inflection point occurrence produces a reconstruction of the cascade of gene activation and repression.

In differentiating cells along the cortical NSC → IP → neuron trajectory in day 30 brain organoids, 112 out of 1,185 (9%) transcription factors that were expressed in at least 1% of cells had a non-uniform fit. These genes are highlighted in Figure 6.12. Initially, ASCL1, a gene that is central to the differentiation of neuroblasts into neurons (Sanes et al., 2012), is up-regulated, and the HES target genes of Notch signaling, HES1 and HES4, are repressed. These genes are known to play a critical role in the maintenance of NSCs, with their respective inactivation leading to the acceleration of neurogenesis (Kageyama et al., 2008). This repression is accompanied by an up-regulation of HES6, a known repressor of HES1 (Bae et al., 2000), as well as SOX4 and NEUROG2, which are required for IP cell specification and maintenance via activation of EOMES (Chen et al., 2015). Downstream of these initial changes, EOMES is up-regulated, constituting the neuronal lineage commitment of radial glia to IP cells. Following EOMES up-regulation, the transcription factors BCL11B and TBR1, markers of deep-layer cortical neurons generated during early cortical neurogenesis (Molyneaux et al., 2007), are up-regulated. These results demonstrate that the relative ordering of inflection point estimates for dynamically expressed transcription factors along the cortical NSC → IP → neuron trajectory in day 30 brain organoids accurately recapitulates known temporal orderings and regulatory interactions that are essential for the differentiation of cortical neurons in vivo, and these relationships are recapitulated in brain organoids.

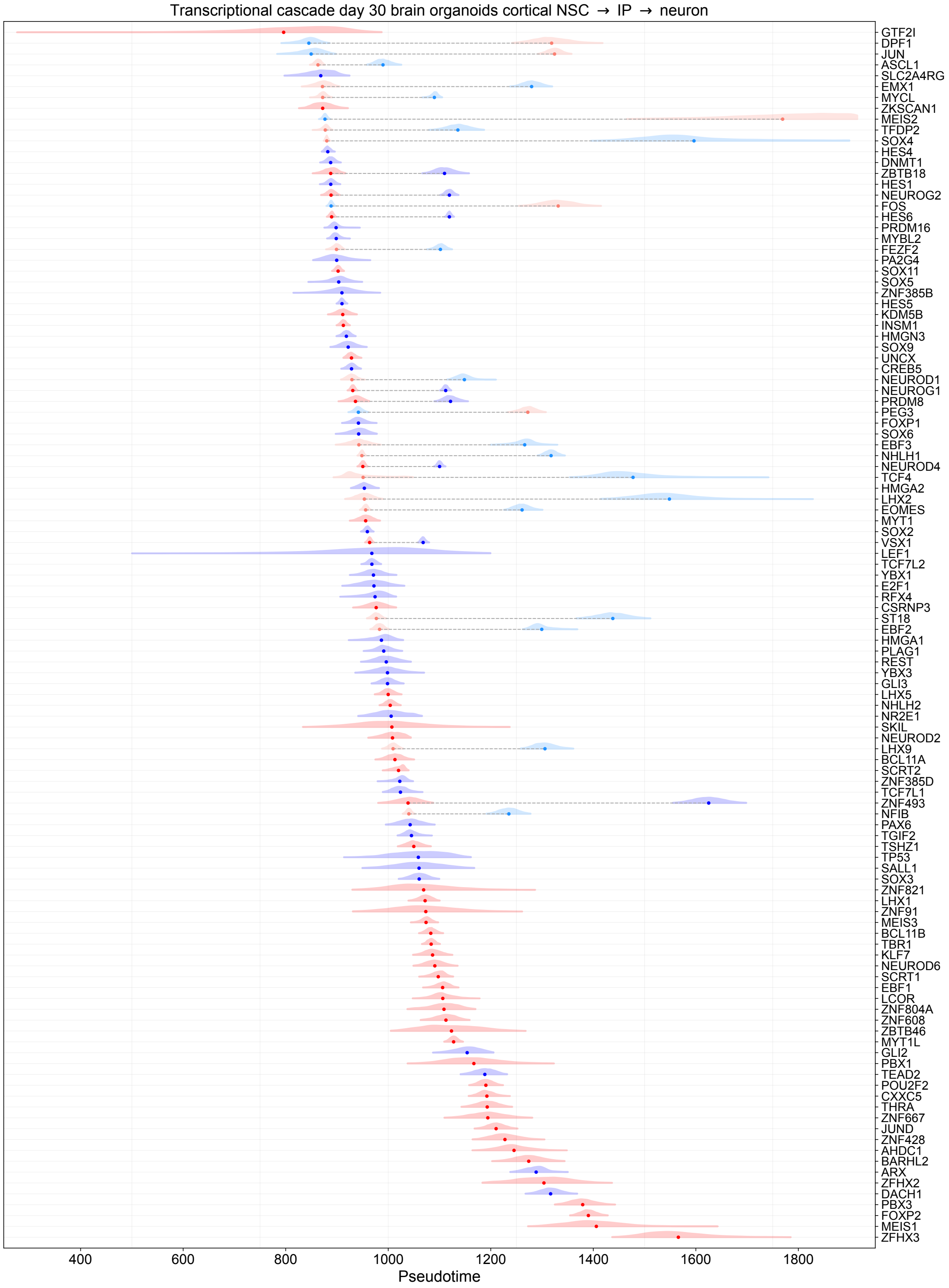

**Figure 6.12: Differentiation cascade for cortical NSC → IP → neuron trajectory in day 30 brain organoids.** Genes are ordered according to the mean of the relative pseudotemporal occurrence of the first inflection point, and the distributions of inflection point estimates across all MCMC iterations are shown. Positive inflection points are shown in red and negative inflection points are shown in blue. Genes with a sigmoidal fit have one inflection point while genes with a double sigmoidal or Gaussian fit have two, and the inflection points are connected with a dashed line. Inflection point estimates from double sigmoidal fits are shown in light blue and light red, and those from Gaussian fits in blue and red.

### 6.2.2 Constructing regulatory interactions in cortical cells of brain organoids

Using this sorting, potential upstream regulators of specific genes can be explored, as well as genes important for the maturation and differentiation of a given cell type. For example, the factors that repress EOMES in late-stage IP cells, thereby enabling the differentiation of these cell types into neurons, is poorly understood (Hevner, 2019). The transcription factor JUN is known to repress the expression of EOMES in differentiating T helper Th17 cells (Ichiyama et al., 2011), however, its role in differentiating cortical IP cells has not been characterized. The infection point estimates for these genes are highlighted in Figure 6.13. From the overlap of the inflection point estimates in these genes, it is possible to calculate a p-value to measure whether the inflection points occur simultaneously or not. The overlap is estimated by binning the inflection point estimates to 100 equally spaced bins starting at the minimum inflection point estimate across both genes and ending at the maximum inflection point estimate across both genes. The overlap of the first two inflection point estimates between EOMES and JUN is < 0.00143% (p-value < 1.43e-5), signifying that JUN is down-regulated prior to EOMES up-regulation. The overlap of the second two inflection point estimates is 3.7% (p-value = 0.037), signifying that EOMES is down-regulated nearly simultaneously as JUN is up-regulated. This strongly suggests a potential role for JUN to down-regulate EOMES during IP maturation into a fully differentiated neuron. A similar interaction between EOMES and FOS is observed, highlighted in Figure A3. FOS and JUN proteins are known to function as dimeric transcription factors that bind to AP-1 regulatory elements, and play a critical role in a variety of cellular processes including cell proliferation and differentiation (Chinenov et al., 2001). The results presented here suggest they may play a role in the differentiation of cortical neurons specifically by down-regulating EOMES in IPs.

By comparison, the Pearson correlation coefficient between the expression levels of EOMES and JUN is -0.056, reflecting a very low-level of anti-correlation. By comparing the inflection points of these two genes, a mutual repressive regulatory interaction can be observed, along with the relative timing of state change dynamics in both genes. This example highlights the unique insights this method offers to measure the regulatory interactions between genes along differentiation trajectories.

Furthermore, potential upstream regulators can be measured by extracting the genes with a positive inflection point occurring before, or simultaneously with, the positive inflection point of a given gene, as well as genes with a negative inflection point occurring after the positive inflection point of a given gene. To test whether inflection points occurred before or simultaneously with EOMES, a p-value was estimated as follows. As before, a histogram was generated for each gene by binning the inflection point estimates to 100 equally spaced bins starting at the minimum inflection point estimate (for the first inflection point in the case of Gaussian or double sigmoidal fit) and ending at the maximum inflection point estimate across both genes. Let $\{x_i\}_{i\in[1,100]}$ represent this binning domain. If $p_A(x_i)$ is the percent of counts in the histogram in bin $x_i$ for gene A, and $p_B(x_i)$ is the percent of counts in the histogram in bin $x_i$ for gene B, and $F_A(x_i)$ and $F_B(x_i)$ the cumulative distributions for gene $A$ and $B$ over the binned space respectively, then

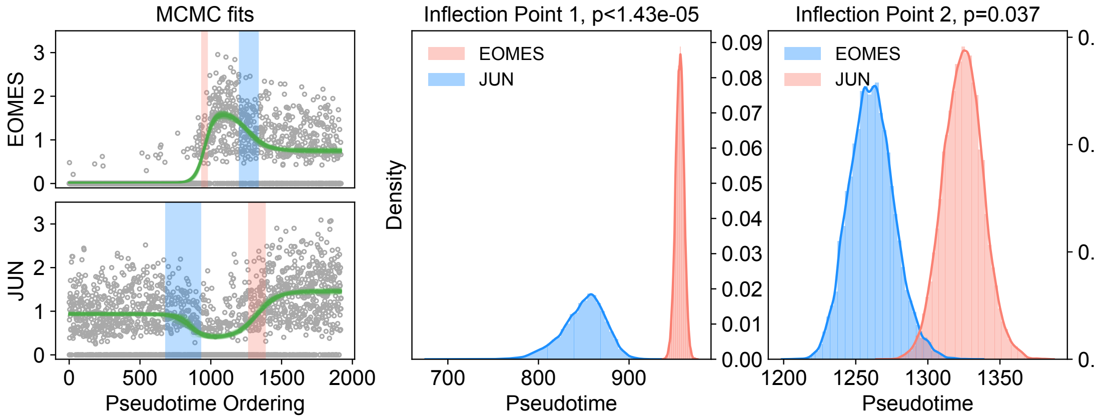

**Figure 6.13: Inflection point comparison of JUN and EOMES in day 30 cortical cells.** The left panel in the plot displays a random sampling of the parameters from 100 iterations of the MCMC traces for the double sigmoidal model, the best-fitting model for both genes. The full range of first and second inflection point estimates for both genes is highlighted as a shaded region, with blue indicating a negative inflection point and red a positive inflection point. The middle panel highlights the distribution of first inflection point estimates across MCMC iterations. The overlap in the distributions is <0.00143%, indicating a non-simultaneous inflection point. The right panel highlights the distribution of second inflection point estimates across MCMC iterations. The overlap in the distributions is 3.7%, indicating the two inflection points occur nearly simultaneously.

$$P(A \leq B) = \max_{i \in [1,100]} (F_A(x_i) - F_B(x_i)) + \sum_{i=1}^{100} \min(p_A(x_i), p_B(x_i)). \tag{6.13}$$

In Equation 6.13, the first term corresponds to $P(A < B)$ and the second term corresponds to the overlap in the two distributions (i.e. $P(A = B)$). Then, to test if the inflection point of gene A occurs before gene B, a p-value was estimated as $P(A \leq B)$. This procedure was performed for all transcription factors compared against EOMES, followed by a Benjamini–Hochberg multiple hypothesis correction (Benjamini et al., 1995). Genes with an adjusted p-value < 0.01 were labeled as positive regulators if they had a negative inflection point, and genes with an adjusted p-value ≥ 0.01 were labeled as positive regulators if they had a positive inflection point. This provided a list of 45 potential upstream positive regulators of EOMES, highlighted in Figure 6.14. Amongst the genes with a positive inflection point occurring before or simultaneously with EOMES is NEUROG2, which directly activates EOMES in the developing mouse neocortex (Kovach et al., 2013), as well as INSM1, which is also known to induce expression of EOMES in the developing mouse neocortex (Farkas et al., 2008). PAX6, an essential activator of EOMES gene expression (Quinn et al., 2007), was down-regulated after EOMES up-regulation, signifying that the expression of PAX6 enabled EOMES expression. Interestingly, directly after EOMES and NEUROG2 are up-regulated, PAX6 is down-regulated, suggesting a negative feedback loop, whereby PAX6 activates both EOMES and NEUROG2, which then in turn repress PAX6, a relationship which has been previously described in the developing mouse cortex (Kovach et al., 2013). These results further validate the utility of this method in discovering upstream regulators of a given gene of interest. The remaining potential activators of EOMES listed here warrant further experimental validation.

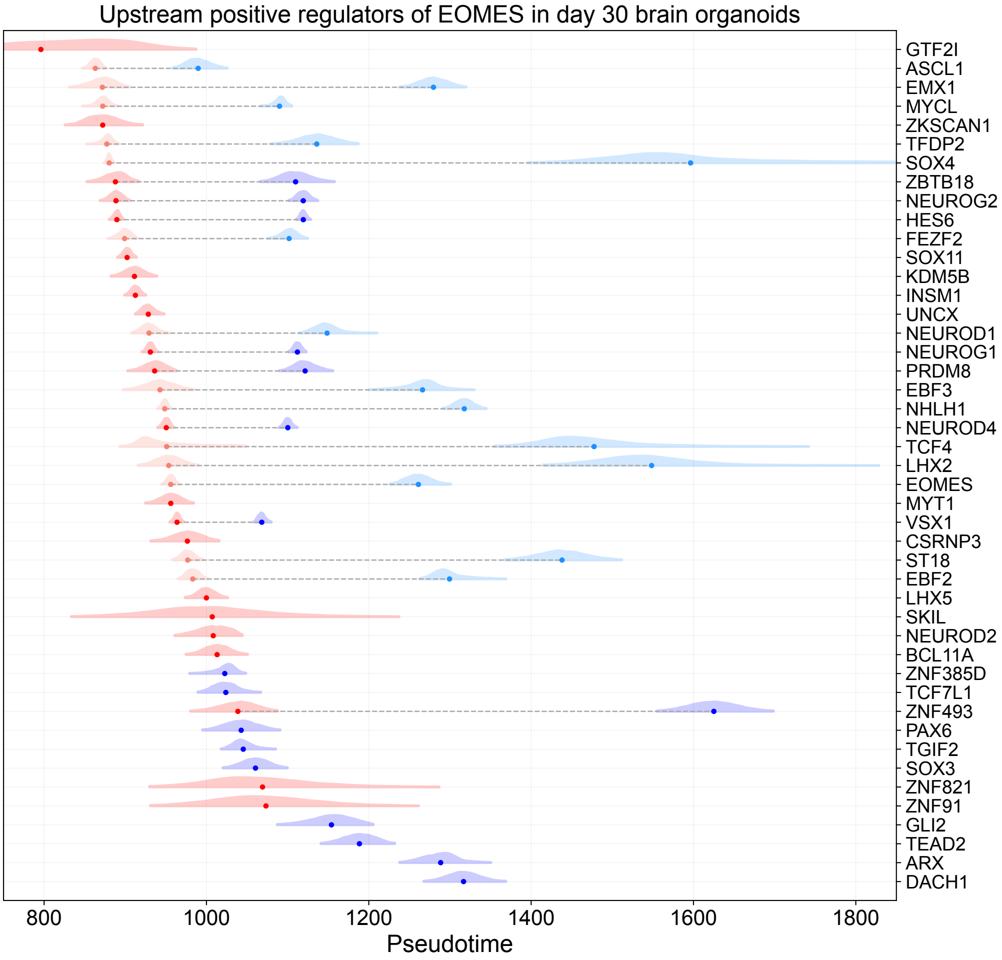

**Figure 6.14: Differentiation cascade of potential upstream regulators of EOMES.** A transcriptional cascade of the potential upregulators of EOMES is shown. Genes with a positive inflection point occurring before, or simultaneously with, the positive inflection point of EOMES, as well as genes with a negative inflection point occurring after EOMES are shown. Inflection point estimates from double sigmoidal fits are shown in light blue and light red, and those from Gaussian and sigmoidal fits in blue and red.

### 6.2.3 Comparing regulatory interactions in day 30 and 50 brain organoids

In order to increase confidence in the regulatory interactions derived in differentiating cortical NSCs in day 30 brain organoids, the same analysis was performed on day 50 brain organoids. Figure 6.15 highlights the pseudotime measurements within cortical cells derived using diffusion pseudotime (Haghverdi et al., 2016) in day 50 organoids derived under Triple-i and sequenced using the 10X protocol.

From the pseudotemporal ordering of cortical cells along a cortical NSC → IP → neuron differentiation trajectory, genes were fit to a double sigmoidal, Gaussian, sigmoidal and uniform function using the procedure described in Section 6.1. A transcriptional cascade of the shared

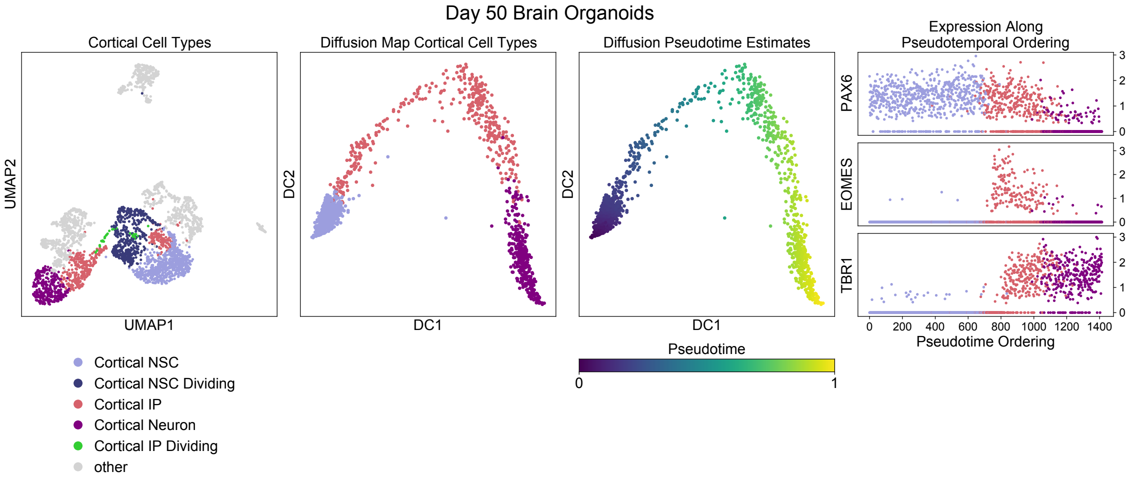

**Figure 6.15: Ordering day 50 organoid cells along a cortical neuronal differentiation trajectory.** The left-most panel displays a UMAP of cells in a scRNA-Seq experiment of day 50 cerebral organoids derived using the Triple-i protocol. Cortical cell types are highlighted in the plot, including dividing and non-dividing cortical NSCs, IPs, and neurons. All other cells are highlighted as "other". The second panel highlights a diffusion map embedding of the cortical NSC, IP and neuronal populations, after removal of all dividing cell types, with the third panel highlighting the same cells in the diffusion map colored by diffusion pseudotime estimates, scaled from 0 to 1. The rightmost panel displays the expression of key developmental transcription factors for cortical neurogenesis, PAX6, EOMES and TBR1, within cells after ordering them according to their relative pseudotemporal ordering using a fixed time step of 1. Cells are colored by their cell type annotations.

potential upstream regulators of EOMES is highlighted in Figure 6.16. EOMES has a similar expression pattern in both day 30 and day 50 brain organoids, with the double sigmoidal model having the best fit in both cases. However, in day 50 organoids, EOMES expression becomes fully suppressed in cortical neurons as seen in Figure 6.15, while in day 30 cortical neurons, a low level of EOMES expression is still detected in cortical neurons, potentially signifying a less mature neuronal stage at day 30. Among the 45 transcription factors found to be potential upstream regulators of EOMES in cortical cells of day 30 brain organoids, 26 (58%) were shared between both datasets. Notably, the experimentally validated activators of EOMES in the developing mouse cortex, PAX6, NEUROG2 and INSM1, are present in both datasets.

Furthermore, it is possible to further increase the confidence of the role of JUN as a transcriptional repressor of EOMES in cortical IP cells. In differentiating cortical cells in day 50 organoids, JUN has a double sigmoidal fit, with the first and second inflection points having a 22.3% and 11.5% overlap with the first and second inflection points of EOMES, as shown in Figure 6.17. In day 30 organoids, JUN was repressed prior to EOMES up-regulation (Figure 6.13), however, in both datasets, the second inflection points overlap significantly. This indicates that in maturing cortical IP cells of days 30 and 50 of organoid development, the reactivation of JUN co-occurs with the repression of EOMES, highlighting its potential role as a transcriptional repressor of EOMES.

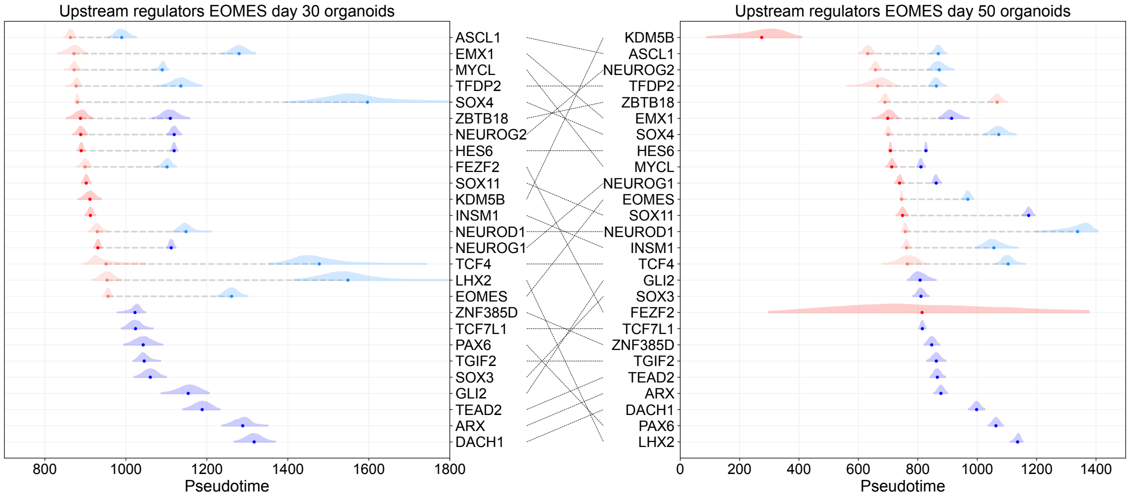

**Figure 6.16: Differentiation cascade of shared potential upstream regulators of EOMES across day 30 and day 50 brain organoids.** The left and right plots show a transcriptional cascade of the shared potential upregulators of EOMES in day 30 and day 50 brain organoids. Genes with a positive inflection point occurring before, or simultaneously with, the positive inflection point of EOMES, as well as genes with a negative inflection point occurring after the positive inflection point of EOMES, in both datasets, are shown. Inflection point estimates from double sigmoidal fits are shown in light blue and light red, and those from Gaussian and sigmoidal fits in blue and red. The dashed lines connect the genes across both datasets.

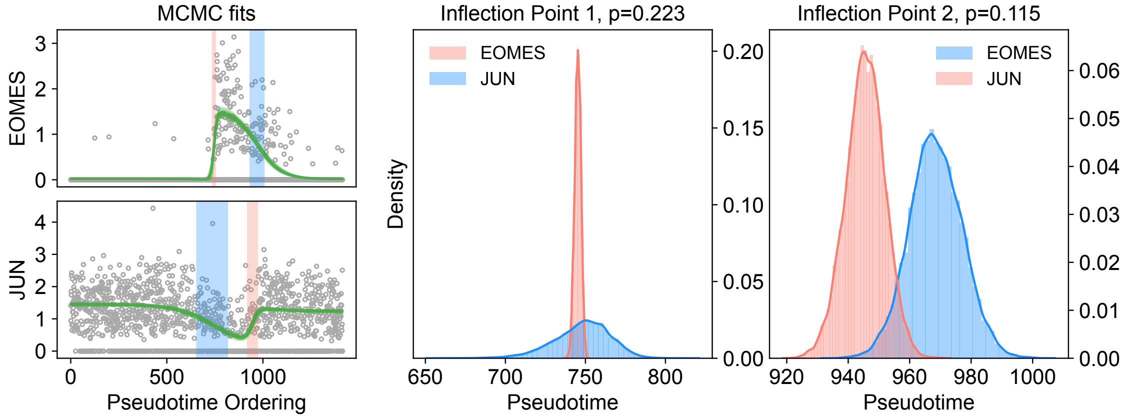

**Figure 6.17: Inflection point comparison of JUN and EOMES in day 50 cortical cells.** The left panel in the plot displays a random sampling of the parameters from 100 iterations of the MCMC traces for the double sigmoidal model, the best-fitting model for both genes. The full range of first and second inflection point estimates for both genes is highlighted as a shaded region, with blue indicating a negative inflection point and red a positive inflection point. The middle and right panels highlight the distribution of first and second inflection point estimates across MCMC iterations, respectively, with overlaps of 22.3% and 11.5%, indicating simultaneous inflection points.

### 6.2.4 Dissecting Notch signaling in day 30 and 50 brain organoids

Different signaling pathways are also known to play an essential role in cortical development. For example, Notch activation in radial glia instructs these progenitors to remain as neural stem cells (Nye et al., 1994; la et al., 1997). IPs interact with radial glia via the Delta-Notch signaling pathway, whereby Notch receptors are activated on the surface of radial glia by binding to ligands, such as DLL1 and DLL3, present on the surface of IPs (Hevner, 2019). The binding of DLL1 to NOTCH1 is further enhanced by the glycosyltransferase MFNG, which modifies the extracellular domain of NOTCH1, resulting in an increase ability to bind to DLL1. Hence, IPs are essential for maintaining the balance between proliferation and differentiation of radial glia in the developing cortex. Furthermore, while DLL1 is selectively expressed in IPs located more apically in the developing cortex, or closer to the ventricular zone harboring radial glia, DLL3 is selectively expressed in more basal IPs, those which have already migrated away from the ventricular zone (Hevner, 2019).

To measure these dynamics along the cortical NSC → IP → neuron trajectory in day 30 and day 50 brains organoids, shared dynamically expressed genes involving ligand-receptor pairs of Notch receptors (Shao et al., 2021) in both samples were estimated, and are highlighted in Figure 6.18. In both trajectories, DLL1 is up-regulated in early IP cells, followed by the up-regulation DLL3 in later stage IPs, confirming the selective basal expression of DLL3 from in vivo studies. This is accompanied by MFNG up-regulation and NOTCH3 repression. Interestingly, NOTCH3 is down-regulated in early-stage IPs followed by NOTCH1 down-regulation in late-stage IPs, indicating that the time-dependent down-regulation of different Notch receptor genes may play an important role in IP maturation.

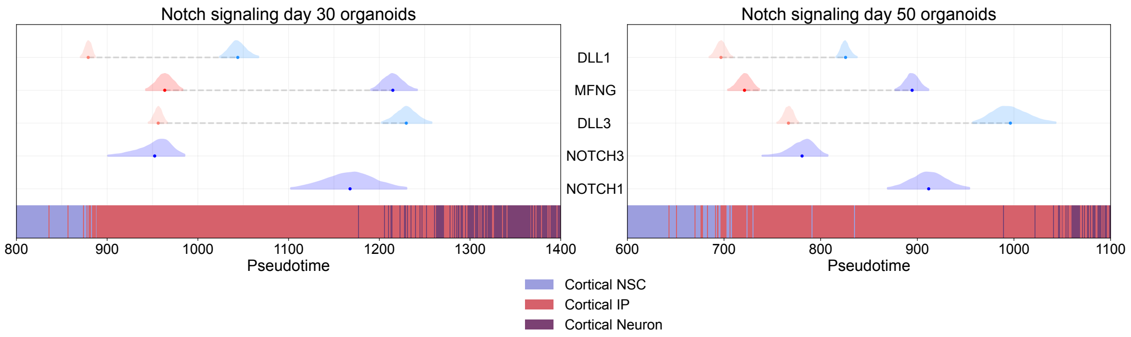

**Figure 6.18: Notch signaling cascade in day 30 and 50 cortical cells.** The left and right plots show a transcriptional cascade of the shared ligand-receptor pairs involved in Notch signaling in day 30 and day 50 brain organoids. The annotated cell type for each cell in the trajectory is highlighted in the bottom of each plot.

## 6.3 SUMMARY

In this chapter, we explored an approach to model the transcriptional dynamics of differentiating cell types along a pseudotime trajectory. Pseudotemporally ordered gene expression profiles were fit to double sigmoidal, sigmoidal, Gaussian and uniform functions using an affine-invariant MCMC approach to explore the posterior distributions over the parameter space of the different models. The best-fitting model was chosen by comparing the BIC estimated on random subsets of the input data for each gene, using the uniform fits as a background model, reflecting the absence of gene expression dynamics. The utility of using these specific functions to model gene expression dynamics is in part due to the ease of interpretation of the model parameters, in particular the ability to measure inflection points, which measure when a gene exhibits a local maximal or minimal change in gene expression. The inflection point estimates can then be used to compare the relative timing of when genes exhibit a state transition along a pseudotime trajectory.

We applied this method to differentiating cortical neural stem cells into neurons via IPs in day 30 and day 50 brain organoids. By ordering transcription factors along a pseudotime trajectory by the relative occurrence of inflection points, we were able to measure the transcriptional cascades underlying neuronal differentiation within the developing cortex. We then identified potential upstream positive regulators of EOMES, an essential gene for the formation of IPs, in both datasets. This analysis revealed a set of high confidence regulators, some of which have been experimentally validated, and others have yet to be fully explored in the context of EOMES regulation during cortical development. We also identify a potential negative regulator of EOMES, the transcription factor JUN, which is known to directly repress the expression of EOMES during T cell differentiation, but whose role in cortical IP maturation has not been fully explored. We also compared the expression dynamics of genes involved in the Notch signaling pathway across day 30 and day 50 brain organoids. This revealed a sequential up-regulation of the Notch receptor ligands DLL1 and DLL3 in progressing IPs, reflecting the in-vivo selective expression of these genes in apical and basal IPs respectively. This analysis further revealed the sequential down-regulation of NOTCH3 followed by NOTCH1, highlighting a potentially selective role for different Notch receptors in IP maturation.

# 7 DISCUSSION AND CONCLUSION

The human cortex is arguably the most complex structure in the known universe. Understanding the transcriptional programs underlying cell-fate determination and differentiation in the developing cortex are of great interest. Furthermore, access to in vivo samples of the developing human brain is extremely limited. While animal models can be very useful in studying brain development, they are limited in their ability to model the complexities of the human brain.

Brain organoids offer the ability to probe cortical development using an in vitro system at any time point of interest. These in vitro systems offer a means to model brain development under both healthy and disease conditions, enabling scientists to understand the genetic processes underlying the formation of the brain in a controlled environment. This further enables genetic perturbation screening and clinical applications, such as testing the therapeutic potential of small molecules, among other applications. However, in order to use brain organoids to model healthy development and disorders of cortical development, robust methods for deriving highly specified cortical neural stem cell populations and their progeny across many iPSC and ESC lines must be established.

In Chapter 4, we showed that a brain organoid derivation protocol consisting of a combination of TGF-$\beta$, BMP and WNT inhibition (referred to as Triple-i throughout this work) specifically enriches for cortical cell types and gives rise to robust cortical populations across four iPSC lines and one ESC line. We highlighted the insights bulk RNA-Seq and scRNA-Seq technologies offer to measure these enrichments and dissect the regional and cell-type heterogeneity arising in brain organoid systems. From bulk RNA-Seq of individual day 30 organoids, we highlighted a significant enrichment of cortical genes in Triple-i organoids when compared against Dual SMAD-i and Inhibitor-free organoids derived from the H9 ESC line. Furthermore, using scRNA-Seq of day 50 organoids derived from four different iPSC lines, we showed that the Triple-i protocol is the only method which gives rise to substantial and consistent cortical populations across all cell lines. This work establishes the Triple-i protocol as a robust method to model cortical development.

scRNA-Seq technologies enable an unprecedented view into the identity of individual cells within a tissue, as defined by their transcriptomic output. However, in order to draw conclusions from the data generated by this technology, potential artifacts in the data must be properly addressed. In Chapter 5, we describe a method we developed to identify and remove empty droplets. These empty droplets contain ambient mRNA, or free-floating mRNA from the sample, and can have a strong negative impact in downstream analyses, leading to the identification of artifcatual cell populations, as well as the contamination of ambient mRNA signal in real cell populations. We use the method on inDrops and 10X scRNA-Seq datasets of day 50 brain organoids to classify droplets as cell-containing or empty, and show that removing the empty droplets identified by the method produces more accurate results, and represents an important pre-processing step in the analysis of any scRNA-Seq dataset. Furthermore, we show that scRNA-

Seq datasets containing cells from multiple individuals can be demultiplexed, in which UMIs are genotyped to each individual based on the presence of SNPs in the sequencing reads corresponding to a UMI. This information can be aggregated across all UMIs in each cell barcode to assign a given cell to an individual, or identify droplets which contain a mixture of genetic material reflecting the mixture seen in the ambient mRNA, as shown in Section 5.3.3. This provides an orthogonal piece of information to identify empty droplets, validating the results using the method we developed to identify empty droplets based on the raw count data alone.

In Chapter 6, we describe a downstream application in scRNA-Seq analysis that involves measuring transcriptional dynamics along developmental trajectories. Specifically, we investigated the dynamics of gene expression in differentiating cells along the cortical NSC → IP → neuron trajectory within brain organoids derived by the Triple-i method at days 30 and 50 of development. Based on the pseudotemporal ordering of cells along this trajectory, expression profiles of individual genes are fit to a variety of curves that reflect biological state switches using an affine-invariant MCMC inference approach, presented in Section 6.1.4. By ordering the genes based on the relative pseudotemporal occurrence of inflection points, genes can be ordered according to a transcriptional cascade, as shown in Section 6.2. The transcriptional cascades measured along these trajectories match known temporal orderings from in vivo studies.

We further investigated how to use these cascades to determine potential positive upstream regulators of genes of interest, focusing on the transcription factor EOMES, which is essential for the formation of intermediate progenitors in the developing cortex. Not only do we recover validated activators of EOMES, such as PAX6 and NEUROG2, we also detect a number of other transcription factors whose roles in EOMES activation and repression have not been fully characterized. Further studies are needed in order to validate the roles of these transcription factors with respect to the regulation of EOMES expression.

Here, we formulate a variety of experiments which can be used to validate the regulatory roles of these genes. Chromatin Immunprecipitation followed by high-throughput sequencing (ChIP-Seq) is an experimental approach for genome-wide profiling of DNA-binding proteins, such as transcription factors or histone marks (Johnson et al., 2007; Robertson et al., 2007). By conducting ChIP-Seq experiments, it is possible to measure to what extent an individual transcription factor binds to the promoter or enhancer regions of EOMES, thereby validating its regulatory role in EOMES expression. It would be essential to isolate the EOMES+ cells in brain organoids before performing a ChIP-Seq experiment in order to measure this regulation specifically in cortical IP cells. EOMES+ cells can be isolated with a reporter cell line, in which a reporter gene is inserted next to the gene promoter of EOMES with a reporter gene such as luciferase or fluorescent proteins like GFP and RFP, which can be isolated using Fluorescence-activated Cell Sorting (FACS). Promoter and enhancers for EOMES can be detected using ChIP-Seq for histone marks H3K4me1, H3K4me3 and H3K27ac, in combination with Hi-C data to measure the TAD boundaries containing the genomic location of *EOMES*.

Other approaches to determine the regulatory roles of the transcription factors highlighted in Section 6.2.3 with respect to EOMES expression could involve gene knockout (whereby gene expression is disrupted by deleting part of the DNA sequence or inserting irrelevant DNA sequence within the gene body), knockdown (whereby gene expression is reduced via the introduction of an oligonucleotide which binds to the mRNA of a gene of interest, leading to mRNA degradation or preventing translation), or cell transfection, such as virus-mediated or liposome-mediated overexpression, experiments. If the expression of EOMES is reduced in a knockdown or knockout experiment of an upstream positive regulator, this provides evidence of the gene's positive regulatory role in EOMES expression. Similarly, if the overexpression of a positive regulator of EOMES leads to the up-regulation of EOMES, this provides experimental support of that gene's regulatory function in EOMES expression. For suppressors of EOMES, the opposite is true.

Furthermore, other datasets can be incorporated to increase the evidence for a given regulatory interaction. For example, the enrichment of known transcription factor motifs in the promoter and enhancer regions of a given target gene provides a further piece of evidence of the presence of regulatory function of the upstream activator or repressor. Tools such as SCENIC (Aibar et al., 2017) incorporate this information to measure regulatory networks from scRNA-Seq datasets, albeit without incorporating pseudotime dependent gene interactions.

Finally, the MCMC approach presented in Section 6.1 is computationally expensive, and determining the total number of iterations to run the MCMC to ensure it accurately samples from the posterior distribution over the parameter space is impossible. Therefore, it is worth investigating whether a similar performance can be achieved using other Bayesian inference approaches, such as variational inference. Also, in differentiating cell types where there are more than two state-changes present, other curves can be easily introduced into the model besides those in Equation 6.1, enabling the modeling of more complex transcriptional dynamics.

While we focused specifically on cells along a cortical NSC → IP → neuron trajectory in this thesis, the method presented in Section 6.1 can be applied to any scRNA-Seq dataset where cells are ordered along a differentiation trajectory. The method can predict novel regulatory interactions within differentiating cells, as well as measure transcriptional cascades to deduce critical genes for cell maturation and gene interactions involved in different signaling pathways. Therefore, we believe this approach can provide useful insight into the molecular underpinnings involved in many developmental biology contexts.

# ABBREVIATIONS

| | |
|---|---|
| **A** | adenine |
| **AIC** | Akaike information criterion |
| **aRG** | apical radial glia |
| **BIC** | Bayesian information criterion |
| **BMP** | bone morphogenetic proteins |
| **BMPR** | bone morphogenetic protein receptors |
| **bp** | base pair |
| **C** | cytosine |
| **cDNA** | complementary DNA |
| **CNS** | central nervous system |
| **CP** | cortical plate |
| **DIC** | deviance information criterion |
| **DNA** | deoxyribo-nucleic acid |
| **ESC** | embryonic stem cell |
| **ESS** | effective sample size |
| **FPKM** | fragments per kilobase of exon per million mapped fragments |
| **G** | guanine |
| **HVG** | highly variable genes |
| **ICM** | inner cell mass |
| **i.i.d.** | independent identically distributed |
| **IP** | intermediate progenitors |
| **iPSC** | induced PSCs |
| **kNN** | *k*-nearest neighbors |
| **LAD** | lamina-associating domains |
| **LOO-CV** | leave-one-out cross validation |
| **LpO-CV** | leave-p-out cross validation |
| **MAP** | maximum a-posterior |
| **MCMC** | Markov chain Monte Carlo |
| **mRNA** | Messenger RNA |
| **MLE** | maximum likelihood estimation |
| **NGS** | next-generation sequencing |

| | |
|---|---|
| **NSC** | neural stem cells |
| **oRG** | outer radial glia |
| **pre-mRNA** | precursor mRNA |
| **PCA** | Principal component analysis |
| **PCR** | polymerase chain reaction |
| **PNS** | peripheral nervous system |
| **PSC** | pluripotent stem cells |
| **RNA** | ribonucleic acid |
| **RNAP II** | RNA Polymerase II |
| **RNA-Seq** | RNA sequencing |
| **scRNA-Seq** | single-cell RNA-Sequencing |
| **SNP** | single nucleotide polymorphisms |
| **SVZ** | subventricular zone |
| **TFBS** | transcription factor binding sites |
| **TGF-$\beta$** | transforming growth factor beta |
| **T** | thymine |
| **U** | uracil |
| **UMI** | unique molecular identifier |
| **UMAP** | Uniform Manifold Approximation and Projection |
| **VZ** | ventricular zone |
| **TAD** | topologically associating domains |
| **5’ UTR** | 5’ untranslated region |
| **3’ UTR** | 3’ untranslated region |

# LIST OF FIGURES

Figure 2.1 **Central dogma of molecular biology.** 4

Figure 2.2 **Transcription in eukaryotic cells.** 6

Figure 2.3 **Waddington's epigenetic landscape.** 9

Figure 2.4 **Early stages of mammalian embryogenesis.** 10

Figure 2.5 **Neurulation in mammalian embryogenesis.** 11

Figure 2.6 **Formation of primary and secondary brain vesicles.** 12

Figure 2.7 **Neocortical expansion and lamination in mouse neocortex.** 14

Figure 2.8 **Overview of RNA-Seq.** 18

Figure 2.9 **Overview of scRNA-Seq.** 20

Figure 3.1 **Bayesian inference.** 26

Figure 3.2 **Rejection sampling.** 27

Figure 3.3 **Markov process.** 29

Figure 3.4 **Metropolis–Hastings examples.** 32

Figure 3.5 **Ensemble sampler stretch move.** 33

Figure 4.1 **PCA of bulk RNA-Seq day 30 brain organoids.** 40

Figure 4.2 **Differential gene expression and gene set enrichment analysis of bulk RNA-Seq of day 30 brain organoids.** 41

Figure 4.3 **scRNA-Seq UMAP embedding of day 50 organoids.** 42

Figure 4.4 **scRNA-Seq relative expression of marker genes across clusters in day 50 organoids.** 43

Figure 4.5 **Regional composition of day 50 organoids from scRNA-Seq.** 44

Figure 5.1 **Cell barcode rank plot from scRNA-Seq data generated using inDrops protocol.** 46

Figure 5.2 **Distribution of UMI counts per cell barcode.** 46

Figure 5.3 **UMI counts vs. mitochondrial content per cell barcode.** 47

Figure 5.4 **Artifactual modes in scRNA-Seq data.** 47

Figure 5.5 **UMAP embedding of real cells and simulated cells.** 50

Figure 5.6 **Co-clustering real and simulated cells with varying Louvain resolution parameter.** 51

Figure 5.7 $P_{si}$ **values in both real and simulated cells.** 52

Figure 5.8 $D_r$ **estimates with varying Louvain resolution parameter** $r$**.** 53

Figure 5.9 **Classification of cells as empty droplets.** 54

Figure 5.10 **UMI counts in empty droplets.** 54

Figure 5.11 $\hat{c}_i$ **estimates for corresponding** $N_{i1}$ **and** $N_{i2}$ **estimates in each cell barcode.** 57

Figure 5.12 $\hat{c}_i$ **estimates per cluster.** 57

Figure 5.13 **Percentage of empty droplets per cluster estimated by emptyDrops.** 58

Figure 5.14 **SoupX contamination estimates.** 59

Figure 5.15 **DecontX contamination estimates.** 60

Figure 5.16 **Clustering inDrops scRNA-Seq of day 50 organoid samples with and without empty droplets.** 62
Figure 5.17 **Distribution of UMI counts per cell barcode in 10x scRNA-Seq.** 63
Figure 5.18 **Including cells with low UMI counts in 10X ZIP8K8-derived day 50 Triple-i organoids.** 65
Figure 6.1 **Ordering cells along cortical neuronal differentiation trajectory.** 71
Figure 6.2 **Modeling gene expression dynamics.** 72
Figure 6.3 **Estimating $\phi$ in scRNA-Seq UMI count data.** 74
Figure 6.4 **MCMC trace plot for EOMES in cortical cells derived from day 30 brain organoids.** 78
Figure 6.5 **Acceptance rates for individual MCMC walkers for EOMES in cortical cells derived from day 30 brain organoids.** 78
Figure 6.6 **Distributions of log-likelihoods for MCMC fits for EOMES in cortical cells derived from day 30 brain organoids.** 79
Figure 6.7 **MCMC fits and BIC estimates for EOMES in cortical cells derived from day 30 brain organoids.** 81
Figure 6.8 $\rho_f(\tau)$ **estimates in MCMC double sigmoidal run for EOMES in cortical cells derived from day 30 brain organoids.** 83
Figure 6.9 $\hat{\tau}_f$ **estimates in MCMC double sigmoidal run for EOMES in cortical cells derived from day 30 brain organoids.** 83
Figure 6.10 $\hat{\tau}_f$ **estimates for all genes in cortical cells derived from day 30 brain organoids.** 84
Figure 6.11 **Corner plot of MCMC double sigmoidal run for EOMES in cortical cells derived from day 30 brain organoids.** 85
Figure 6.12 **Differentiation cascade for cortical NSC → IP → neuron trajectory in day 30 brain organoids.** 88
Figure 6.13 **Inflection point comparison of JUN and EOMES in day 30 cortical cells.** 90
Figure 6.14 **Differentiation cascade of potential upstream regulators of EOMES.** 91
Figure 6.15 **Ordering day 50 organoid cells along a cortical neuronal differentiation trajectory.** 92
Figure 6.16 **Differentiation cascade of shared potential upstream regulators of EOMES across day 30 and day 50 brain organoids.** 93
Figure 6.17 **Inflection point comparison of JUN and EOMES in day 50 cortical cells.** 93
Figure 6.18 **Notch signaling cascade in day 30 and 50 cortical cells.** 94
Figure A1 **Differential expression analysis comparing real cells with simulated background cells.** 109
Figure A2 **Empty droplets per cluster in scRNA-Seq of day 50 organoids.** 110
Figure A3 **Inflection point comparison of FOS and EOMES in day 30 cortical cells.** 111

# LIST OF TABLES

Table 5.1 **Empty droplets in 10X and inDrops scRNA-Seq of day 50 brain organoids.** . . . . . 66
Table 6.1 **Priors on Function Parameters.** . . . . . 76

# A APPENDIX

## A.1 EXPERIMENTAL PROCEDURES AND DATA PROCESSING

All organoids described in this thesis were grown in Yechiel Elkabetz's lab at the Max Planck Institute for Molecular Genetics, using the Triple-i, Dual SMAD-i, and Inhibitor-free protocols, as described in (Elkabetz et al., 2022). All raw and processed data of the bulk RNA-Seq datasets of individual day 30 organoids and 10X scRNA-Seq datasets of pooled day 50 organoids from the paper (Rosebrock et al., 2022) are deposited in the Gene Expression Omnibus under the accession code GSE189981.

**Bulk RNA-Seq of day 30 brain organoids.** The day 30 organoids from the paper (Rosebrock et al., 2022) and described in Chapter 4 were derived by Triple-i, Dual SMAD-i, and Inhibitor-free protocols from the ESC line H9 (N=8 for each protocol). RNA from individual organoids was purified using an miRNeasy RNA MiniPrep kit (Qiagen). RNA-Seq libraries were generated using Illumina TruSeq RNA library preparation kits and sequenced on an Illumina HiSeq 2500 sequencer as 100-bp paired-end reads. Reads were then trimmed using Trimmomatic (Bolger et al., 2014) (version 0.36; parameters: LEADING, 3; TRAILING, 3; SLIDINGWINDOW, 4:15; MINLEN, 36). The trimmed reads were then aligned to the human reference genome GRCh37 using STAR mapper version 2.6.1d (Dobin et al., 2013) and Gencode v19 gene annotations. Read counts and FPKM values were then estimated using RSEM version 1.3.1 (Li et al., 2011).

**10X scRNA-Seq of day 30 and day 50 brain organoids.** The day 50 organoids from the paper (Rosebrock et al., 2022) and described in Chapter 4 were derived by Triple-i, Dual SMAD-i, and Inhibitor-free protocols from ZIP8K8, ZIP13K5, KUCG2 and FOK1 iPSC cell lines (N=4-5 pooled organoids per sample). The day 30 and day 50 organoids described in Chapter 6 were derived by the Triple-i protocol from the ZIP8K8 iPSC line (N=3 pooled organoids per sample). The samples were then dissociated into single cells using a papin dissociation kit (Worthington). The organoids were dissected into small pieces, incubated with papin and DNaseq I solution for 35-45 minutes, triturated and the cell suspension was filtered twice through 40-$\mu$m filter to obtain a single-cell suspension. The cells were centrifuged at 300g for 5 minutes, resuspended in Dulbecco's phosphate buffer solution containing 0.4% BSA and counted for viability (>80%). Roughly 17,400 single live cells (1,000 cells $\mu l^{-1}$) in Dulbecco's phosphate buffer solution containing 0.4% BSA were used for Gel Beads-in-emulstion (GEM) generation, barcoding and library preparation according the manufacturer's recommendations for the 10X Chromium single cell 3' reagent kit v3.1. Nine cycles were used for cDNA amplification, whereas 12 cycles were performed for library construction, and the libraries were sequenced using an Illumina NovaSeq 6000 sequencer. The fastq data was then processed using the Cell Ranger software version 3.1.0 (Zheng et al.,

2017) using default parameters with human reference genome version GRCh38 and ensemble v92 reference transcriptome.

**inDrops scRNA-Seq of day 50 brain organoids.** The day 50 organoids described in Chapter 5 were derived by Triple-i, Dual SMAD-i, and Inhibitor-free protocols from cell lines ZIP8K8 and ZIP13K5 (N=2 for each iPSC line and each protocol). The samples were dissociated into single cells using papain dissociation kit (Worthington). Organoids were first incubated with papain and DNase I solution for 35 min and then triturated and filtered through 40-micron filter to obtain single cell suspension. Cells were centrifuged at 300g for 5 min, re-suspended in Hank's Balanced Salt Solution (HBSS), counted for viability (>80% were viable) and FACS sorted using FACS Aria III (Becton Dickinson), while excluding dead cells labeled with DAPI. Roughly 50,000 single organoid derived and sorted live cells were collected in PBS buffer (100 cells/$\mu$L) and subjected to an inDrop v2 procedure (Klein et al., 2015). Briefly, collected cells were injected in the commercially available inDrop system from 1CellBio. Hydrogel bead-cell co-encapsulation, in-drop synthesis of barcoded cDNA and library preparation was performed according to the recommended procedures of the manufacturer. Obtained libraries were sequenced using Illumina short read sequencing. Paired-end sequencing was generated using an Illumina NextSeq 500 sequencing device. Read 1 was used to obtain the sample barcode and UMI sequences, and read 2 was then mapped to a reference transcriptome using the indrops pipeline (https://github.com/indrops/indrops) as described below. The reads were first filtered based on presence in read 1 of two sample barcode components separated by the W1 adaptor sequence. Read 2 was then trimmed using Trimomatic (Bolger et al., 2014) (version 0.39; parameters: LEADING:28 SLIDINGWINDOW:4:20 MINLEN:16). Barcodes for each read were matched against a list of the inDrops v2 pre-determined barcodes from 1CellBio, and errors of up to two nucleotides mismatch were corrected. Reads with a barcode separated by more than two nucleotides from the reference list were discarded. The reads were then split into barcode specific files for mapping and UMI filtering, with UMI alignment performed according to (Macosko et al., 2015). Reads split into barcode-specific files were then aligned using Bowtie (Langmead et al., 2012) (version 1.0.0, parameters: -n 1 –l 15 –e 300 –m 200) to the human transcriptome using human reference genome version GRCh38 and ensemble v92 reference transcriptome.

## A.2 SUPPLEMENTAL FIGURES

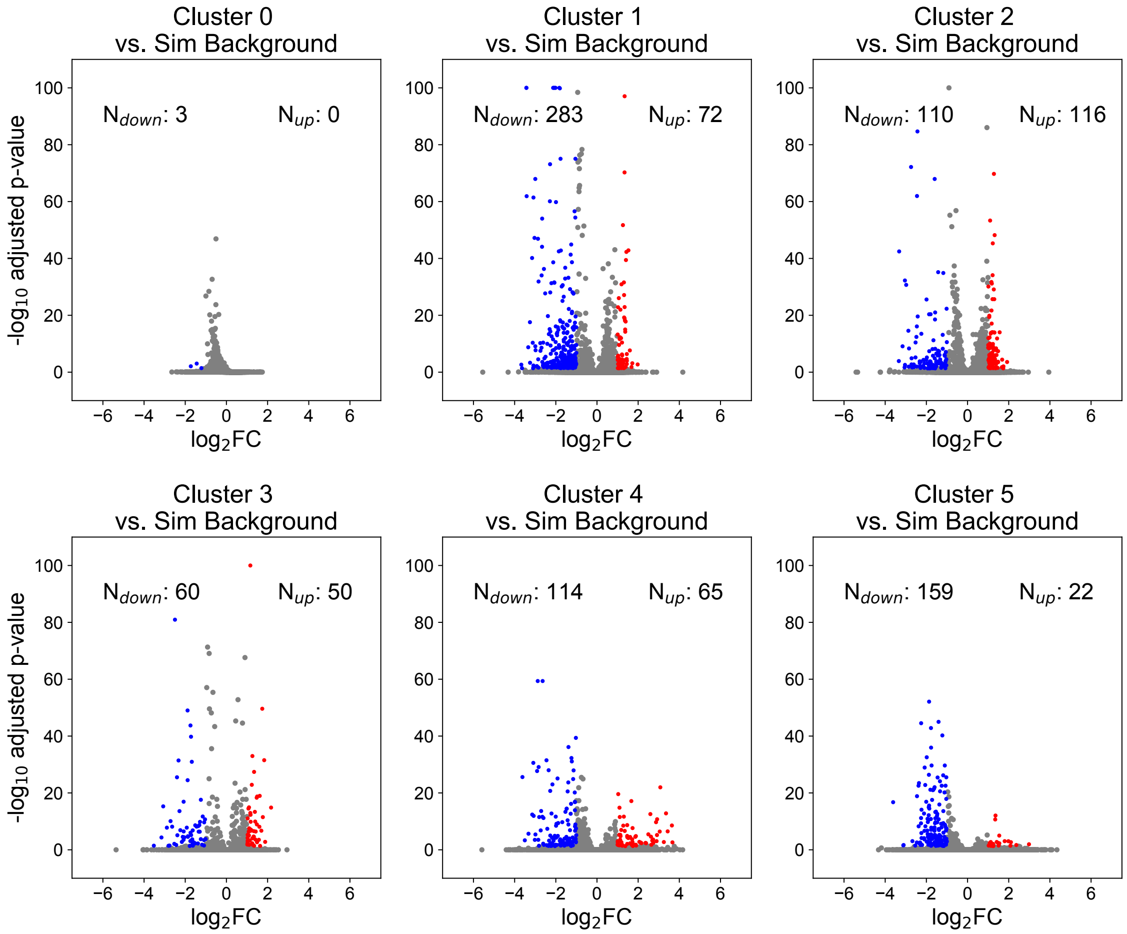

**Figure A1: Differential expression analysis comparing real cells with simulated background cells.** For each cluster from Figure 5.9, a differential expression analysis was performed comparing all real cells from the dataset in that cluster with the set of simulated background cells using a Wilcoxon rank sum-test. Cells were labeled as significantly differentially expressed if they had an absolute log-fold change $> 1$ and an adjusted p-value of $< 0.05$, after Benjamini–Hochberg multiple hypothesis correction.

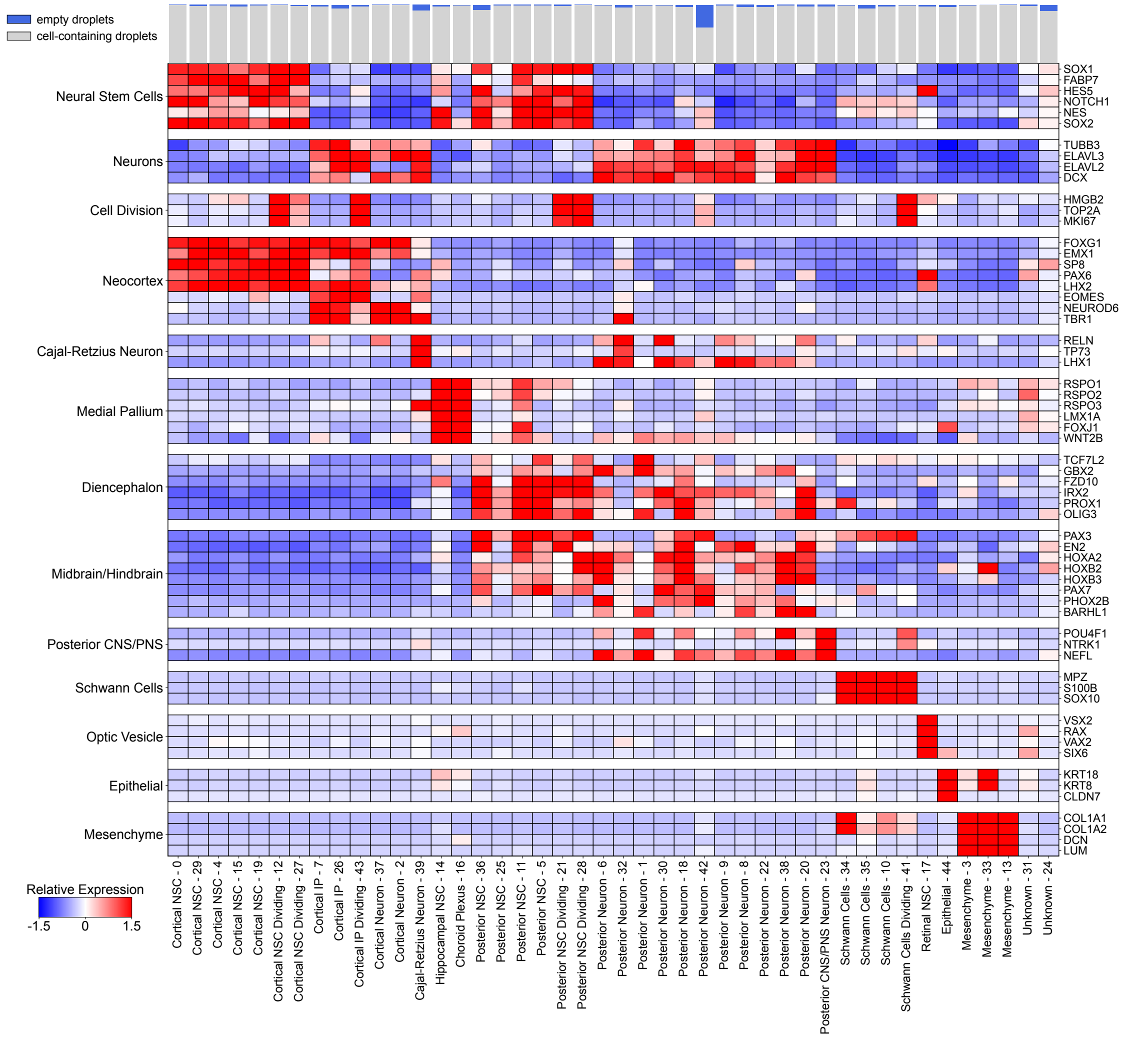

**Figure A2: Empty droplets per cluster in scRNA-Seq of day 50 organoids.** The heatmap displays the relative expression values after z-score normalization of the average log-normalized expression values for each gene across clusters after doublet removal for selected genes categorized according to cell state, cell type, and brain region from the dataset in (Rosebrock et al., 2022). The relative percentage of annotated empty droplets across all samples per cluster is highlighted in the top bar chart.

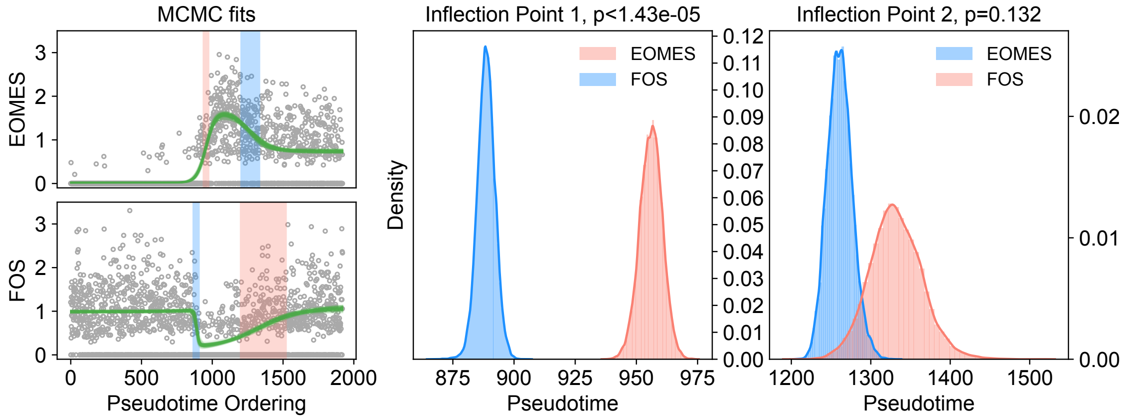

**Figure A3: Inflection point comparison of FOS and EOMES in day 30 cortical cells.** The left panel in the plot displays a random sampling of the parameters from 100 iterations from the MCMC traces for the double sigmoidal model, the best-fitting model for both genes. The full range of first and second inflection point estimates for both genes is highlights as a shaded region, with blue indicating a negative inflection point and red a positive inflection point. The middle panel highlights the distribution of first inflection point estimates across MCMC iterations. The overlap in the distributions is <0.00143%, indicating a non-simultaneous inflection point. The right panel highlights the distribution of second inflection point estimates across MCMC iterations. The overlap in the distributions if 13.2%, indicating a simultaneous inflection point.

# BIBLIOGRAPHY

Aibar, Sara et al. (Nov. 2017). "SCENIC: single-cell regulatory network inference and clustering." In: *Nature Methods* 14.11, pp. 1083–1086. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4463.

Aizarani, Nadim, Antonio Saviano, Sagar, Laurent Mailly, Sarah Durand, Josip S. Herman, Patrick Pessaux, Thomas F. Baumert, and Dominic Grün (Aug. 2019). "A human liver cell atlas reveals heterogeneity and epithelial progenitors." In: *Nature* 572.7768, pp. 199–204. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1373-2.

Akaike, H. (Dec. 1974). "A new look at the statistical model identification." In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723. ISSN: 0018-9286. DOI: 10.1109/TAC.1974.1100705.

Anders, S., P. T. Pyl, and W. Huber (Jan. 2015). "HTSeq–a Python framework to work with high-throughput sequencing data." In: *Bioinformatics* 31.2, pp. 166–169. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btu638.

Andoniadou, Cynthia Lilian and Juan Pedro Martinez-Barbera (Oct. 2013). "Developmental mechanisms directing early anterior forebrain specification in vertebrates." In: *Cellular and Molecular Life Sciences* 70.20, pp. 3739–3752. ISSN: 1420-682X, 1420-9071. DOI: 10.1007/s00018-013-1269-5.

Andrieu, Christophe, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan (2003). "An introduction to MCMC for machine learning." In: *Machine Learning* 50.1/2, pp. 5–43. ISSN: 08856125. DOI: 10.1023/A:1020281327116.

Bae, S., Y. Bessho, M. Hojo, and R. Kageyama (July 2000). "The bHLH gene Hes6, an inhibitor of Hes1, promotes neuronal differentiation." In: *Development* 127.13, pp. 2933–2943. ISSN: 1477-9129, 0950-1991. DOI: 10.1242/dev.127.13.2933.

Baione, Fabio, Davide Biancalana, and Paolo De Angelis (June 2021). "An application of Sigmoid and Double-Sigmoid functions for dynamic policyholder behaviour." In: *Decisions in Economics and Finance* 44.1, pp. 5–22. ISSN: 1593-8883, 1129-6569. DOI: 10.1007/s10203-020-00279-7.

Bannister, Andrew J and Tony Kouzarides (Mar. 2011). "Regulation of chromatin by histone modifications." In: *Cell Research* 21.3, pp. 381–395. ISSN: 1001-0602, 1748-7838. DOI: 10.1038/cr.2011.22.

Bar-Joseph, Ziv, Anthony Gitter, and Itamar Simon (Aug. 2012). "Studying and modelling dynamic biological processes using time-series gene expression data." In: *Nature Reviews Genetics* 13.8, pp. 552–564. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3244.

Becht, Etienne, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell (Jan. 2019). "Dimensionality reduction for visualizing single-cell data using UMAP." In: *Nature Biotechnology* 37.1, pp. 38–44. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.4314.

Benjamini, Yoav and Yosef Hochberg (Jan. 1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1, pp. 289–300. ISSN: 00359246. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Berge, Koen Van den, Hector Roux de Bézieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement (Dec. 2020). "Trajectory-based differential expression analysis for single-cell sequencing data." In: *Nature Communications* 11.1, p. 1201. ISSN: 2041-1723. DOI: 10.1038/s41467-020-14766-3.

Bershteyn, Marina, Tomasz J. Nowakowski, Alex A. Pollen, Elizabeth Di Lullo, Aishwarya Nene, Anthony Wynshaw-Boris, and Arnold R. Kriegstein (Apr. 2017). "Human iPSC-Derived Cerebral Organoids Model Cellular Features of Lissencephaly and Reveal Prolonged Mitosis of Outer Radial Glia." In: *Cell Stem Cell* 20.4, 435–449.e4. ISSN: 19345909. DOI: 10.1016/j.stem.2016.12.007.

Bishop, Christopher M. (2006). *Pattern recognition and machine learning*. Information science and statistics. New York: Springer. ISBN: 978-0-387-31073-2.

Blaess, Sandra, JoMichelle D. Corrales, and Alexandra L. Joyner (May 2006). "Sonic hedgehog regulates Gli activator and repressor functions with spatial and temporal precision in the mid/hindbrain region." In: *Development* 133.9, pp. 1799–1809. ISSN: 1477-9129, 0950-1991. DOI: 10.1242/dev.02339.

Blakeley, Paul, Norah M. E. Fogarty, Ignacio del Valle, Sissy E. Wamaitha, Tim Xiaoming Hu, Kay Elder, Philip Snell, Leila Christie, Paul Robson, and Kathy K. Niakan (Oct. 2015). "Defining the three cell lineages of the human blastocyst by single-cell RNA-seq." In: *Development* 142.20, pp. 3613–3613. ISSN: 1477-9129, 0950-1991. DOI: 10.1242/dev.131235.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (Oct. 2008). "Fast unfolding of communities in large networks." In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel (Aug. 2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." In: *Bioinformatics* 30.15, pp. 2114–2120. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btu170.

Bonn, Stefan, Robert P Zinzen, Charles Girardot, E Hilary Gustafson, Alexis Perez-Gonzalez, Nicolas Delhomme, Yad Ghavi-Helm, Bartek Wilczyński, Andrew Riddell, and Eileen E M Furlong (Feb. 2012). "Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development." In: *Nature Genetics* 44.2, pp. 148–156. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.1064.

Brennecke, Philip et al. (Nov. 2013). "Accounting for technical noise in single-cell RNA-seq experiments." In: *Nature Methods* 10.11, pp. 1093–1095. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.2645.

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija (May 2018). "Integrating single-cell transcriptomic data across different conditions, technologies, and species." In: *Nature Biotechnology* 36.5, pp. 411–420. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.4096.

Calo, Eliezer and Joanna Wysocka (Mar. 2013). "Modification of Enhancer Chromatin: What, How, and Why?" In: *Molecular Cell* 49.5, pp. 825–837. ISSN: 10972765. DOI: 10.1016/j.molcel.2013.01.038.

Campbell, Kieran R and Christopher Yau (Jan. 2019). "A descriptive marker gene approach to single-cell pseudotime inference." In: *Bioinformatics* 35.1. Ed. by Inanc Birol, pp. 28–35. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bty498.

Cao, Junyue et al. (Feb. 2019). "The single-cell transcriptional landscape of mammalian organogenesis." In: *Nature* 566.7745, pp. 496–502. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-0969-x.

Chambers, Stuart M, Christopher A Fasano, Eirini P Papapetrou, Mark Tomishima, Michel Sadelain, and Lorenz Studer (Mar. 2009). “Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling.” In: *Nature Biotechnology* 27.3, pp. 275–280. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.1529.

Chen, Chao, Garrett A. Lee, Ariel Pourmorady, Elisabeth Sock, and Maria J. Donoghue (July 2015). “Orchestration of Neuronal Differentiation and Progenitor Pool Expansion in the Developing Cortex by SoxC Genes.” In: *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 35.29, pp. 10629–10642. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.1663-15.2015.

Chen, Taiping and Sharon Y. R. Dent (Feb. 2014). “Chromatin modifiers and remodellers: regulators of cellular differentiation.” In: *Nature Reviews Genetics* 15.2, pp. 93–106. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3607.

Chinenov, Yurii and Tom K Kerppola (Apr. 2001). “Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity.” In: *Oncogene* 20.19, pp. 2438–2452. ISSN: 0950-9232, 1476-5594. DOI: 10.1038/sj.onc.1204385.

Chung, Neo Christopher and John D. Storey (Feb. 2015). “Statistical significance of variables driving systematic variation in high-dimensional data.” In: *Bioinformatics* 31.4, pp. 545–554. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btu674.

Clouaire, Thomas et al. (Oct. 2018). “Comprehensive Mapping of Histone Modifications at DNA Double-Strand Breaks Deciphers Repair Pathway Chromatin Signatures.” In: *Molecular Cell* 72.2, 250–262.e6. ISSN: 10972765. DOI: 10.1016/j.molcel.2018.08.020.

Coifman, R. R., S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker (May 2005). “Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps.” In: *Proceedings of the National Academy of Sciences* 102.21, pp. 7426–7431. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0500334102.

Cramér, Harald (1946). *Mathematical methods of statistics.* Princeton Mathematical Series 9. Princeton: Princeton University Press. ISBN: 978-0-691-08004-8.

Crick, F. H. (1958). “On protein synthesis.” In: *Symposia of the Society for Experimental Biology* 12, pp. 138–163. ISSN: 0081-1386.

Deaton, Aimée M. and Adrian Bird (May 2011). “CpG islands and the regulation of transcription.” In: *Genes & Development* 25.10, pp. 1010–1022. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.2037511.

Defays, D. (Apr. 1977). “An efficient algorithm for a complete link method.” In: *The Computer Journal* 20.4, pp. 364–366. ISSN: 0010-4620, 1460-2067. DOI: 10.1093/comjnl/20.4.364.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren (May 2012). “Topological domains in mammalian genomes identified by analysis of chromatin interactions.” In: *Nature* 485.7398, pp. 376–380. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11082.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras (Jan. 2013). “STAR: ultrafast universal RNA-seq aligner.” In: *Bioinformatics* 29.1, pp. 15–21. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/bts635.

Ehrlich, Melanie, Miguel A. Gama-Sosa, Lan-Hsiang Huang, Rose Marie Midgett, Kenneth C. Kuo, Roy A. McCune, and Charles Gehrke (1982). “Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells." In: *Nucleic Acids Research* 10.8, pp. 2709–2721. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/10.8.2709.

Eiraku, Mototsugu, Nozomu Takata, Hiroki Ishibashi, Masako Kawada, Eriko Sakakura, Satoru Okuda, Kiyotoshi Sekiguchi, Taiji Adachi, and Yoshiki Sasai (Apr. 2011). "Self-organizing optic-cup morphogenesis in three-dimensional culture." In: *Nature* 472.7341, pp. 51–56. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature09941.

Elkabetz, Yechiel, Georgia Panagiotakos, George Al Shamy, Nicholas D. Socci, Viviane Tabar, and Lorenz Studer (Jan. 2008). "Human ES cell-derived neural rosettes reveal a functionally distinct early neural stem cell stage." In: *Genes & Development* 22.2, pp. 152–165. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.1616208.

Elkabetz, Yechiel, Sneha Arora, Anastasios Balaskas, Amèlia Aragonés Hernández, and Daniel Rosebrock (June 2022). *Generation of Cerebral Organoids with Enriched Cortical Cellular Diversity and Outer Radial Glial Cell Identity from Human Pluripotent Stem Cells.* preprint. Protocol Exchange. DOI: 10.21203/rs.3.pex-1877/v1.

Elsen, Gina E., Francesco Bedogni, Rebecca D. Hodge, Theo K. Bammler, James W. MacDonald, Susan Lindtner, John L. R. Rubenstein, and Robert F. Hevner (Aug. 2018). "The Epigenetic Factor Landscape of Developing Neocortex Is Regulated by Transcription Factors Pax6→ Tbr2→ Tbr1." In: *Frontiers in Neuroscience* 12, p. 571. ISSN: 1662-453X. DOI: 10.3389/fnins.2018.00571.

Fame, Ryann M., Jessica L. MacDonald, and Jeffrey D. Macklis (Jan. 2011). "Development, specification, and diversity of callosal projection neurons." In: *Trends in Neurosciences* 34.1, pp. 41–50. ISSN: 01662236. DOI: 10.1016/j.tins.2010.10.002.

Farkas, Lilla M., Christiane Haffner, Thomas Giger, Philipp Khaitovich, Katja Nowick, Carmen Birchmeier, Svante Pääbo, and Wieland B. Huttner (Oct. 2008). "Insulinoma-Associated 1 Has a Panneurogenic Role and Promotes the Generation and Expansion of Basal Progenitors in the Developing Mouse Neocortex." In: *Neuron* 60.1, pp. 40–55. ISSN: 08966273. DOI: 10.1016/j.neuron.2008.09.020.

Ferent, Julien, Donia Zaidi, and Fiona Francis (Oct. 2020). "Extracellular Control of Radial Glia Proliferation and Scaffolding During Cortical Development and Pathology." In: *Frontiers in Cell and Developmental Biology* 8, p. 578341. ISSN: 2296-634X. DOI: 10.3389/fcell.2020.578341.

Florio, Marta and Wieland B. Huttner (June 2014). "Neural progenitors, neurogenesis and the evolution of the neocortex." In: *Development* 141.11, pp. 2182–2194. ISSN: 1477-9129, 0950-1991. DOI: 10.1242/dev.090571.

Foreman-Mackey, Daniel, David W. Hogg, Dustin Lang, and Jonathan Goodman (Mar. 2013). "emcee: The MCMC Hammer." In: *Publications of the Astronomical Society of the Pacific* 125.925. arXiv:1202.3665 [astro-ph, physics:physics, stat], pp. 306–312. ISSN: 00046280, 15383873. DOI: 10.1086/670067.

Fortunato, Santo and Marc Barthélemy (Jan. 2007). "Resolution limit in community detection." In: *Proceedings of the National Academy of Sciences* 104.1, pp. 36–41. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0605965104.

Fortunato, Santo and Darko Hric (Nov. 2016). "Community detection in networks: A user guide." In: *Physics Reports* 659. arXiv:1608.00163 [physics], pp. 1–44. ISSN: 03701573. DOI: 10.1016/j.physrep.2016.09.002.

GTEx Consortium (Oct. 2017). "Genetic effects on gene expression across human tissues." In: *Nature* 550.7675, pp. 204–213. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature24277.

Galluzzi, Lorenzo, Oliver Kepp, and Guido Kroemer (Dec. 2012). "Mitochondria: master regulators of danger signalling." In: *Nature Reviews Molecular Cell Biology* 13.12, pp. 780–788. ISSN: 1471-0072, 1471-0080. DOI: 10.1038/nrm3479.

Gelman, Andrew (2014). *Bayesian data analysis.* Third edition. Chapman & Hall/CRC texts in statistical science. Boca Raton: CRC Press. ISBN: 978-1-4398-4095-5.

Gilbert, Scott F (2009). *Developmental biology.* OCLC: 946202902. Place of publication not identified: Sinauer Associates. ISBN: 978-0-87893-371-6.

Goodman, Jonathan and Jonathan Weare (Jan. 2010). "Ensemble samplers with affine invariance." In: *Communications in Applied Mathematics and Computational Science* 5.1, pp. 65–80. ISSN: 2157-5452, 1559-3940. DOI: 10.2140/camcos.2010.5.65.

Greig, Luciano Custo, Mollie B. Woodworth, Maria J. Galazo, Hari Padmanabhan, and Jeffrey D. Macklis (Nov. 2013). "Molecular logic of neocortical projection neuron specification, development and diversity." In: *Nature Reviews Neuroscience* 14.11, pp. 755–769. ISSN: 1471-003X, 1471-0048. DOI: 10.1038/nrn3586.

Guelen, Lars et al. (June 2008). "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions." In: *Nature* 453.7197, pp. 948–951. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature06947.

Guo, Hongshan et al. (July 2014). "The DNA methylation landscape of human early embryos." In: *Nature* 511.7511, pp. 606–610. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature13544.

Haghverdi, Laleh, Florian Buettner, and Fabian J. Theis (Sept. 2015). "Diffusion maps for high-dimensional single-cell analysis of differentiation data." In: *Bioinformatics* 31.18, pp. 2989–2998. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btv325.

Haghverdi, Laleh, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis (Oct. 2016). "Diffusion pseudotime robustly reconstructs lineage branching." In: *Nature Methods* 13.10, pp. 845–848. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.3971.

Haines, Duane E. and Gregory A. Mihailoff, eds. (2018). *Fundamental neuroscience for basic and clinical applications.* Fifth edition. Philadelphia, PA: Elsevier. ISBN: 978-0-323-39632-5.

Hansen, David V., Jan H. Lui, Philip R. L. Parker, and Arnold R. Kriegstein (Mar. 2010). "Neurogenic radial glia in the outer subventricular zone of human neocortex." In: *Nature* 464.7288, pp. 554–561. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature08845.

Harms, Robbert L. and Alard Roebroeck (Dec. 2018). "Robust and Fast Markov Chain Monte Carlo Sampling of Diffusion MRI Microstructure Models." In: *Frontiers in Neuroinformatics* 12, p. 97. ISSN: 1662-5196. DOI: 10.3389/fninf.2018.00097.

Hashimshony, Tamar et al. (Dec. 2016). "CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq." In: *Genome Biology* 17.1, p. 77. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0938-8.

Hastings, W. K. (Apr. 1970). "Monte Carlo sampling methods using Markov chains and their applications." In: *Biometrika* 57.1, pp. 97–109. ISSN: 1464-3510, 0006-3444. DOI: 10.1093/biomet/57.1.97.

Heintzman, Nathaniel D. et al. (May 2009). "Histone modifications at human enhancers reflect global cell-type-specific gene expression." In: *Nature* 459.7243, pp. 108–112. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature07829.

Hevner, Robert F. (Sept. 2019). "Intermediate progenitors and Tbr2 in cortical development." In: *Journal of Anatomy* 235.3, pp. 616–625. ISSN: 0021-8782, 1469-7580. DOI: 10.1111/joa.12939.

Hill, Caroline S. (Oct. 2016). "Transcriptional Control by the SMADs." In: *Cold Spring Harbor Perspectives in Biology* 8.10, a022079. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a022079.

Hogg, David W. and Daniel Foreman-Mackey (May 2018). "Data Analysis Recipes: Using Markov Chain Monte Carlo." In: *The Astrophysical Journal Supplement Series* 236.1, p. 11. ISSN: 1538-4365. DOI: 10.3847/1538-4365/aab76e.

Holzwarth, Christina, Martin Vaegler, Friederike Gieseke, Stefan M Pfister, Rupert Handgretinger, Gunter Kerst, and Ingo Müller (Dec. 2010). "Low physiologic oxygen tensions reduce proliferation and differentiation of human multipotent mesenchymal stromal cells." In: *BMC Cell Biology* 11.1, p. 11. ISSN: 1471-2121. DOI: 10.1186/1471-2121-11-11.

Hotelling, H. (Sept. 1933). "Analysis of a complex of statistical variables into principal components." In: *Journal of Educational Psychology* 24.6, pp. 417–441. ISSN: 1939-2176, 0022-0663. DOI: 10.1037/h0071325.

Hou, Fengji, Jonathan Goodman, David W. Hogg, Jonathan Weare, and Christian Schwab (2011). "An Affine-Invariant Sampler for Exoplanet Fitting and Discovery in Radial Velocity Data." In: Publisher: arXiv Version Number: 2. DOI: 10.48550/ARXIV.1104.2612.

Ichiyama, Kenji et al. (May 2011). "Transcription Factor Smad-Independent T Helper 17 Cell Induction by Transforming-Growth Factor-$\beta$ Is Mediated by Suppression of Eomesodermin." In: *Immunity* 34.5, pp. 741–754. ISSN: 10747613. DOI: 10.1016/j.immuni.2011.02.021.

Ilicic, Tomislav, Jong Kyoung Kim, Aleksandra A. Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John C. Marioni, and Sarah A. Teichmann (Dec. 2016). "Classification of low quality cells from single-cell RNA-seq data." In: *Genome Biology* 17.1, p. 29. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0888-1.

International Human Genome Sequencing Consortium et al. (Feb. 2001). "Initial sequencing and analysis of the human genome." In: *Nature* 409.6822, pp. 860–921. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/35057062.

Jaitin, D. A. et al. (Feb. 2014). "Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types." In: *Science* 343.6172, pp. 776–779. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1247651.

Ji, Zhicheng and Hongkai Ji (July 2016). "TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis." In: *Nucleic Acids Research* 44.13, e117–e117. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkw430.

Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold (June 2007). "Genome-Wide Mapping of in Vivo Protein-DNA Interactions." In: *Science* 316.5830, pp. 1497–1502. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1141319.

Kadoshima, Taisuke, Hideya Sakaguchi, Tokushige Nakano, Mika Soen, Satoshi Ando, Mototsugu Eiraku, and Yoshiki Sasai (Dec. 2013). "Self-organization of axial polarity, inside-out layer pattern, and species-specific progenitor dynamics in human ES cell–derived neocortex." In: *Proceedings of the National Academy of Sciences* 110.50, pp. 20284–20289. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1315710110.

Kageyama, Ryoichiro, Toshiyuki Ohtsuka, and Taeko Kobayashi (Apr. 2008). "Roles of Hes genes in neural development: Hes genes in neural development." In: *Development, Growth & Differentiation* 50, S97–S103. ISSN: 00121592, 1440169X. DOI: 10.1111/j.1440-169X.2008.00993.x.

Kang, Hyo Jung et al. (Oct. 2011). "Spatio-temporal transcriptome of the human brain." In: *Nature* 478.7370, pp. 483–489. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature10523.

Kang, Hyun Min et al. (Jan. 2018). "Multiplexed droplet single-cell RNA-sequencing using natural genetic variation." In: *Nature Biotechnology* 36.1, pp. 89–94. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.4042.

Kelava, Iva and Madeline A. Lancaster (June 2016). “Stem Cell Models of Human Brain Development.” In: *Cell Stem Cell* 18.6, pp. 736–748. ISSN: 19345909. DOI: 10.1016/j.stem.2016.05.022.

Klein, Allon M., Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner (May 2015). “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells.” In: *Cell* 161.5, pp. 1187–1201. ISSN: 00928674. DOI: 10.1016/j.cell.2015.04.044.

Kotrys, Anna V. and Roman J. Szczesny (Dec. 2019). “Mitochondrial Gene Expression and Beyond—Novel Aspects of Cellular Physiology.” In: *Cells* 9.1, p. 17. ISSN: 2073-4409. DOI: 10.3390/cells9010017.

Kovach, Christopher, Rajiv Dixit, Saiqun Li, Pierre Mattar, Grey Wilkinson, Gina E. Elsen, Deborah M. Kurrasch, Robert F. Hevner, and Carol Schuurmans (Aug. 2013). “Neurog2 Simultaneously Activates and Represses Alternative Gene Expression Programs in the Developing Neocortex.” In: *Cerebral Cortex* 23.8, pp. 1884–1900. ISSN: 1460-2199, 1047-3211. DOI: 10.1093/cercor/bhs176.

Kriegstein, Arnold R. and Magdalena Götz (July 2003). “Radial glia diversity: a matter of cell fate.” In: *Glia* 43.1, pp. 37–43. ISSN: 0894-1491. DOI: 10.1002/glia.10250.

L. Lun, Aaron T., Karsten Bach, and John C. Marioni (Dec. 2016). “Pooling across cells to normalize single-cell RNA sequencing data with many zero counts.” In: *Genome Biology* 17.1, p. 75. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0947-7.

Lancaster, Madeline A., Magdalena Renner, Carol-Anne Martin, Daniel Wenzel, Louise S. Bicknell, Matthew E. Hurles, Tessa Homfray, Josef M. Penninger, Andrew P. Jackson, and Juergen A. Knoblich (Sept. 2013). “Cerebral organoids model human brain development and microcephaly.” In: *Nature* 501.7467, pp. 373–379. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12517.

Lange, Marius et al. (Feb. 2022). “CellRank for directed single-cell fate mapping.” In: *Nature Methods* 19.2, pp. 159–170. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-021-01346-6.

Langmead, Ben and Steven L Salzberg (Apr. 2012). “Fast gapped-read alignment with Bowtie 2.” In: *Nature Methods* 9.4, pp. 357–359. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1923.

Lause, Jan, Philipp Berens, and Dmitry Kobak (Dec. 2021). “Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data.” In: *Genome Biology* 22.1, p. 258. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02451-7.

Li, Bo, Sei Kuriyama, Mauricio Moreno, and Roberto Mayor (Oct. 2009). “The posteriorizing gene *Gbx2* is a direct target of Wnt signalling and the earliest factor in neural crest induction.” In: *Development* 136.19, pp. 3267–3278. ISSN: 1477-9129, 0950-1991. DOI: 10.1242/dev.036954.

Li, Bo and Colin N Dewey (Dec. 2011). “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.” In: *BMC Bioinformatics* 12.1, p. 323. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-323.

Link, William A. and Mitchell J. Eaton (Feb. 2012). “On thinning of chains in MCMC: *Thinning of MCMC chains*.” In: *Methods in Ecology and Evolution* 3.1, pp. 112–115. ISSN: 2041210X. DOI: 10.1111/j.2041-210X.2011.00131.x.

Logan, Catriona Y. and Roel Nusse (Nov. 2004). “The Wnt signaling pathway in development and disease.” In: *Annual Review of Cell and Developmental Biology* 20.1, pp. 781–810. ISSN: 1081-0706, 1530-8995. DOI: 10.1146/annurev.cellbio.20.010403.113126.

Love, Michael I, Wolfgang Huber, and Simon Anders (Dec. 2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” In: *Genome Biology* 15.12, p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8.

Luecken, Malte D and Fabian J Theis (June 2019). “Current best practices in single-cell RNA-seq analysis: a tutorial.” In: *Molecular Systems Biology* 15.6. ISSN: 1744-4292, 1744-4292. DOI: 10.15252/msb.20188746.

Lumsden, Andrew and Robb Krumlauf (Nov. 1996). “Patterning the Vertebrate Neuraxis.” In: *Science* 274, pp. 1109–1115. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.274.5290.1109.

Lun, Aaron T. L., Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, and John C. Marioni (Dec. 2019). “EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data.” In: *Genome Biology* 20.1, p. 63. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1662-y.

MacQueen, J. B. (1967). “Some Methods for Classification and Analysis of MultiVariate Observations.” In: *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by L. M. Le Cam and J. Neyman. Vol. 1. University of California Press, pp. 281–297.

Macosko, Evan Z. et al. (May 2015). “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” In: *Cell* 161.5, pp. 1202–1214. ISSN: 00928674. DOI: 10.1016/j.cell.2015.05.002.

Marcelino, Jose, Christopher M. Sciortino, Michael F. Romero, Lynn M. Ulatowski, R. Tracy Ballock, Aris N. Economides, Peter M. Eimon, Richard M. Harland, and Matthew L. Warman (Sept. 2001). “Human disease-causing *NOG* missense mutations: Effects on noggin secretion, dimer formation, and bone morphogenetic protein binding.” In: *Proceedings of the National Academy of Sciences* 98.20, pp. 11353–11358. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.201367598.

McGinnis, Christopher S., Lyndsay M. Murrow, and Zev J. Gartner (Apr. 2019). “DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors.” In: *Cell Systems* 8.4, 329–337.e4. ISSN: 24054712. DOI: 10.1016/j.cels.2019.03.003.

McInnes, Leland, John Healy, and James Melville (2018). “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” In: Publisher: arXiv Version Number: 3. DOI: 10.48550/ARXIV.1802.03426.

McMahon, A.P., A.L. Joyner, A. Bradley, and J.A. McMahon (May 1992). “The midbrain-hindbrain phenotype of Wnt-1-Wnt-1- mice results from stepwise deletion of engrailed-expressing cells by 9.5 days postcoitum.” In: *Cell* 69.4, pp. 581–595. ISSN: 00928674. DOI: 10.1016/0092-8674(92)90222-X.

McMahon, Jill A., Shinji Takada, Lyle B. Zimmerman, Chen-Ming Fan, Richard M. Harland, and Andrew P. McMahon (May 1998). “Noggin-mediated antagonism of BMP signaling is required for growth and patterning of the neural tube and somite.” In: *Genes & Development* 12.10, pp. 1438–1452. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.12.10.1438.

Metropolis, Nicholas and S. Ulam (Sept. 1949). “The Monte Carlo Method.” In: *Journal of the American Statistical Association* 44.247, pp. 335–341. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.1949.10483310.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller (June 1953). “Equation of State Calculations by Fast Computing Machines.” In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.1699114.

Molyneaux, Bradley J., Paola Arlotta, Joao R. L. Menezes, and Jeffrey D. Macklis (June 2007). “Neuronal subtype specification in the cerebral cortex.” In: *Nature Reviews Neuroscience* 8.6, pp. 427–437. ISSN: 1471-003X, 1471-0048. DOI: 10.1038/nrn2151.

Montoro, Daniel T. et al. (Aug. 2018). “A revised airway epithelial hierarchy includes CFTR-expressing ionocytes.” In: *Nature* 560.7718, pp. 319–324. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0393-7.

Moon, Kevin R., Jay S. Stanley, Daniel Burkhardt, David van Dijk, Guy Wolf, and Smita Krishnaswamy (Feb. 2018). “Manifold learning-based methods for analyzing single-cell RNA-sequencing data.” In: *Current Opinion in Systems Biology* 7, pp. 36–46. ISSN: 24523100. DOI: 10.1016/j.coisb.2017.12.008.

Morrison, Sean J. and Judith Kimble (June 2006). “Asymmetric and symmetric stem-cell divisions in development and cancer.” In: *Nature* 441.7097, pp. 1068–1074. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature04956.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold (July 2008). “Mapping and quantifying mammalian transcriptomes by RNA-Seq.” In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1226.

Newman, M. E. J. and M. Girvan (Feb. 2004). “Finding and evaluating community structure in networks.” In: *Physical Review E* 69.2, p. 026113. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.69.026113.

Nott, Alexi et al. (Nov. 2019). “Brain cell type–specific enhancer–promoter interactome maps and disease - risk association.” In: *Science* 366.6469, pp. 1134–1139. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aay0793.

Nye, J.S., R. Kopan, and R. Axel (Sept. 1994). “An activated Notch suppresses neurogenesis and myogenesis but not gliogenesis in mammalian cells.” In: *Development* 120.9, pp. 2421–2430. ISSN: 1477-9129, 0950-1991. DOI: 10.1242/dev.120.9.2421.

Ohno, S. (1972). “So much "junk" DNA in our genome.” In: *Brookhaven Symposia in Biology* 23, pp. 366–370. ISSN: 0068-2799.

Paşca, Anca M et al. (July 2015). “Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture.” In: *Nature Methods* 12.7, pp. 671–678. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.3415.

Patestas, Maria Antoniou and Leslie P. Gartner (2016). *A textbook of neuroanatomy*. Second edition. Hoboken, New Jersey: Wiley Blackwell. ISBN: 978-1-118-67722-3 978-1-118-67735-3.

Pearson, Karl (Nov. 1901). “LIII. *On lines and planes of closest fit to systems of points in space*.” In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, pp. 559–572. ISSN: 1941-5982, 1941-5990. DOI: 10.1080/14786440109462720.

Pennacchio, Len A., Wendy Bickmore, Ann Dean, Marcelo A. Nobrega, and Gill Bejerano (Apr. 2013). “Enhancers: five essential questions.” In: *Nature Reviews Genetics* 14.4, pp. 288–295. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3458.

Picelli, Simone, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg (Nov. 2013). “Smart-seq2 for sensitive full-length transcriptome profiling in single cells.” In: *Nature Methods* 10.11, pp. 1096–1098. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.2639.

Plasschaert, Lindsey W., Rapolas Žilionis, Rayman Choo-Wing, Virginia Savova, Judith Knehr, Guglielmo Roma, Allon M. Klein, and Aron B. Jaffe (Aug. 2018). “A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte." In: *Nature* 560.7718, pp. 377–381. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0394-6.

Pombo, Ana and Niall Dillon (Apr. 2015). "Three-dimensional genome architecture: players and mechanisms." In: *Nature Reviews Molecular Cell Biology* 16.4, pp. 245–257. ISSN: 1471-0072, 1471-0080. DOI: 10.1038/nrm3965.

Ptashne, Mark (Aug. 1986). "Gene regulation by proteins acting nearby and at a distance." In: *Nature* 322.6081, pp. 697–701. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/322697a0.

Qian, Xuyu et al. (May 2016). "Brain-Region-Specific Organoids Using Mini-bioreactors for Modeling ZIKV Exposure." In: *Cell* 165.5, pp. 1238–1254. ISSN: 00928674. DOI: 10.1016/j.cell.2016.04.032.

Qiu, Xiaojie, Arman Rahimzamani, Li Wang, Bingcheng Ren, Qi Mao, Timothy Durham, José L. McFaline-Figueroa, Lauren Saunders, Cole Trapnell, and Sreeram Kannan (Mar. 2020). "Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe." In: *Cell Systems* 10.3, 265–274.e11. ISSN: 24054712. DOI: 10.1016/j.cels.2020.02.003.

Quinn, Jane C., Michael Molinek, Ben S. Martynoga, Paulette A. Zaki, Andrea Faedo, Alessandro Bulfone, Robert F. Hevner, John D. West, and David J. Price (Feb. 2007). "Pax6 controls cerebral cortical cell number by regulating exit from the cell cycle and specifies cortical cell identity by a cell autonomous mechanism." In: *Developmental Biology* 302.1, pp. 50–65. ISSN: 00121606. DOI: 10.1016/j.ydbio.2006.08.035.

Ramírez, Fidel, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A. Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke (Dec. 2018). "High-resolution TADs reveal DNA sequences underlying genome organization in flies." In: *Nature Communications* 9.1, p. 189. ISSN: 2041-1723. DOI: 10.1038/s41467-017-02525-w.

Regev, Aviv et al. (Dec. 2017). "The Human Cell Atlas." In: *eLife* 6, e27041. ISSN: 2050-084X. DOI: 10.7554/eLife.27041.

Reichardt, Jörg and Stefan Bornholdt (July 2006). "Statistical mechanics of community detection." In: *Physical Review E* 74.1, p. 016110. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.74.016110.

Ribes, V. and J. Briscoe (Aug. 2009). "Establishing and Interpreting Graded Sonic Hedgehog Signaling during Vertebrate Neural Tube Patterning: The Role of Negative Feedback." In: *Cold Spring Harbor Perspectives in Biology* 1.2, a002014–a002014. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a002014.

Richmond, Timothy J. and Curt A. Davey (May 2003). "The structure of DNA in the nucleosome core." In: *Nature* 423.6936, pp. 145–150. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature01595.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth (Apr. 2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." In: *Nucleic Acids Research* 43.7, e47–e47. ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gkv007.

Robert, Christian P. and George Casella (1999). *Monte Carlo statistical methods.* Springer texts in statistics. New York: Springer. ISBN: 978-0-387-98707-1.

Robertson, Gordon et al. (Aug. 2007). "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." In: *Nature Methods* 4.8, pp. 651–657. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth1068.

Roh, Tae-Young, Suresh Cuddapah, Kairong Cui, and Keji Zhao (Oct. 2006). "The genomic landscape of histone modifications in human T cells." In: *Proceedings of the National Academy of Sciences* 103.43, pp. 15782–15787. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0607617103.

Rose, Nathan R. and Robert J. Klose (Dec. 2014). "Understanding the relationship between DNA methylation and histone lysine methylation." In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1839.12, pp. 1362–1372. ISSN: 18749399. DOI: 10.1016/j.bbagrm.2014.02.007.

Rosebrock, Daniel et al. (June 2022). "Enhanced cortical neural stem cell identity through short SMAD and WNT inhibition in human cerebral organoids facilitates emergence of outer radial glial cells." In: *Nature Cell Biology* 24.6, pp. 981–995. ISSN: 1465-7392, 1476-4679. DOI: 10.1038/s41556-022-00929-5.

Roy, Vivekananda (Mar. 2020). "Convergence Diagnostics for Markov Chain Monte Carlo." In: *Annual Review of Statistics and Its Application* 7.1, pp. 387–412. ISSN: 2326-8298, 2326-831X. DOI: 10.1146/annurev-statistics-031219-041300.

Saelens, Wouter, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys (May 2019). "A comparison of single-cell trajectory inference methods." In: *Nature Biotechnology* 37.5, pp. 547–554. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-019-0071-9.

Sanes, Dan Harvey, Thomas A. Reh, and William A. Harris (2012). *Development of the nervous system*. 3rd ed. OCLC: ocn667213240. Amsterdam ; Boston : Burlington, MA: Elsevier ; Academic Press. ISBN: 978-0-12-374539-2.

Sansom, Stephen N., Dean S. Griffiths, Andrea Faedo, Dirk-Jan Kleinjan, Youlin Ruan, James Smith, Veronica van Heyningen, John L. Rubenstein, and Frederick J. Livesey (June 2009). "The Level of the Transcription Factor Pax6 Is Essential for Controlling the Balance between Neural Stem Cell Self-Renewal and Neurogenesis." In: *PLoS Genetics* 5.6. Ed. by Jean M. Hébert, e1000511. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000511.

Saxonov, Serge, Paul Berg, and Douglas L. Brutlag (Jan. 2006). "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters." In: *Proceedings of the National Academy of Sciences* 103.5, pp. 1412–1417. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0510310103.

Schena, Mark, Dari Shalon, Ronald W. Davis, and Patrick O. Brown (Oct. 1995). "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray." In: *Science* 270.5235, pp. 467–470. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.270.5235.467.

Schwarz, Gideon (Mar. 1978). "Estimating the Dimension of a Model." In: *The Annals of Statistics* 6.2. ISSN: 0090-5364. DOI: 10.1214/aos/1176344136.

Serfozo, Richard (2009). *Basics of Applied Stochastic Processes*. Ed. by Joe Gani, Chris Heyde, Peter Jagers, and Thomas G. Kurtz. Probability and Its Applications. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-89331-8 978-3-540-89332-5. DOI: 10.1007/978-3-540-89332-5.

Sessa, Alessandro, Ernesto Ciabatti, Daniela Drechsel, Luca Massimino, Gaia Colasante, Serena Giannelli, Takashi Satoh, Shizuo Akira, Francois Guillemot, and Broccoli Vania (Sept. 2016). "The Tbr2 Molecular Network Controls Cortical Neuronal Differentiation Through Complementary Genetic and Epigenetic Pathways." In: *Cerebral Cortex*, cercor;bhw270v1. ISSN: 1047-3211, 1460-2199. DOI: 10.1093/cercor/bhw270.

Setty, Manu, Vaidotas Kiseliovas, Jacob Levine, Adam Gayoso, Linas Mazutis, and Dana Pe'er (Apr. 2019). "Characterization of cell fate probabilities in single-cell data with Palantir." In:

*Nature Biotechnology* 37.4, pp. 451–460. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-019-0068-4.

Shao, Xin, Jie Liao, Chengyu Li, Xiaoyan Lu, Junyun Cheng, and Xiaohui Fan (July 2021). “CellTalkDB: a manually curated database of ligand–receptor interactions in humans and mice.” In: *Briefings in Bioinformatics* 22.4, bbaa269. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbaa269.

Sibson, R. (Jan. 1973). “SLINK: An optimally efficient algorithm for the single-link cluster method.” In: *The Computer Journal* 16.1, pp. 30–34. ISSN: 0010-4620, 1460-2067. DOI: 10.1093/comjnl/16.1.30.

Smith, Zachary D., Michelle M. Chan, Kathryn C. Humm, Rahul Karnik, Shila Mekhoubad, Aviv Regev, Kevin Eggan, and Alexander Meissner (July 2014). “DNA methylation dynamics of the human preimplantation embryo.” In: *Nature* 511.7511, pp. 611–615. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature13581.

Specht, Alicia T. and Jun Li (Dec. 2016). “LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering.” In: *Bioinformatics*, btw729. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btw729.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde (June 2014). “The deviance information criterion: 12 years on.” In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.3, pp. 485–493. ISSN: 13697412. DOI: 10.1111/rssb.12062.

Street, Kelly, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit (Dec. 2018). “Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics.” In: *BMC Genomics* 19.1, p. 477. ISSN: 1471-2164. DOI: 10.1186/s12864-018-4772-0.

Subramanian, Aravind et al. (Oct. 2005). “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.” In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0506580102.

Svensson, Valentine (Feb. 2020). “Droplet scRNA-seq is not zero-inflated.” In: *Nature Biotechnology* 38.2, pp. 147–150. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-019-0379-5.

Takahashi, Kazutoshi and Shinya Yamanaka (Aug. 2006). “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors.” In: *Cell* 126.4, pp. 663–676. ISSN: 00928674. DOI: 10.1016/j.cell.2006.07.024.

Takahashi, Kazutoshi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka (Nov. 2007). “Induction of pluripotent stem cells from adult human fibroblasts by defined factors.” In: *Cell* 131.5, pp. 861–872. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.11.019.

Taminato, Tomohito, Daisuke Yokota, Soh Araki, Hiroki Ovara, Kyo Yamasu, and Akinori Kawamura (Aug. 2016). “Enhancer activity-based identification of functional enhancers using zebrafish embryos.” In: *Genomics* 108.2, pp. 102–107. ISSN: 08887543. DOI: 10.1016/j.ygeno.2016.05.005.

Tang, Fuchou et al. (May 2009). “mRNA-Seq whole-transcriptome analysis of a single cell.” In: *Nature Methods* 6.5, pp. 377–382. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1315.

The ENCODE Project Consortium (Sept. 2012). “An integrated encyclopedia of DNA elements in the human genome.” In: *Nature* 489.7414, pp. 57–74. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11247.

The Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators (Oct. 2018). "Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris." In: *Nature* 562.7727, pp. 367–372. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0590-4.

Thomson, J. A., J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones (Nov. 1998). "Embryonic stem cell lines derived from human blastocysts." In: *Science (New York, N.Y.)* 282.5391, pp. 1145–1147. ISSN: 0036-8075. DOI: 10.1126/science.282.5391.1145.

Tirosh, Itay et al. (Apr. 2016). "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq." In: *Science* 352.6282, pp. 189–196. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aad0501.

Traag, V. A., L. Waltman, and N. J. van Eck (Dec. 2019). "From Louvain to Leiden: guaranteeing well-connected communities." In: *Scientific Reports* 9.1, p. 5233. ISSN: 2045-2322. DOI: 10.1038/s41598-019-41695-z.

Treutlein, Barbara, Doug G. Brownfield, Angela R. Wu, Norma F. Neff, Gary L. Mantalas, F. Hernan Espinoza, Tushar J. Desai, Mark A. Krasnow, and Stephen R. Quake (May 2014). "Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq." In: *Nature* 509.7500, pp. 371–375. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature13173.

Vallejos, Catalina A, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni (June 2017). "Normalizing single-cell RNA sequencing data: challenges and opportunities." In: *Nature Methods* 14.6, pp. 565–571. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4292.

Vallier, Ludovic, Daniel Reynolds, and Roger A. Pedersen (Nov. 2004). "Nodal inhibits differentiation of human embryonic stem cells along the neuroectodermal default pathway." In: *Developmental Biology* 275.2, pp. 403–421. ISSN: 00121606. DOI: 10.1016/j.ydbio.2004.08.031.

Velasco, Silvia et al. (June 2019). "Individual brain organoids reproducibly form cell diversity of the human cerebral cortex." In: *Nature* 570.7762, pp. 523–527. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1289-x.

Waddington, C. H. (1957). *The strategy of the genes: A discussion of some aspects of theoretical biology*. London: Allen & Unwin.

Wang, Xiaohu, Lu Ni, Siyuan Wan, Xiaohong Zhao, Xiao Ding, Anne Dejean, and Chen Dong (Feb. 2020). "Febrile Temperature Critically Controls the Differentiation and Pathogenicity of T Helper 17 Cells." In: *Immunity* 52.2, 328–341.e5. ISSN: 10747613. DOI: 10.1016/j.immuni.2020.01.006.

Wang, Zhong, Mark Gerstein, and Michael Snyder (Jan. 2009). "RNA-Seq: a revolutionary tool for transcriptomics." In: *Nature Reviews Genetics* 10.1, pp. 57–63. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg2484.

Watanabe, Kiichi, Daisuke Kamiya, Ayaka Nishiyama, Tomoko Katayama, Satoshi Nozaki, Hiroshi Kawasaki, Yasuyoshi Watanabe, Kenji Mizuseki, and Yoshiki Sasai (Mar. 2005). "Directed differentiation of telencephalic precursors from embryonic stem cells." In: *Nature Neuroscience* 8.3, pp. 288–296. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn1402.

Weinreb, Caleb, Samuel Wolock, and Allon M Klein (Apr. 2018). "SPRING: a kinetic interface for visualizing high dimensional single-cell expression data." In: *Bioinformatics* 34.7. Ed. by Bonnie Berger, pp. 1246–1248. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btx792.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis (Dec. 2018). "SCANPY: large-scale single-cell gene expression data analysis." In: *Genome Biology* 19.1, p. 15. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0.

Wolock, Samuel L., Romain Lopez, and Allon M. Klein (Apr. 2019). "Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data." In: *Cell Systems* 8.4, 281–291.e9. ISSN: 24054712. DOI: 10.1016/j.cels.2018.11.005.

Wu, Guangming et al. (Dec. 2010). "Initiation of trophectoderm lineage specification in mouse embryos is independent of Cdx2." In: *Development* 137.24, pp. 4159–4169. ISSN: 1477-9129, 0950-1991. DOI: 10.1242/dev.056630.

Yang, Shiyi, Sean E. Corbett, Yusuke Koga, Zhe Wang, W Evan Johnson, Masanao Yajima, and Joshua D. Campbell (Dec. 2020). "Decontamination of ambient RNA in single-cell RNA-seq with DecontX." In: *Genome Biology* 21.1, p. 57. ISSN: 1474-760X. DOI: 10.1186/s13059-020-1950-6.

Ying, Qi-Long, Marios Stavridis, Dean Griffiths, Meng Li, and Austin Smith (Feb. 2003). "Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture." In: *Nature Biotechnology* 21.2, pp. 183–186. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt780.

Young, Matthew D and Sam Behjati (Dec. 2020). "SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data." In: *GigaScience* 9.12, giaa151. ISSN: 2047-217X. DOI: 10.1093/gigascience/giaa151.

Zhang, Su-Chun, Marius Wernig, Ian D. Duncan, Oliver Brüstle, and James A. Thomson (Dec. 2001). "In vitro differentiation of transplantable neural precursors from human embryonic stem cells." In: *Nature Biotechnology* 19.12, pp. 1129–1133. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt1201-1129.

Zhang, Xiannian, Tianqi Li, Feng Liu, Yaqi Chen, Jiacheng Yao, Zeyao Li, Yanyi Huang, and Jianbin Wang (Jan. 2019). "Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems." In: *Molecular Cell* 73.1, 130–142.e5. ISSN: 10972765. DOI: 10.1016/j.molcel.2018.10.020.

Zheng, Grace X. Y. et al. (Apr. 2017). "Massively parallel digital transcriptional profiling of single cells." In: *Nature Communications* 8.1, p. 14049. ISSN: 2041-1723. DOI: 10.1038/ncomms14049.

Ziller, Michael J. et al. (Aug. 2013). "Charting a dynamic DNA methylation landscape of the human genome." In: *Nature* 500.7463, pp. 477–481. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12433.

la, J.L. Pompa de et al. (Mar. 1997). "Conservation of the Notch signalling pathway in mammalian neurogenesis." In: *Development* 124.6, pp. 1139–1148. ISSN: 1477-9129, 0950-1991. DOI: 10.1242/dev.124.6.1139.

van der Maaten, L.J.P. and G.E. Hinton (2008). "Visualizing High-Dimensional Data Using t-SNE." In: *Journal of Machine Learning Research* 9.nov. Pagination: 27, pp. 2579–2605. ISSN: 1532-4435.

# ABSTRACT

Over the past decade, there has been a rapid expansion in the development and utilization of brain organoid models, enabling three-dimensional in vivo-like views of fundamental neurodevelopmental features of corticogenesis in health and disease. Nonetheless, the methods used for generating cortical organoid fates exhibit widespread heterogeneity across different cell lines. Here, we show that a combination of dual SMAD and WNT inhibition (Triple-i protocol) establishes a robust cortical identity in brain organoids, while other widely used derivation protocols are inconsistent with respect to regional specification. In order to measure this heterogeneity, we employ single-cell RNA-sequencing (scRNA-Seq), enabling the sampling of the gene expression profiles of thousands of cells in an individual sample. However, in order to draw meaningful conclusions from scRNA-Seq data, technical artifacts must be identified and removed. In this thesis, we present a method to detect one such artifact, empty droplets that do not contain a cell and consist mainly of free-floating mRNA in the sample. Furthermore, from their expression profiles, cells can be ordered along a developmental trajectory which recapitulates the progression of cells as they differentiate. Based on this ordering, we model gene expression using a Bayesian inference approach in order to measure transcriptional dynamics within differentiating cells. This enables the ordering of genes along transcriptional cascades, statistical testing for differences in gene expression changes, and measuring potential regulatory gene interactions. We apply this approach to differentiating cortical neural stem cells into cortical neurons via an intermediate progenitor cell type in brain organoids to provide a detailed characterization of the endogenous molecular processes underlying neurogenesis.

# ZUSAMMENFASSUNG

Im letzten Jahrzent hat die Entwicklung und Nutzung von Organoidmodellen des Gehirns stark zugenommen. Diese Modelle erlauben dreidimensionale, in-vivo ähnliche Einblicke in fundamentale Aspekte der neurologischen Entwicklung des Hirnkortex in Gesundheit und Krankheit. Jedoch weisen die Methoden, um die Entwicklung kortikaler Organoide zu verfolgen, starke Heterogenität zwischen verschiedenen Zelllinien auf. Hier weisen wir nach, dass eine Kombination dualer SMAD und WNT Hemmung (Triple-i Protokoll) eine konstante kortikale Zuordnung in Hirnorganoiden erzeugt, während andere, weit verbreitete und genutzte Protokolle in Bezug auf kortikale Spezifizierung keine konstanten Ergebnisse liefern. Um die Heterogenität zu messen, haben wir Einzelzell-RNA Sequenzierung (scRNA-Seq) benutzt, wodurch die Erfassung der Genexpression von Tausenden von Zellen in einer Probe möglich ist. Um jedoch sinnvolle Schlüsse aus diesen scRNA-Seq Daten zu ziehen, müssen technische Artifakte identifiziert und aus den Daten entfernt werden. In dieser Dissertation stellen wir eine Methode vor, um eines solcher Artifakte zu erkennen: leere Tröpfchen (ohne Zellen), die hauptsächlich aus freischwebender mRNA in der Probe bestehen. Weiterhin können Zellen anhand ihrer Genexpressionsprofile entlang einer Entwicklungsschiene angeordnet werden, die die Entwicklung der Zellen während ihrer Differenzierung rekapituliert. Auf der Grundlage dieser Entwicklungsreihenfolge modellieren wir die Genexpression mit einem Bayes'schen Inferenzansatz, um die Dynamik der Transkription in sich differenzierenden Zellen zu messen. Dies ermöglicht das Anordnen von Genen entlang einer

Transkriptionskaskade, sowie statistische Untersuchungen in Hinblick auf Unterschiede in der Veränderung von Genexpression, und das Messen des Einflusses möglicher Regulationsgene. Wir wenden diese Methode an, um kortikale neuronale Stammzellen zu untersuchen, die sich über einen intermediären Vorläuferzelltyp in kortikale Neuronen in Hirnorganoiden differenzieren, und um eine detaillierte Charakterisierung der molekularen Prozesse zu liefern, die der Neurogenese zugrunde liegen.

# SELBSTSTÄNDIGKEITSERKLÄRUNG

Name: Rosebrock
Vornamen: Daniel

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht.

Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

*Berlin, 2023*

Daniel Rosebrock